SIAM J. APPL. MATH.
Vol. 68, No. 1, pp. 1–23

FIBER DYNAMICS IN TURBULENT FLOWS: SPECIFIC TAYLOR DRAG*

NICOLE MARHEINEKE[†] AND RAIMUND WEGENER[‡]

Abstract. In [N. Marheineke and R. Wegener, *SIAM J. Appl. Math.*, 66 (2006), pp. 1703–1726], an aerodynamic force concept for a general air drag model based on a stochastic k-ϵ description for a turbulent flow field is derived. The turbulence effects on the dynamics of a long, slender, elastic fiber are specifically modeled by a correlated random Gaussian force and in its asymptotic limit on a macroscopic fiber scale by Gaussian white noise with flow-dependent amplitude. The present paper states quantitative similarity estimates and numerical comparisons for the choice of a Taylor drag model in a given application.

Key words. flexible fibers, k-ϵ turbulence model, fiber-turbulence interaction scales, air drag, random Gaussian aerodynamic force, white noise, stochastic differential equations, ARMA process

AMS subject classifications. 74F10, 76F60, 35R60, 65C20

DOI. 10.1137/06065489X

1. Introduction.

The understanding of the motion of long flexible fibers suspended in highly turbulent air flows is of great interest for textile manufacturing in the melt-spinning process of nonwoven materials. Disregarding the fiber's influence on the flow, the authors of [13] stated a stochastic partial differential system that describes the dynamics of a single slender elastic fiber in a turbulent flow. The turbulence effects are modeled by a correlated Gaussian aerodynamic force. Applying a global-from-local force concept for general air drag models, we can derive these effects, particularly, on the basis of homogeneous Gaussian fields for the randomly fluctuating local velocity components of the flow. Their construction satisfies the requirements of the stochastic k-ϵ turbulence model and Kolmogorov's universal equilibrium theory on local isotropy. On macroscopic scales, white noise with flow-dependent amplitude turns out be a good approximation for the original correlated force according to $\mathcal{L}^2$- and $\mathcal{L}^\infty$-similarity estimates. In the following, we show the applicability of this general force concept under conditions of a real melt-spinning process by choosing an empirically motivated Taylor drag; see Figure 1. Then, the simplified force model satisfies the demands of accuracy on the relevant fiber scale while drastically facilitating the numerical computations at the same time.

For convenience we start with a brief summary of the models for fiber dynamics and aerodynamic force. Dimensional analysis of turbulence and fiber behavior reveals the characteristic interaction scales for our application in section 2. On the fiber macroscale the mean flow dominates the swinging of the fiber, whereas the energy-bearing turbulent vortices of the mesoscale cause the entanglement and fine-loop forming on the fiber that are crucial for the quality of the resulting nonwoven materials. The interest in a macroscopic description of the fiber dynamics justifies

*Received by the editors March 23, 2006; accepted for publication (in revised form) May 24, 2007; published electronically October 10, 2007. This work has been supported by the Kaiserslautern Excellence Cluster *Dependable Adaptive Systems and Mathematical Modeling*.

http://www.siam.org/journals/siap/68-1/65489.html

[†] Department of Mathematics, Technical University Kaiserslautern, P.O. Box 3049, D-67653 Kaiserslautern, Germany (nicole@mathematik.uni-kl.de).

[‡] Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), Fraunhofer-Platz 1, D-67663 Kaiserslautern, Germany (wegener@itwm.fhg.de).

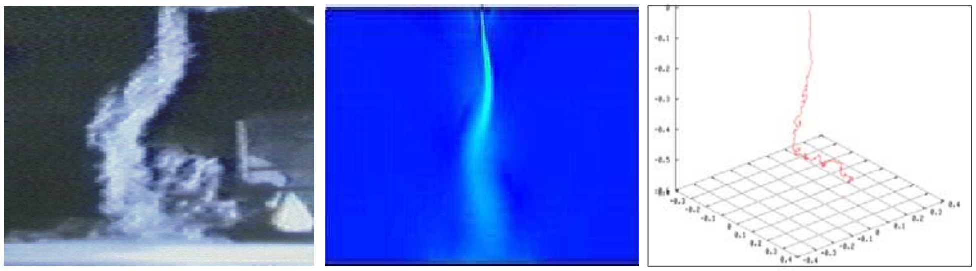


FIG. 1. *From left to right: Turbulent flow in a melt-spinning process, mean velocity flow field by the k-ϵ model, and turbulence effects on fiber dynamics. Photo by industrial partner.*

the use of the simplified force model, as it contains all crucial correlation information of the mesoscale according to the stated quantitative similarity estimates. From the choice of the Taylor drag model, we derive a linear drag operator and thus the concrete correlated and uncorrelated global forces in section 3. Their effects on the fiber dynamics are numerically compared in section 4 by using an introduced curvature measure which yields very convincing results.

1.1. General aerodynamic force model. In the following, we recall the basic models from [13] that are crucial for the description of the fiber dynamics in a turbulent flow. Consider a single long flexible fiber that is fixed at one end and suspended in a subsonic highly turbulent air flow with small pressure gradients and Mach number Ma $< 1/3$. Let l denote the fiber length and d its diameter with aspect ratio $d/l \ll 1$. Whereas the fiber influence on the turbulence is negligibly small due to the slender geometry, the turbulent flow essentially determines the dynamics of the fiber. The motion is modeled by a system of stochastic partial differential equations with algebraic constraint of inextensibility that is deduced from the dynamical Kirchhoff–Love equations for a Cosserat rod being capable of large, geometrically nonlinear deformations,

$$\rho A\, \partial_{tt}\mathbf{r}(s,t) = \partial_s[T(s,t)\, \partial_s\mathbf{r}(s,t)] - EI\, \partial_{ssss}\mathbf{r}(s,t) + \rho A\, \mathbf{g} + \mathbf{f}^{air}(\mathbf{r}(.),s,t), \tag{1.1}$$

$$\|\partial_s\mathbf{r}(s,t)\|_2 = 1, \tag{1.2}$$

with Dirichlet boundary conditions at the fixed end, Neumann at the free end, and the position of rest as the initial condition. Here, $\mathbf{r} : [0,l] \times \mathbb{R}_0^+ \to \mathbb{R}^3$ might be interpreted as the center line of the fiber with arc-length s and time t; its constant line weight is denoted by ρA. The internal line forces stem from bending stiffness indicated by Young's modulus E and moment of inertia I as well as from traction. In this spirit, the Lagrangian multiplier $T : [0,l] \times \mathbb{R}_0^+ \to \mathbb{R}$ can be viewed as the modified tractive force $T = T_t + EI\|\partial_{ss}\mathbf{r}\|_2^2$ containing tension T_t and curvature $\|\partial_{ss}\mathbf{r}\|_2^2$ due to bending. The external line forces acting on the fiber arise from gravity $\mathbf{g}$ and aerodynamics $\mathbf{f}^{air}$.

The aerodynamic force term acts as the additive Gaussian noise in (1.1) due to the applied general global-from-local force concept that is based on the stochastic k-ϵ description of the underlying turbulent flow. In particular, we consider here a correlated Gaussian aerodynamic force $\mathbf{f}_{cc}^{air}$ and its uncorrelated asymptotic limit on macroscopic scales $\mathbf{f}_{uc}^{air}$,

$$\mathbf{f}_{cc}^{air}(\mathbf{r}(.),s,t) = \mathbf{f}(\bar{\mathbf{v}}(s,t),\partial_s\mathbf{r}(s,t)) + \mathbf{L}^{\mathbf{f}}(s,t)\ \frac{\int_{N(\mathbf{r}(.),s,t)} \mathbf{w}_f^{\sigma,\tau}(s,t)\,d\mathcal{W}_{\sigma,\tau}}{(\int_{N(\mathbf{r}(.),s,t)} d\sigma\,d\tau)^{1/2}}, \tag{1.3}$$

$$\mathbf{f}_{uc}^{air}(\mathbf{r}(.),s,t) = \mathbf{f}(\bar{\mathbf{v}}(s,t),\partial_s\mathbf{r}(s,t)) + \mathbf{L}^{\mathbf{f}}(s,t)\ \mathbf{D}^{s,t}\ \mathbf{p}(s,t), \tag{1.4}$$

which depend on the chosen air drag model $\mathbf{f} : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$ and its respective linear drag operator $\mathbf{L}^{\mathbf{f}}$. A feasible air drag model is prescribed as a function of the mean relative velocity between fluid and fiber, i.e., $\bar{\mathbf{v}}(s,t) = \bar{\mathbf{u}}(\mathbf{r}(s,t),t) - \partial_t\mathbf{r}(s,t)$, and the fiber tangent $\partial_s\mathbf{r}(s,t)$. In analogy to the k-ϵ turbulence model, the forces are split into a deterministic part $\bar{\mathbf{f}}$, resulting from the mean flow velocity $\bar{\mathbf{u}} : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^3$, and a stochastic part $\mathbf{f}'$ coming from the turbulent fluctuations that are characterized by the turbulent kinetic energy $k : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^+$ and the dissipation rate $\epsilon : \mathbb{R}^3 \times \mathbb{R}_0^+ \to \mathbb{R}^+$. In (1.3) the random fluctuations are modeled as Ito-integral over a family of locally isotropic, homogeneous, incompressible Gaussian velocity fields along the fiber $\{(\mathbf{w}_f^{\sigma,\tau})_{s,t},\ (s,t) \in [0,l] \times \mathbb{R}_0^+),\ (\sigma,\tau) \in [0,l] \times \mathbb{R}_0^+\}$, where $(\mathcal{W}_{\sigma,\tau},\ (\sigma,\tau) \in [0,l] \times \mathbb{R}_0^+)$ denotes a Wiener process (Brownian motion). The underlying fiber region $N(\mathbf{r}(.),s,t) = \{(\sigma,\tau) \in [0,l] \times \mathbb{R}_0^+ \mid \|\mathbf{r}(s,t) - \mathbf{r}(\sigma,\tau) - \bar{\mathbf{u}}(\mathbf{r}(s,t),t)(t-\tau)\|_2 \leq l_{\mathrm{T}} \wedge |t-\tau| \leq t_{\mathrm{T}}\}$ is determined by the turbulent large-scale length l_{T} and time t_{T}. Moreover, the construction of the correlation tensors $\boldsymbol{\gamma}_0^{\sigma,\tau}(s_1-s_2,\ t_1-t_2) := \mathbb{E}[\mathbf{w}_f^{\sigma,\tau}(s_1,t_1) \otimes \mathbf{w}_f^{\sigma,\tau}(s_2,t_2)]$ corresponding to the centered velocity fields complies with the requirements of the k-ϵ model, Kolmogorov's universal equilibrium theory on local isotropy, as well as Taylor's hypothesis of frozen turbulence pattern, by choosing the following energy spectra $E^{\sigma,\tau} \in \mathcal{C}^2(\mathbb{R}_0^+)$:

$$E^{\sigma,\tau}(\kappa) = \begin{cases} K^{\sigma,\tau}\ \kappa_1^{-5/3}\ \sum_{j=4}^{6} a_j\,(\frac{\kappa}{\kappa_1})^j, & \kappa < \kappa_1, \\ K^{\sigma,\tau}\ \kappa^{-5/3}, & \kappa_1 \leq \kappa \leq \kappa_2, \\ K^{\sigma,\tau}\ \kappa_2^{-5/3}\ \sum_{j=7}^{9} b_j\,(\frac{\kappa}{\kappa_2})^{-j}, & \kappa > \kappa_2, \end{cases} \tag{1.5}$$

$$\int_0^\infty E^{\sigma,\tau}(\kappa)\,d\kappa = k(\mathbf{r}(\sigma,\tau),\tau), \qquad \int_0^\infty E^{\sigma,\tau}(\kappa)\,\kappa^2\,d\kappa = \frac{\epsilon(\mathbf{r}(\sigma,\tau),\tau)}{2\nu} \tag{1.6}$$

with viscosity ν, Kolmogorov constant $K^{\sigma,\tau} = C_{\mathrm{K}}\ \epsilon(\mathbf{r}(\sigma,\tau),\tau)^{2/3}$, and further prescribed constant fitting parameters a_j, b_j. In contrast, in (1.4) the integral effects of the localized random fluctuations are incorporated into the amplitude $\mathbf{D}^{s,t}$ of the Gaussian white noise $(\mathbf{p}_{s,t},\ (s,t) \in [0,l] \times \mathbb{R}_0^+)$, i.e., the $\mathbb{R}^3$-valued random variable $\lim_{(\triangle s,\triangle t)\to\mathbf{0}} \sqrt{\triangle s \triangle t}\,\mathbf{p}(s,t) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ is centered, reduced, and normally distributed. In particular,

$$\mathbf{D}^{s,t} = \left(\frac{2\pi}{\bar{v}_{\mathrm{n}}(s,t)}\ \int_0^\infty \frac{E^{s,t}(\kappa)}{\kappa^2}\,d\kappa\right)^{1/2} \quad \mathbf{P}_{\mathbf{t},\mathbf{n}(s,t)} \tag{1.7}$$

is proportional to the projector $\mathbf{P}_{\mathbf{t},\mathbf{n}}$ onto the plane spanned by fiber tangent $\mathbf{t} = \partial_s\mathbf{r}$ and normal $\mathbf{n} = (\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t})/\|\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t}\|_2$, where $\bar{v}_{\mathrm{n}} = \bar{\mathbf{v}} \cdot \mathbf{n}$. Note that the existence of the amplitude in (1.7), and thus of the uncorrelated force in (1.4), presupposes the linear independence of fiber tangent and mean relative velocity.

2. Fluid-fiber interaction scales. The handling of fiber-turbulence interaction is very difficult, as it is governed by many complex factors, including nature of the flow field, turbulent length scales, and size and behavior of the fiber. The applicability of the uncorrelated aerodynamic force $\mathbf{f}_{uc}^{air}$ particularly depends on the characteristic

Table 1
Typical fiber and flow parameter values in melt-spinning processes.

Fiber			
Diameter	d	$3.0 \cdot 10^{-5}$	m
Length	l	2.5	m
Line weight	ρA	$9.0 \cdot 10^{-7}$	kg/m
Bending stiffness	EI	$4.7 \cdot 10^{-10}$	Nm^2
Absolute velocity	W	$1.0 \cdot 10^{1}$	m/s
Acceleration of gravity	g	9.81	m/s^2
Suspended height	H	1	m

Flow			
Density	ρ^{air}	1.22	kg/m^3
Absolute mean velocity	$\bar{u}$	$1.0 \cdot 10^{2}$	m/s
Turbulent kinetic energy	k	$1.0 \cdot 10^{2}$	m^2/s^2
Dissipation rate	ϵ	$1.0 \cdot 10^{5}$	m^2/s^3
Viscosity	ν	$1.5 \cdot 10^{-5}$	m^2/s

interaction scales of the considered fiber-flow problem. In a typical melt-spinning process, fiber and flow are specified by the parameter values of Table 1. These yield the following quantitative scales and similarity estimates between the correlated and uncorrelated force by using dimensional analysis.

2.1. Turbulence scales. Turbulence is characterized by its wide range of length and time scales. As their significance plays a decisive role in the coming analysis, we focus on them and their interpretation.

Due to the underlying k-ϵ turbulence model, we already distinguish between the length and time scales of the mean motion and those of the fluctuations. The mean motion and its scales are derived from the boundary conditions (geometry) and the absolute mean flow velocity $\bar{u}$. On the other hand, the fluctuations might be interpreted as the turbulent effects of overlapping vortices of different sizes that are indicated by the turbulent kinetic energy k, dissipation rate ϵ, and viscosity ν. The smallest, viscously determined vortices are given by the Kolmogorov scales

$$\eta = \left(\frac{\nu^3}{\epsilon}\right)^{1/4}, \qquad t_{\mathrm{K}} = \left(\frac{\nu}{\epsilon}\right)^{1/2}.$$

Apart from that, the local correlation tensor $\boldsymbol{\gamma}_0$ [12, 13] provides additional information about the size of the present turbulent structures. The structures in the dissipation area (small lengths, thus high frequencies) are determined by the run of the one-dimensional longitudinal correlation function $c_1(z) = 2/z^3 \int_0^\infty \partial_\kappa(E(\kappa)/\kappa) \sin(\kappa z)\, d\kappa$, $z \in \mathbb{R}_0^+$ around the origin and hence by k and ϵ; see (1.6). For $z \ll 1$, $c_1(z) = 2/3k - \epsilon/(30\,\nu)\, z^2 + \mathcal{O}(z^4)$ then describes a parabola that intersects the z-axis at the dissipation length λ_{T}, i.e., $c_1(\lambda_{\mathrm{T}}) = 0$. Thus,

$$\lambda_{\mathrm{T}} = \left(\frac{20k\nu}{\epsilon}\right)^{1/2}$$

represents the turbulent fine or microscale for the decay of the correlations.

In contrast, the large, macro, or integral scale

$$\Lambda_{\mathrm{T}} = \frac{\int_0^\infty \mathrm{tr}\boldsymbol{\gamma}_0(z)\, dz}{\mathrm{tr}\boldsymbol{\gamma}_0(0)} = \frac{\pi}{2}\, \frac{\int_0^\infty E(\kappa)/\kappa\, d\kappa}{\int_0^\infty E(\kappa)\; d\kappa}$$

characterizes the mean coherence scale independently of longitudinal and lateral correlations and can be interpreted as the typical size of the energy-bearing vortices. In this context, the turbulent length proposed by the k-ϵ model,

$$l_{\mathrm{T}} = \frac{k^{3/2}}{\epsilon},$$

can be understood as the leading order term of Λ_{T}; compare with the modeled energy spectrum of (1.5). The energy spectrum gives

$$\Lambda_{\mathrm{T}} = \frac{\pi\, C_{\mathrm{K}}}{2} \frac{\epsilon^{2/3}}{k} \left(A_1 \kappa_1^{-5/3} + B_1 \kappa_2^{-5/3} \right),$$

where κ_1 and κ_2 with $\kappa_2 > \kappa_1 > 0$ are the solutions of the nonlinear system

$$\begin{aligned} A_k \kappa_1^{-2/3} + B_k \kappa_2^{-2/3} &= \frac{k}{C_{\mathrm{K}} \epsilon^{2/3}} = f_k, \\ A_\epsilon \kappa_1^{4/3} + B_\epsilon \kappa_2^{4/3} &= \frac{\epsilon^{1/3}}{2 C_{\mathrm{K}} \nu} = f_\epsilon = \frac{f_k}{\delta^2}, \end{aligned} \tag{2.1}$$

stemming from (1.6). After nondimensionalizing, $\delta = (2k\nu/\epsilon)^{1/2} \sim \mathcal{O}(\lambda_{\mathrm{T}})$ with $\lambda_{\mathrm{T}}/H \ll 1$ turns out to be small, whereas the other coefficients $A_i, B_i, f_k \sim \mathcal{O}(1)$. Thereby, $A_i, B_i, i = 1, k, \epsilon$, denote linear combinations of the fitting parameters arising in (1.5), and $C_{\mathrm{K}} = 0.5$ is the Kolmogorov constant. Substituting $x_i = \kappa_i^{2/3}$, $i = 1, 2$, we write $x_1 = f_k/A_k - B_k/A_k\, x_2$. Inserting this expression into (2.1) yields a 4th order equation for x_2 that has two complex as well as two real solutions—a negative and a positive. The feasible positive solution can be expanded in δ as $x_2 = x_2^{(1)} \delta + x_2^{(3)} \delta^3 + \mathcal{O}(\delta^4)$, which results directly in a δ-series for Λ_{T},

$$\Lambda_{\mathrm{T}} = F_1\, l_{\mathrm{T}} + \mathcal{O}(\delta) \qquad \text{with } F_1 = \frac{\pi}{2} \frac{A_1}{C_{\mathrm{K}}^{3/2} A_k^{5/2}} \approx 1.05. \tag{2.2}$$

In spite of the use of A_i in (2.2), the magnitude of F_1 can be treated as independent of the differentiability order of the underlying chosen energy model. An ansatz for a smoother energy spectrum, $E \in \mathcal{C}^l(\mathbb{R}_0^+)$, $l \geq 3$, certainly contains more fitting parameters, but their influence cancels out in the definition of F_1. In this work, we refer to l_{T} as the turbulent large-scale length.

Concerning the turbulent time scale for the decay of the energy-bearing vortices, the length l_{T} and velocity scale $u_{\mathrm{T}} = k^{1/2}$ of the k-ϵ model imply

$$t_{\mathrm{T}} = \frac{k}{\epsilon}.$$

As this scale does not take into account the advective influence of the mean flow, we suggest additionally

$$t_{\mathrm{A}} = \frac{l_{\mathrm{T}}}{\bar{u}} = \frac{k^{3/2}}{\epsilon\, \bar{u}}.$$

Moreover, the amplitude $\mathbf{D}$ of the uncorrelated force in (1.7) might also be expressed by k and ϵ, since it contains a moment of the energy spectrum. In our case, we get $\int_0^\infty E(\kappa)/\kappa^2\, d\kappa = C_{\mathrm{K}} \epsilon^{2/3} (A_2 \kappa_1^{-8/3} + B_2 \kappa_2^{-8/3})$, where $A_2, B_2 \sim \mathcal{O}(1)$

are linear combinations of the fitting parameters. Following the approach above and introducing the small parameter δ, the expansion for the energy moment reads

$$\int_0^\infty \frac{E(\kappa)}{\kappa^2}\,d\kappa = F_2\,\frac{k^4}{\epsilon^2} + \mathcal{O}(\delta) \qquad \text{with } F_2 = \frac{A_2}{C_{\mathrm{K}}^3\,A_k^4} \approx 0.80. \tag{2.3}$$

In leading order, the amplitude is consequently given by

$$\mathbf{D}^{(0)} = \left(\frac{2\pi F_2}{\bar{v}_{\mathrm{n}}}\right)^{1/2} \frac{k^2}{\epsilon} \quad \mathbf{P}_{\mathbf{t},\mathbf{n}}\,. \tag{2.4}$$

Thus, the resulting correlations along the fiber $(\mathbf{D}^{(0)})^2\,\delta_0(s)\,\delta_0(t)$ can be interpreted as being proportional to the turbulent energy k acting over the mean coherence length l_{T} and over the characteristic turbulent fiber time $\tau_{\mathrm{T}}^f = l_{\mathrm{T}}/\bar{v}_{\mathrm{n}}$ that depends on the geometrical relation between fiber orientation and mean relative velocity.

2.2. Fiber scales. For a better understanding of the fiber behavior in the turbulent flow, dimensional analysis is applied on the fiber system (1.1), (1.2). Therefore, we introduce a dimensionless zooming parameter $h = L/H$ as a ratio of the typical varying length of interest $L \in [0, l]$ and the fixed height of the suspended fiber H, where l denotes the fiber length.

Apart from H, the problem contains nine parameters: diameter d, line weight ρA, bending stiffness EI, fiber velocity W, acceleration of gravity g, flow density ρ^{air}, mean flow velocity $\bar{u}$, mean relative velocity between flow and fiber $\bar{v}$, and kinetic turbulent energy k. The number of parameters can be reduced to four dimensionless:

$$\mathrm{Fr} = \frac{W^2}{g\,H}, \quad \mathrm{Gr} = \frac{\rho A\,g\,H^3}{EI}, \quad \bar{\mathrm{P}} = \frac{d\,\rho^{air}\,H^3\,\bar{v}^2}{EI}, \quad \mathrm{P}' = \frac{d\,\rho^{air}\,H^3\,k^{1/2}\,\bar{v}}{EI}.$$

The Froude number Fr states the ratio of kinetic and gravitational potential energy, the dimensionless gravity Gr the ratio of gravitational and flexural energies, and the dimensionless mean $\bar{\mathrm{P}}$ and fluctuating aerodynamic force P' the ratio of aerodynamic and flexural energies. Introducing dimensionless variables gives

$$\begin{aligned} \mathbf{r}(s,t) &= H\; r^*(s^*,t^*), & T_t(s,t) &= \frac{EI}{L\,H}\; T_t^*(s^*,t^*), \\ \bar{\mathbf{f}}(s,t) &= d\,\rho^{air}\left(\frac{\bar{v}}{h}\right)^2 \bar{\mathbf{f}}^*(s^*,t^*), & \mathbf{f}'(s,t) &= d\,\rho^{air} k^{1/2}\left(\frac{\bar{v}}{h}\right) \mathbf{f}'^*(s^*,t^*), \end{aligned}$$

with $s = L\,s^*$ and $t = (L/W)\,t^*$. Here, two different scalings are used for fiber curve $\mathbf{r}$ and arc-length s, $\mathbf{r}$ is scaled by the suspended height of the fiber H, and s by the typical length of interest L. This choice is motivated by our interest in the whole spatial domain of the fiber line while zooming in on certain fiber lengths. This allows us to investigate the characteristic fiber behavior, e.g., bending, loop forming, crimping or stiffness, arising on typical scales. Hence, the interplay of the fixed outer H and the varying inner length L appears also in the factor of the tension part T_t. The bending part is treated separately due to the composed structure of T. For the scaling of the aerodynamic force $\mathbf{f}^{air}$, it is sufficient to utilize its proportionality to the dynamic pressure, since $\|\mathbf{f}^{air}\|_2 \sim d\rho^{air}\|\mathbf{v}\|_2^2$ in the following. Thereby, the deterministic force part $\bar{\mathbf{f}}$ is based on the quadratic mean relative velocity, and the stochastic part $\mathbf{f}'$ on the product of mean relative velocity and flow fluctuations that are expressed by $k^{1/2}$.

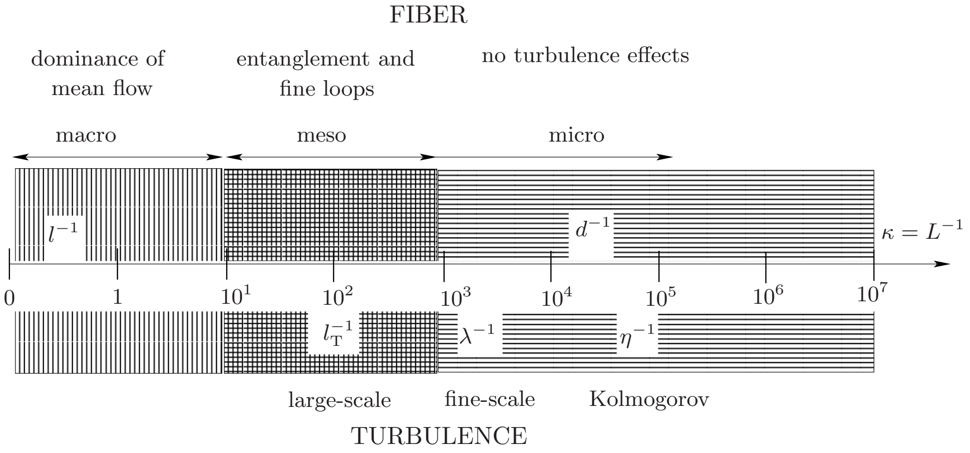


FIG. 2. *Scales of fiber-turbulence interactions in melt-spinning processes (cf. Tables* 1 *and* 2*).*

The magnitude of the mean relative velocity $\bar v$ depends particularly on the direction of mean flow and fiber velocity according to

$$\bar{\mathbf{v}}(s,t) = \|\bar u\,\bar{\mathbf{u}}^* - W/h\;\partial_{t^*}\mathbf{r}^*\|_2\;\bar{\mathbf{v}}^*(s^*,t^*) = (\bar v/h)\,\bar{\mathbf{v}}^*(s^*,t^*). \tag{2.5}$$

It is minimal if $\bar{\mathbf{u}}^\star$ and $\partial_{t^\star}\mathbf{r}^\star$ are similarly directed, and maximal if they are opposite directed; thus $\bar v \in [\,|h\bar u - W|,\,|h\bar u + W|\,]$. The time scaling in (2.5) that is chosen with respect to the fiber dynamics of the typical length L incorporates here the zooming ratio h into the definition of $\bar v$. Then, the dimensionless fiber system reads

$$\begin{aligned}\mathrm{Fr}\,\mathrm{Gr}\,\partial_{tt^\star}\mathbf{r}^\star =&\ \partial_{s^\star}\big((h^{-1}\,T_t^\star + h^{-4}\,\|\partial_{ss^\star}\mathbf{r}^\star\|_2^2)\,\partial_{s^\star}\mathbf{r}^\star\big) - h^{-2}\,\partial_{ssss^\star}\mathbf{r}^\star - h^2\,\mathrm{Gr}\,\mathbf{e_3}\\ &+\bar{\mathrm{P}}\,\bar{\mathbf{f}}^\star + h\,\mathrm{P}'\,\mathbf{f}'^\star,\\ (\partial_{s^\star}\mathbf{r}^\star)^2 =&\ h^2.\end{aligned}$$

For a melt-spinning process, the typical fiber and flow parameter values listed in Table 1 yield

$$\mathrm{Fr}\sim 10^1,\quad \mathrm{Gr}\sim 10^4,\quad \bar{\mathrm{P}}\sim\begin{cases}10^8-10^9, & h\sim 1,\\ 10^6-10^7, & h\ll 1,\end{cases}\qquad \mathrm{P}'\sim\begin{cases}10^7-10^8, & h\sim 1,\\ 10^6-10^7, & h\ll 1,\end{cases}$$

where the aerodynamic similarity quantities $\bar{\mathrm{P}}$ and P' are roughly estimated by means of the range of $\bar v$. Varying the length of interest L and thus the zooming parameter $h = L/H$ reveals three characteristic scales for the fiber-turbulence problem in the technical application that are worth considering in more detail; cf. Figure 2. In the following we suppress the superscript * to keep the expressions short.

Macroscale: $1 \geq h > 10^{-1}$.

$$\mathrm{Fr}\,\mathrm{Gr}\,\partial_{tt}\mathbf{r} = -h^2\,\mathrm{Gr}\,\mathbf{e_3} + \bar{\mathrm{P}}\,\bar{\mathbf{f}} + h\,\mathrm{P}'\,\mathbf{f}'.$$

Over the whole length of the fiber l, the fiber dynamics is caused by the external forces. In particular, the mean flow affects the fiber swinging.

Mesoscale: $10^{-1} \geq h > 10^{-3}$.

$$\mathrm{Fr}\,\mathrm{Gr}\,\partial_{tt}\mathbf{r} = \partial_s\big((h^{-1}T_t + h^{-4}\|\partial_{ss}\mathbf{r}\|_2^2)\,\partial_s\mathbf{r}\big) - h^{-2}\,\partial_{ssss}\mathbf{r} + \bar{\mathrm{P}}\,\bar{\mathbf{f}} + h\,\mathrm{P}'\,\mathbf{f}'.$$

Table 2

Overview of turbulence and fiber scales as well as dimensionless numbers for the fiber-turbulence problem deduced from the typical parameter values in melt-spinning processes (cf. Table 1).

Turbulence scales			
Large length scale	l_{T}	$1.0 \cdot 10^{-2}$	m
Fine length scale	λ_{T}	$5.4 \cdot 10^{-4}$	m
Kolmogorov length scale	η	$1.4 \cdot 10^{-5}$	m
Turbulent time scale	t_{T}	$1.0 \cdot 10^{-3}$	s
Advection time scale	t_{A}	$1.0 \cdot 10^{-4}$	s

Fiber scale			
Typical length of interest	L	$[0, 2.5]$	m

Dimensionless numbers of fiber-turbulence problem			
Zooming ratio	h	$[0, 2.5]$	
Froude number	Fr	10^1	
Gravity number	Gr	10^4	
Mean force number	$\bar{\mathrm{P}}$	$[10^8 - 10^9]$, $h \sim 1$,	$[10^6 - 10^7]$, $h \ll 1$
Fluctuating force number	P'	$[10^7 - 10^8]$, $h \sim 1$,	$[10^6 - 10^7]$, $h \ll 1$
Spatial smoothing parameter	α_s	10^{-2}	
Temporal smoothing parameter	α_t	10^{-3}	

This fiber scale coincides with the turbulent large-scale l_{T} of the energy-bearing vortices. Here, the inner and outer forces acting on the fiber balance each other. But the fluctuating part of the aerodynamic force $\mathbf{f}'$ causes entanglement and fine-loop forming which crucially determine the fiber dynamics.

Microscale: $h \leq 10^{-3}$.

$$\partial_s(h^{-4}\|\partial_{ss}\mathbf{r}\|_2^2\,\partial_s\mathbf{r}) = \mathbf{0}, \qquad \mathrm{Fr}\,\mathrm{Gr}\,\partial_{tt}\mathbf{r} = -h^{-2}\,\partial_{ssss}\mathbf{r} + \bar{\mathrm{P}}\,\bar{\mathbf{f}}.$$

The inner forces, in particular the bending stiffness, dominate the total fiber behavior. In contrast, the effects of the fine-scale λ_{T} and Kolmogorov vortices of size η are irrelevant for the fiber dynamics; here $\eta < d$ (cf. Tables 1 and 2).

The time scales of the problem provide no further information, as they are related to the length scales using the reciprocal of the fiber velocity W as a proportionality factor. Due to its inertia, the fiber shows thus no reaction to turbulent structures decaying faster than $t_{inertia} = h_{micro}H/W \sim 10^{-4}$ s, which includes the whole fine-scale turbulence. The natural decay of the large-scale vortices in contrast is indicated by $t_{\mathrm{T}} \sim 10^{-3}$ s and, under consideration of advection, by the mean flow by $t_{\mathrm{A}} \sim 10^{-4}$ s.

Summing up, fine-scale vortices do not affect a fiber in the melt-spinning process because of bending stiffness of the fiber. Thus, their influence (correlations) might be neglected in the model of the stochastic aerodynamic force. In contrast, the turbulent large-scale vortices cause entanglement and loop-forming, which play a decisive role in the fiber behavior. But instead of resolving their effects explicitly, it is sufficient to model them on the macroscale, as our interest focuses exclusively on a macroscopic description for the fiber dynamics. This motivates the idea of approximating the correlated force by an integrated—still correlated—force. In the following, the introduced uncorrelated aerodynamic force $\mathbf{f}_{uc}^{air}$ of (1.4) that contains the mean turbulent coherences (integral correlations) turns out to satisfy the stated demands on approximability.

2.3. Quantitative similarity estimates. To justify the applicability of the uncorrelated force as a substitute for the original correlated force in our problem, we analyze its approximation properties by means of the similarity estimates taken from [13].

Similarity estimates (see [13]). *Let $\alpha_s, \alpha_t \in \mathbb{R}_0^+$ be spatial and temporal smoothing parameters of the fiber-flow problem. Define $\mathcal{E}(\kappa_1,\kappa_2) := \int_{\mathbb{R}} E(\|\kappa_1,\kappa_2,l\|_2)/(\kappa_1,\kappa_2,l)^2\,dl$ with $\mathcal{E}_0 := \mathcal{E}(0,0)$ and $\mathcal{S} := \sup_{\boldsymbol{\kappa}\in[0,1]^2}\|\nabla_{\boldsymbol{\kappa}}\mathcal{E}(\kappa_1,\kappa_2)\|_2$. Then, the approximability of the correlated by the uncorrelated aerodynamic force given in* (1.3), (1.4) *is expressed by the following estimates:*

$\mathcal{L}^2$-similarity:

$$\mathcal{I}_{\mathcal{L}^2} \leq \frac{\sqrt{\alpha_s\,\alpha_t}}{\sqrt{6}\,\pi\,\bar{v}_{\mathrm{n}}}\sqrt{\mathcal{S}^2\left(\alpha_s^2\left(1+\frac{\bar{v}_{\mathrm{t}}^2}{\bar{v}_{\mathrm{n}}^2}\right)+\frac{\alpha_t^2}{\bar{v}_{\mathrm{n}}^2}\right)+\frac{8\mathcal{E}_0^2}{3\pi}\left(\alpha_s^3+\frac{\alpha_t^3}{(\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|)^3}\right)}; \tag{2.6}$$

$\mathcal{L}^\infty$-similarity:

$$\begin{aligned}\mathcal{I}_{\mathcal{L}^\infty} \leq \frac{\sqrt{2}\,\alpha_s\,\alpha_t}{\pi^2\,\bar{v}_{\mathrm{n}}}\Bigg[&\mathcal{S}\left(\alpha_s\left(1+\frac{\bar{v}_{\mathrm{t}}}{\bar{v}_{\mathrm{n}}}\right)\left(\frac{c}{2}+\ln\left(\frac{1}{\alpha_s}\right)\right)+\frac{\alpha_t}{\bar{v}_{\mathrm{n}}}\left(\frac{c}{2}+\ln\left(\frac{\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|}{\alpha_t}\right)\right)\right)\\ &+\mathcal{E}_0\left(\alpha_s+\frac{\alpha_t}{\bar{v}_{\mathrm{n}}+|\bar{v}_{\mathrm{t}}|}\right)\Bigg],\end{aligned} \tag{2.7}$$

where $\bar{v}_{\mathrm{t}}$, $\bar{v}_{\mathrm{n}}$ are the tangential and normal component of the mean relative velocity with respect to the $(\mathbf{t},\bar{\mathbf{v}})$-induced fiber basis of section 1.1 *and $c = \int_0^1 (1-\cos\iota)/\iota\,d\iota$.*

The limit $\alpha_i \to 0$, $i = s,t$, describes the smoothing over the whole $\mathbb{R}^2$. This is unrealistic, as the fiber length l prescribes a natural upper bound for the spatial smoothing parameter α_s. Thus, $\alpha_s = l_{\mathrm{T}}/l$ is certainly a reasonable value for the macroscopic description of the turbulent flow effects on the fiber. The temporal flow and fiber scales are related to the spatial ones by the respective velocities $\bar{u}$ and W, which motivates the choice of $\alpha_t = t_{\mathrm{A}}W/l = \alpha_s W/\bar{u}$.

Inserting the typical parameter values of Tables 1 and 2 yields for the dimensionless smoothing values $\alpha_s \sim 10^{-2}$ and $\alpha_t \sim 10^{-3}$ as well as for the quantities $\mathcal{S}$ and $\mathcal{E}_0$ in standard international units (SI-units) $\mathcal{S} \sim 1$ $[m^5/s^2]$ and $\mathcal{E}_0 \sim k^4/\epsilon^2 \sim 10^{-2}$ $[m^4/s^2]$ according to (2.3). The order of the relative velocity can be approximated by $\bar{v} \sim 10^2$ which implies $|\bar{v}_{\mathrm{t}}| \in [0,10^2]$ and $\bar{v}_{\mathrm{n}} \in [0,10^2]$. Thus, quantitative similarity estimates in SI-units depend drastically on the relation between fiber direction $\mathbf{t} = \partial_s\mathbf{r}$ and mean relative velocity $\bar{\mathbf{v}}$, as they are expressed by

$$\mathcal{I}_{\mathcal{L}^2}^2 \lesssim 10^{-10}\,\bar{v}_{\mathrm{n}}^{-2} + 10^{-6}\,\bar{v}_{\mathrm{n}}^{-4}, \qquad \mathcal{I}_{\mathcal{L}^\infty} \lesssim 10^{-8}\,\bar{v}_{\mathrm{n}}^{-1} + 10^{-6}\,\bar{v}_{\mathrm{n}}^{-2}$$

with $\mathbf{n} = (\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t})/\|\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t}\|_2$. In the case of $\mathbf{t} \perp \bar{\mathbf{v}}$, we have $\bar{v}_{\mathrm{n}} \sim 10^2$ such that $\mathcal{I}_{\mathcal{L}^2} \lesssim 10^{-7}$ and $\mathcal{I}_{\mathcal{L}^\infty} \lesssim 10^{-10}$ indicate very good approximation properties. But even for smaller normal velocity components—down to $\bar{v}_{\mathrm{n}}^{crit} \sim 10^{-1}$—the uncorrelated force is a good substitute for the correlated one, since the deviations are little, i.e., $\mathcal{I}_{\mathcal{L}^2} \lesssim 10^{-1}$, $\mathcal{I}_{\mathcal{L}^\infty} \lesssim 10^{-4}$. In fact, $\bar{v}_{\mathrm{n}} \sim 1$ in general, and the events $\bar{v}_{\mathrm{n}} < \bar{v}_{\mathrm{n}}^{crit}$ might be viewed as elements of a nullset, because the perturbing influence of turbulence and fiber inertia prevents the fiber from moving continuously within the mean streamlines. However, further numerical realization also requires their treatment, so in section 3.3 we will deal with the arising singularity for $\bar{v}_{\mathrm{n}} \to 0$ which results from the definition of the force amplitude $\mathbf{D}$, (1.7).

3. Air drag model and its consequences.

The numerical simulations of the fiber dynamics imposed by the correlated and/or uncorrelated aerodynamic force rely essentially on the choice of an appropriate air drag model $\mathbf{f}$ and the derivation of the corresponding linear drag operator $\mathbf{L}^{\mathbf{f}}$. We particularly distinguish between linear and quadratic drag relations and discuss their applicability as well as their consequences for our application.

3.1. Choice of drag model.

Stokes drag for turbulent flow. For slow viscous flows with $Re < 1$, Cox [4] has developed an insightful analytical series approximation for the force distribution along the length l of a straight fiber. As the Reynolds number based on the fiber diameter d approaches zero, the drag force per unit length along the fiber is proportional to the relative velocity between fluid and fiber $\mathbf{v}(s,t) = \mathbf{u}(\mathbf{r}(s,t),t) - \partial_t \mathbf{r}(s,t)$ at fiber point s and time t. So,

$$\mathbf{f}(\mathbf{v},\mathbf{t}) = \mathbf{C}^{drag}(\mathbf{t})\ \mathbf{v}, \qquad \mathbf{C}^{drag}(\mathbf{t}) = c_{\mathrm{t}}\, \mathbf{t}\otimes\mathbf{t} + c_{\mathrm{n}}\, (\mathbf{I} - \mathbf{t}\otimes\mathbf{t}) \tag{3.1}$$

gives the linear Stokes drag relation, where the drag tensor $\mathbf{C}^{drag}$ depends on the fiber orientation $\mathbf{t} = \partial_s \mathbf{r}$ in the surrounding flow. From the Stokes flow approximation, Keller and Rubinow [10] have determined the drag coefficients $c_{\mathrm{n}}, c_{\mathrm{t}}$ up to leading order for smooth ellipsoidal filaments which also conform for small surface variations [1]. Götz and Unterreiter [7], in contrast, have derived an integral equation model for the drag force by applying a matching principle to the asymptotic expansions of the flow field around slender ellipsoidal and cylindrical fibers of circular cross sections in the framework of Stokes' and Oseen's equations. Then with $\mu = \rho^{air}\nu$,

$$c_{\mathrm{n}}^{ellipsoid} = \frac{8\pi\mu}{Re}\left(\ln\left(\frac{2l}{d}\right) + \frac{1}{2}\right)^{-1}, \qquad c_{\mathrm{t}}^{ellipsoid} = \frac{4\pi\mu}{Re}\left(\ln\left(\frac{2l}{d}\right) - \frac{1}{2}\right)^{-1},$$
$$c_{\mathrm{n}}^{cylinder} = \frac{8\pi\mu}{Re}\left(\ln\left(\frac{4l}{d}\right) - \frac{1}{2}\right)^{-1}, \qquad c_{\mathrm{t}}^{cylinder} = \frac{4\pi\mu}{Re}\left(\ln\left(\frac{4l}{d}\right) - \frac{3}{2}\right)^{-1}.$$

However, there is no slender-body theory that is strictly valid for the turbulent flow with high Re that is of interest here, $Re \approx 200$. In the analysis of turbulence effects on particles, a linear Stokes drag has successfully been applied to predict particle motions in turbulent flows [15, 16, 18, 20]. Drag relations based on empirical correlations have also been used [3, 14] as well as a modified Stokes drag that takes into account particle oscillations [9]. As a necessary simplification, the form of the drag force, (3.1), on the fiber under creeping flow conditions is assumed to be retained for high Re turbulent flows. But (3.1) has been derived for a small Reynolds number flow. Thus, it is only valid for infinitely thin, small fibers with $d \leq \eta$ and $l \leq \eta$. Anyhow, the relation is conferrable to longer fibers suspended in turbulent flow by imposing the free-draining approximation, which has been used to model flexible fiber motion [17] and polymer dynamics [6]. In this model, the fiber is considered to be composed of a series of elements of length Δ_l, where $\Delta_l \leq \eta$. Each element meets the necessary conditions for (3.1) to be valid. Assuming hydrodynamic independence of each element allows (3.1) to be applied to all elements and thus to the entire fiber.

Taylor drag. For high Reynolds number flow indicated by $Re \in (20, 10^6)$, Taylor [19] has investigated the behavior of drag forces experimentally. Thereby, he has discovered the nonlinear relation between drag and angle of attack α between the flow direction and center line of an immersed straight slender body as well as the influence of the surface roughness on the drag, which Lee [11] has applied successfully to long, flexible fibers within a carding process.

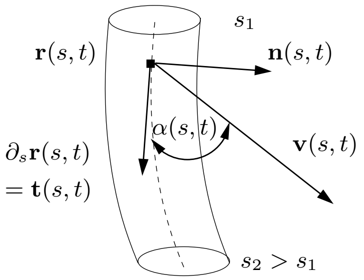


FIG. 3. *Drag-relevant angle* $\alpha \in [0,\pi]$ *enclosed by relative velocity* $\mathbf{v}$ *and fiber tangent* $\partial_s\mathbf{r}$.

As the drag force $\mathbf{f}$ lies in the plane spanned by the fiber tangent and the relative velocity, it can be split into a tangential $\mathbf{f_t}$ and a normal component $\mathbf{f_n}$ with respect to the fiber orientation, i.e., $\mathbf{t} = \partial_s\mathbf{r}$, $\mathbf{n} = (\mathbf{v} - (\mathbf{v}\cdot\mathbf{t})\mathbf{t})/\|\mathbf{v} - (\mathbf{v}\cdot\mathbf{t})\mathbf{t}\|_2$; cf. Figure 3. Then

$$\mathbf{f}(\mathbf{v},\mathbf{t}) = \mathbf{f_n}(\mathbf{v},\mathbf{t}) + \mathbf{f_t}(\mathbf{v},\mathbf{t}), \tag{3.2}$$

where

$$\mathbf{f_n} = 0.5\,\rho^{air}\,d\,\mathbf{v}^2\,\left(\sin^2\alpha + 4\sqrt{\frac{\sin^3\alpha}{Re}}\right)\,\mathbf{n}, \tag{3.3}$$

$$\mathbf{f_t} = 0.5\,\rho^{air}\,d\,\mathbf{v}^2\,\left(5.4\,\cos\alpha\sqrt{\frac{\sin\alpha}{Re}}\right)\,\mathbf{t}, \tag{3.4}$$

with $\sin\alpha = (\mathbf{v}\cdot\mathbf{n})/\|\mathbf{v}\|_2$, $\cos\alpha = (\mathbf{v}\cdot\mathbf{t})/\|\mathbf{v}\|_2$, and $Re = dv/\nu$, respectively. Equations (3.3), (3.4) suggest that a straight fiber with smooth surface does not feel any drag when it is aligned parallel to the direction of the incoming flow. This does not correspond to the experiments in [19] revealing that for small α, $\alpha \to 0$, $\mathbf{f_t}$ can be approximated by $\mathbf{f_t}(\alpha^\circ = \pi/36)$. In contrast, for a rough surface this situation of zero drag does not appear because the Taylor expression reads

$$\mathbf{f} = 0.5\,\rho^{air}\,d\,\mathbf{v}^2\,\left[\left(\sin^2\alpha + \frac{4\sin\alpha}{\sqrt{Re}}\right)\,\mathbf{n} + \cos\alpha\,\mathbf{t}\right]. \tag{3.5}$$

For technical reasons, we rewrite (3.3)–(3.5) as

$$\mathbf{f_n} = 0.5\,\rho^{air}\,d\,c_{\mathrm{n}}\,\|\mathbf{v_n}\|_2\,\mathbf{v_n}, \qquad \mathbf{f_t} = 0.5\,\rho^{air}\,d\,c_{\mathrm{t}}\,\|\mathbf{v_t}\|_2\,\mathbf{v_t} \tag{3.6}$$

with the empirical drag coefficients for smooth, resp., rough, fibers

$$\begin{aligned} c_{\mathrm{n}}^{smooth} &= 1 + 4\sqrt{\nu/(d\|\mathbf{v_n}\|_2)}, & c_{\mathrm{t}}^{smooth} &= 5.4\sqrt{\nu\|\mathbf{v_n}\|_2/(d\|\mathbf{v_t}\|_2^2)}, \\ c_{\mathrm{n}}^{rough} &= 1 + 4\sqrt{\nu\|\mathbf{v}\|_2/(d\|\mathbf{v_n}\|_2^2)}, & c_{\mathrm{t}}^{rough} &= \|\mathbf{v}\|_2/\|\mathbf{v_t}\|_2. \end{aligned} \tag{3.7}$$

The high Reynolds number flow and the presence of very small vortices indicated by the relation $\eta < d$ in our application conflicts with the use of the heuristic linear Stokes drag. Hence, we determine the aerodynamic forces on the smooth polymer fiber under consideration by means of the empirically motivated nearly quadratic Taylor drag (3.6), (3.7), although this concept is examined only for high Re, but still laminar inflow regime. Additionally, to exclude the zero drag in the case of parallelism of $\mathbf{t}$ and $\mathbf{v}$, we suggest a slight modification of the drag coefficient c_{t}^{smooth} that provides more realistic results. As a smooth fiber lying parallel to the direction of the relative velocity experiences the same tangential drag force as one being rotated by α°, and as $\mathbf{v_n} = \mathbf{v} - (\mathbf{v}\cdot\mathbf{t})\mathbf{t}$, we define

$$\mathbf{v}_{\mathbf{n}}^{\circ} := \begin{cases} \mathbf{v_n}, & c^\circ \geq \|\mathbf{v_t}\|_2/\|\mathbf{v}\|_2, \\ \mathbf{v} - \mathrm{sgn}(\mathbf{v}\cdot\mathbf{t})c^\circ\|\mathbf{v}\|_2\,\mathbf{t} & \text{else} \end{cases} \tag{3.8}$$

with $c^\circ = \cos\alpha^\circ$. Here, the sign function, $\mathrm{sgn}(x) = 1$ if $x \geq 0$, $\mathrm{sgn}(x) = -1$ else, includes equal and opposite directed vectors $\mathbf{t}$ and $\mathbf{v}$. We have $\|\mathbf{v}_{\mathbf{n}}^{\circ}\|_2 = 0$ if and only if $\|\mathbf{v}\|_2 = 0$. Setting

$$c_{\mathrm{t}}^{smooth} = 5.4\sqrt{\nu\|\mathbf{v}_{\mathbf{n}}^{\circ}\|_2/(d\|\mathbf{v_t}\|_2^2)} \tag{3.9}$$

thus yields a reasonable tangential drag model that is not only continuous but proves to also be differentiable.

3.2. Linear drag operator. Proceeding with the derivation of the linear drag operator $\mathbf{L^f}$, we consider a generalized linearization approach for the modified Taylor drag model $\mathbf{f}$,

$$\mathbf{f}(\bar{\mathbf{v}} + \mathbf{u}', \mathbf{t}) \approx \mathbf{f}(\bar{\mathbf{v}}, \mathbf{t}) + \mathbf{L^f}(\bar{\mathbf{v}}, \mathbf{t}, k)\ \mathbf{u}', \tag{3.10}$$

with mean relative velocity between fluid and fiber $\bar{\mathbf{v}}$ and random Gaussian fluctuation of the flow velocity $\mathbf{u}'$. In the context of (1.3), (1.4), the first term represents the deterministic part of the aerodynamic forces and the second term the stochastic one.

Model for linear drag operator. *Let* $\mathbf{f} : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$ *be the modified Taylor drag model of* (3.6)–(3.9). *Then construct the linear drag operator* $\mathbf{L^f}$ *as continuous composition*

$$\mathbf{L^f}(\bar{\mathbf{v}}, \mathbf{t}, k) = \begin{cases} \nabla_{\mathbf{v}}\mathbf{f}(\bar{\mathbf{v}}, \mathbf{t}), & \varpi > 1, \\ \\ (1-\varpi)\left(a_{\mathrm{n}_0}(k)\ (\mathbf{I} - \mathbf{P_t}) + \ a_{\mathrm{t}_0}(k)\ \mathbf{P_t}\right) & \\ \quad +\varpi\ \nabla_{\mathbf{v}}\mathbf{f}(\varpi^{-1}\,\bar{\mathbf{v}}, \mathbf{t}), & \varpi \leq 1, \end{cases} \tag{3.11}$$

with $\varpi = \|\bar{\mathbf{v}}\|_2\,(2k)^{-1/2}$. *The parameters are given by*

$$a_{\mathrm{n}_0}(k) = \left(2a_{\mathrm{n}_1}^2 k + 5\sqrt{2^5}/\sqrt{3^{3/2}\pi}\,\mathrm{gam}(5/4)a_{\mathrm{n}_1}a_{\mathrm{n}_2}k^{3/4} + 16/\sqrt{3\pi}\,a_{\mathrm{n}_2}^2 k^{1/2}\right)^{1/2}, \tag{3.12}$$

$$a_{\mathrm{t}_0}(k) = \sqrt{8/(3\pi)^{1/2}}\,a_{\mathrm{t}}k^{1/4} \tag{3.13}$$

with $a_{\mathrm{n}_1} = 0.5\rho^{air}d$, $a_{\mathrm{n}_2} = \rho^{air}\sqrt{d\nu}$, $a_{\mathrm{t}} = 1.35a_{\mathrm{n}_2}$, $c^\circ = \cos\alpha^\circ$, *and gamma function* gam.

Let the projectors on fiber tangent $\mathbf{t} = \partial_s\mathbf{r}$, normal $\mathbf{n} = (\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t})/\|\bar{\mathbf{v}} - (\bar{\mathbf{v}}\cdot\mathbf{t})\mathbf{t}\|_2$, and binormal $\mathbf{b} = \mathbf{t}\times\mathbf{n}$ be described by $\mathbf{P}_{[\mathbf{x},\mathbf{y}]} = \mathbf{x}\otimes\mathbf{y}$. In particular, we abbreviate $\mathbf{P_x} := \mathbf{P}_{[\mathbf{x},\mathbf{x}]}$ and $\mathbf{P}_{\mathbf{x},\mathbf{y}} := \mathbf{P_x} + \mathbf{P_y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$. Then, the operator $\nabla_{\mathbf{v}}\mathbf{f}$ resulting from (3.6)–(3.9) reads

$$\begin{aligned} \nabla_{\mathbf{v}}\mathbf{f}(\bar{\mathbf{v}}, \mathbf{t}) = {} & (a_{\mathrm{n}_1}\|\bar{\mathbf{v}}_{\mathbf{n}}\|_2 + 2a_{\mathrm{n}_2}\|\bar{\mathbf{v}}_{\mathbf{n}}\|_2^{1/2})\ \mathbf{P}_{\mathbf{n},\mathbf{b}} + \ (a_{\mathrm{n}_1}\|\bar{\mathbf{v}}_{\mathbf{n}}\|_2 + a_{\mathrm{n}_2}\|\bar{\mathbf{v}}_{\mathbf{n}}\|_2^{1/2})\ \mathbf{P_n} \\ & + 2a_{\mathrm{t}}\|\bar{\mathbf{v}}_{\mathbf{n}}^{\circ}\|_2^{1/2}\ \mathbf{P_t} + \ a_{\mathrm{t}}\|\bar{\mathbf{v}}_{\mathbf{n}}^{\circ}\|_2^{-1/2}(\bar{\mathbf{v}}\cdot\mathbf{t})\,\mathbf{P}_{[\mathbf{t},\,\bar{\mathbf{v}}_{\mathbf{n}}^{\circ}\|\bar{\mathbf{v}}_{\mathbf{n}}^{\circ}\|^{-1}]} \\ & + \chi(\bar{\mathbf{v}}, \mathbf{t})\,a_{\mathrm{t}}\|\bar{\mathbf{v}}_{\mathbf{n}}^{\circ}\|_2^{-3/2}\,c^\circ(c^\circ - \|\bar{\mathbf{v}}_{\mathbf{t}}\|_2\|\bar{\mathbf{v}}\|_2^{-1})(\|\bar{\mathbf{v}}_{\mathbf{t}}\|_2^2\ \mathbf{P_t} + (\bar{\mathbf{v}}\cdot\mathbf{t})\|\bar{\mathbf{v}}_{\mathbf{n}}\|_2\ \mathbf{P}_{[\mathbf{t},\mathbf{n}]}). \end{aligned}$$

Although the use of the indicator function $\chi(\bar{\mathbf{v}},\mathbf{t}) = 1$ for $(\|\bar{\mathbf{v}}_{\mathbf{t}}\|_2\|\bar{\mathbf{v}}\|_2^{-1}) \geq c^\circ$ and $\chi(\bar{\mathbf{v}},\mathbf{t}) = 0$ else, the introduction of $\mathbf{v}_{\mathbf{n}}^\circ$ in (3.8) yields a continuous Gâteaux derivative. In the limit to $\mathbf{t}\|\bar{\mathbf{v}}$, additionally it stays bounded, which is a big difference from an ansatz based on Taylor's original zero drag model with missing tangential component.

The Gâteaux derivative $\nabla_{\mathbf{v}}\mathbf{f}(\bar{\mathbf{v}},\mathbf{t})\,\mathbf{u}'$ is a good representative for the stochastic part in (3.10) if the mean relative velocity is much higher than the fluctuations that are characterized by the turbulent kinetic energy k, i.e., $\|\bar{\mathbf{v}}\|_2^2 \gg \mathbb{E}[\mathbf{u}'^2] = 2k$. In contrast, in case of $\bar{\mathbf{v}} = \mathbf{0}$ it would provide a zero drag since

$$\mathbf{f}(\bar{\mathbf{v}}+\mathbf{u}',\mathbf{t})|_{\bar{\mathbf{v}}=\mathbf{0}} = \mathbf{f}(\mathbf{0},\mathbf{t}) + \nabla_{\mathbf{v}}\mathbf{f}(\mathbf{0},\mathbf{t})\,\mathbf{u}' + \mathcal{O}((\mathbf{u}')^2) = \mathbf{0} + \mathcal{O}((\mathbf{u}')^2),$$

which is absurd, as the velocity fluctuations affect the fiber though vanishing mean relative velocity

$$\mathbf{f}(\mathbf{u}',\mathbf{t}) = (a_{\mathrm{n}_1}\|\mathbf{u}'_{\mathbf{n}}\|_2 + 2a_{\mathrm{n}_2}\|\mathbf{u}'_{\mathbf{n}}\|_2^{1/2})\;\mathbf{u}'_{\mathbf{n}} + 2a_{\mathrm{t}}\|\mathbf{u}'_{\mathbf{n}}{}^{\circ}\|_2^{1/2}\;\mathbf{u}'_{\mathbf{t}}. \tag{3.14}$$

Note that in (3.14) the direction $\mathbf{n}$ is exceptionally determined by $\mathbf{u}'$, i.e., $\mathbf{u}'_{\mathbf{n}} = \mathbf{u}' - \mathbf{u}'_{\mathbf{t}}$. The fact that the expectations of drag and velocity fluctuations are equal, i.e., $\mathbb{E}[\mathbf{f}(\mathbf{u}',\mathbf{t})] = \mathbb{E}[\mathbf{u}'] = \mathbf{0}$, motivates the stated extension of the linearized approach for $\bar{\mathbf{v}} = \mathbf{0}$. Keeping the directional vectors $\mathbf{u}'_{\mathbf{n}}, \mathbf{u}'_{\mathbf{t}}$, the coefficients with the specific norms are replaced by the respective averaged quantity expressed by the kinetic energy k such that the variance is correctly reproduced. Therefore abbreviate $\mathbf{f} := \mathbf{f}(\mathbf{u}',\mathbf{t})$ and consider

$$\begin{aligned}\mathbb{E}[\mathbf{f}\otimes\mathbf{f}] = {} & \mathbb{E}[(\mathbf{f}\cdot\mathbf{t})^2]\,\mathbf{t}\otimes\mathbf{t} + \mathbb{E}[(\mathbf{f}\cdot\mathbf{n_1})^2]\,\mathbf{n_1}\otimes\mathbf{n_1} + \mathbb{E}[(\mathbf{f}\cdot\mathbf{n_2})^2]\,\mathbf{n_2}\otimes\mathbf{n_2}\\ & + \mathbb{E}[(\mathbf{f}\cdot\mathbf{t})\;\;(\mathbf{f}\cdot\mathbf{n_1})]\;\;(\mathbf{t}\otimes\mathbf{n_1} + \mathbf{n_1}\otimes\mathbf{t})\\ & + \mathbb{E}[(\mathbf{f}\cdot\mathbf{t})\;\;(\mathbf{f}\cdot\mathbf{n_2})]\;\;(\mathbf{t}\otimes\mathbf{n_2} + \mathbf{n_2}\otimes\mathbf{t})\\ & + \mathbb{E}[(\mathbf{f}\cdot\mathbf{n_1})\,(\mathbf{f}\cdot\mathbf{n_2})]\;\;(\mathbf{n_1}\otimes\mathbf{n_2} + \mathbf{n_2}\otimes\mathbf{n_1})\end{aligned}$$

with arbitrarily chosen orthogonal normal vectors $\mathbf{n_1}, \mathbf{n_2}$. The mixed expectations vanish thereby due to the independence and odd appearance of the underlying velocity components, as for $\mathbb{E}[\mathbf{f}]$ above. Because of the identical distribution of the drag in the normal plane, we have $\mathbb{E}[(\mathbf{f}\cdot\mathbf{n_1})^2] = \mathbb{E}[(\mathbf{f}\cdot\mathbf{n_2})^2]$ such that it is sufficient to consider $\mathbb{E}[(\mathbf{f}\cdot\mathbf{n})^2]$. Using $\mathbb{E}[\mathbf{u}'^2] = 2k$ and the identical distribution of the velocity components yields their variance $\mathbb{E}[(\mathbf{u}'\cdot\mathbf{e})^2] = \sigma^2 = 2k/3$ with unit vector $\mathbf{e}$. The general (centered) moments are prescribed by the gamma function according to $\mathbb{E}[|\mathbf{u}'\cdot\mathbf{e}|^{2m}] = (2\pi\sigma^2)^{-1/2}\int x^{2m}e^{-x^2/(2\sigma^2)}\,dx = (2\sigma^2)^m\,\mathrm{gam}(m+1/2)/\sqrt{\pi}$, $m\in\mathbb{R}^+$. Then

$$\begin{aligned}\mathbb{E}[(\mathbf{f}\cdot\mathbf{t})^2] &= 4a_{\mathrm{t}}^2\,\mathbb{E}[|\mathbf{u}'\cdot\mathbf{n}^\circ|]\,\mathbb{E}[(\mathbf{u}'\cdot\mathbf{t})^2] = (a_{\mathrm{t}_0}(k)\,\sigma)^2,\\ \mathbb{E}[(\mathbf{f}\cdot\mathbf{n})^2] &= a_{\mathrm{n}_1}^2\mathbb{E}[(\mathbf{u}'\cdot\mathbf{n})^4] + 4a_{\mathrm{n}_1}a_{\mathrm{n}_2}\mathbb{E}[|\mathbf{u}'\cdot\mathbf{n}|^{7/2}] + 4a_{\mathrm{n}_2}^2\mathbb{E}[|\mathbf{u}'\cdot\mathbf{n}|^3] = (a_{\mathrm{n}_0}(k)\,\sigma)^2\end{aligned}$$

by means of (3.12), (3.13) such that

$$\mathbf{f}_0(\mathbf{u}',\mathbf{t},k) := a_{\mathrm{n}_0}(k)\,\mathbf{u}'_{\mathbf{n}} + a_{\mathrm{t}_0}(k)\,\mathbf{u}'_{\mathbf{t}} = a_{\mathrm{n}_0}(k)\,(\mathbf{u}'-\mathbf{u}'_{\mathbf{t}}) + a_{\mathrm{t}_0}(k)\,\mathbf{u}'_{\mathbf{t}}$$

describes a Gaussian random variable that has the same stochastic parameters, i.e., expectation and variance, as the original drag of (3.14). Moreover, it is linear in $\mathbf{u}'$, although we suggest that its coefficients depend on k. But the turbulent kinetic energy has to be viewed as an input parameter for the generation of the flow fluctuations in the context of this work. Hence, $\mathbf{L}^{\mathbf{f}}(\bar{\mathbf{v}},\mathbf{t},k) = a_{\mathrm{n}_0}(k)\,(\mathbf{I}-\mathbf{P_t}) + a_{\mathrm{t}_0}(k)\;\mathbf{P_t}$ is taken as the drag operator in the case $\bar{\mathbf{v}} = \mathbf{0}$.

For the secant complement that combines the two determined drag operators, all functional dependencies of ϖ might be imaginable, e.g., squared, linear, or quadratic in $\|\bar{\mathbf{v}}\|_2$. But because of the lack of information about this intermediate domain, i.e., $\|\bar{\mathbf{v}}\|_2^2 \in (0, 2k)$, they are mathematically and physically as less motivated as our proposed linear ansatz in (3.11).

3.3. Technical modification of force amplitude. Since the defined drag operator $\mathbf{L}^{\mathbf{f}}$ has a finite, nonvanishing limit for $\bar{v}_{\mathrm{n}} \to 0$, it is unable to balance the arising singularity of the force amplitude in (1.7),

$$\mathbf{D} = \left(\frac{2\pi}{\bar{v}_{\mathrm{n}}}\int_0^\infty \frac{E(\kappa)}{\kappa^2}d\kappa\right)^{1/2} \mathbf{P_{t,n}} \overset{(2.4)}{\approx} \left(\frac{2\pi F_2}{\bar{v}_{\mathrm{n}}}\right)^{1/2}\frac{k^2}{\epsilon}\;\mathbf{P_{t,n}}.$$

Consequently, the uncorrelated aerodynamic force $\mathbf{f}_{uc}^{air}$ diverges in the case of the linear dependence of $\mathbf{t}$ and $\bar{\mathbf{v}}$, whereas the correlated force $\mathbf{f}_{cc}^{air}$ stays bounded, as we have already seen in the similarity estimates (2.6), (2.7). Although the occurrence of this single discrepancy is negligibly small, the further numerical realization requires its handling. Thus, we suggest a slight technical modification of the amplitude that has no influence on the proved approximation quality of the uncorrelated force. Replace $\mathbf{D}$ by

(3.15)
$$\breve{\mathbf{D}} = (2\pi F_2)^{1/2}\,\frac{k^2}{\epsilon}\begin{cases} \bar{v}_{\mathrm{n}}^{-1/2}\;\mathbf{P_{t,n}}, & \omega > 1, \\[2ex] (1-\omega)\,(\bar{v}_{\mathrm{n}}^{crit})^{-1/2}\;(\mathbf{P_t} + (\mathbf{I}-\mathbf{P_t})/2) + \omega\,\bar{v}_{\mathrm{n}}^{-1/2}\;\mathbf{P_{t,n}}, & \omega \le 1, \end{cases}$$

with $\omega = \bar{v}_{\mathrm{n}}/\bar{v}_{\mathrm{n}}^{crit}$; then

$$\lim_{\bar{v}_{\mathrm{n}}\to 0}\mathbf{f}_{uc}^{air} = \mathbf{f} + \left(\frac{2\pi F_2}{\bar{v}_{\mathrm{n}}^{crit}}\right)^{1/2}\frac{k^2}{\epsilon}\begin{cases} l^{||}\,\mathbf{P_t}\,\mathbf{p}, & \varpi > 1, \\[2ex] ((1-\varpi)\,(a_{\mathrm{n}_0}(k)/2\;(\mathbf{I}-\mathbf{P_t}) + \;a_{\mathrm{t}_0}(k)\;\mathbf{P_t}) \\ \qquad +\varpi\, l^{||}\,\mathbf{P_t})\;\mathbf{p}, & \varpi \le 1, \end{cases}$$

coincides with the limit of the correlated force regarding the formal structure of the terms. Here, the deterministic force part given by the modified Taylor drag of (3.6) and (3.9) reads $\mathbf{f} = \mathbf{f_t}$ for $\|\bar{\mathbf{v}}\|_2 \neq 0$ and $\mathbf{f} = \mathbf{0}$ else, and furthermore

$$l^{||} := a_{\mathrm{t}}\left(2\bar{v}_{\mathrm{n}}^{\circ\,1/2} + \frac{\bar{v}_{\mathrm{t}}}{\bar{v}_{\mathrm{n}}^{\circ\,1/2}} + \frac{(c^{\circ\,2}-c^{\circ})\bar{v}_{\mathrm{t}}^2}{\bar{v}_{\mathrm{n}}^{\circ\,3/2}}\right)$$

with $\bar{v}_{\mathrm{n}}^{\circ} = (1-\mathrm{sgn}(\bar{\mathbf{v}}\cdot\mathbf{t})c^{\circ})\|\bar{\mathbf{v}}\|_2 < \infty$. The modification in (3.15) can be interpreted as cutting the amplitude $\mathbf{D}$ at the critical velocity $\bar{v}_{\mathrm{n}} = \bar{v}_{\mathrm{n}}^{crit}$ and matching it continuously with a linear extension. As the underlying $(\mathbf{t},\bar{\mathbf{v}})$-induced set $\{\mathbf{t},\mathbf{n},\mathbf{b}\}$ loses its basis properties in the limit $\bar{v}_{\mathrm{n}} = 0$, we distinguish between the tangential and the remaining projectors and introduce the normal independent splitting $(\mathbf{P_t}+(\mathbf{I}-\mathbf{P_t})/2)$ instead of the original $(\mathbf{P_t}+\mathbf{P_n})$. Thus, the direction of $\mathbf{f}_{uc}^{air}$ is no longer specified by the mean relative velocity for $\varpi \to 0$, as already indicated by $\mathbf{f}_{cc}^{air}$.

The needed technical modification of the amplitude reveals the deficiency of the modeled fluctuation velocity fields $\mathbf{w}_f^{\sigma,\tau}$ whose dynamics is based on a locally frozen turbulence pattern. Hence, the fiber experiences no temporal change of the correlations if it moves within the mean streamlines, i.e., $\bar{v}_\mathrm{n} = 0$. Alternatively to the modification, one might question the underlying concept of frozen turbulence that neglects the natural decay of vortices because of its large time t_T and slow turbulent velocity scale $u_\mathrm{T} = k^{1/2}$ in comparison to the advection scales of the mean flow t_A, $\bar{u}$. However, for a fiber suspended in turbulence, the actual temporal change of the experienced turbulent coherences is prescribed by the velocity $v_\mathrm{T}^f = \max\{\bar{v}_\mathrm{n}, u_\mathrm{T}\}$. This could be incorporated into the definition of the flow-dependent force amplitude $\check{\mathbf{D}}$ by substituting $\bar{v}_\mathrm{n}^{crit}$ with u_T. Then the characteristic turbulent fiber time reads $\tau_\mathrm{T}^f = \min\{l_\mathrm{T}/\bar{v}_\mathrm{n}, t_\mathrm{T}\}$. The consequences of the choice of the parameter $\bar{v}_\mathrm{n}^{crit}$ are illustrated in the numerical results of the next section.

4. Numerical simulations. The input flow data for the following numerical simulations of the fiber dynamics stem from k-ϵ computations of FLUENT 6.1 that have been adapted with user-specific procedures to reflect the realistic turbulent flow behavior of a melt-spinning process. The implementation of the fiber system (1.1), (1.2) is based on a standard method of lines. The use of spatial finite differences of higher order thereby yields the appropriate approximation of the algebraic constraint (1.2). The time integration is realized by a semi-implicit Euler method, where an adaptive time step control ensures stability and accuracy. The arising nonlinear system of equations is iteratively solved by a modified Newton–Raphson method. As the Jacobian matrices show a band structure, the computational effect of an iteration step is proportional to the number of fiber points. Note that the aerodynamic forces are explicitly included. Their quality depends crucially on the available flow data that are linearly interpolated on the spatial and temporal fiber grid.

In the following, we briefly present the numerical algorithms for the realization of the correlated and uncorrelated aerodynamic forces before we compare their effects on the fiber dynamics by means of an introduced curvature measure.

4.1. Algorithms. Let $I_m^n = \{l \in \mathbb{N}_0 \mid m \leq l \leq n\}$. Let the spatial and temporal fiber discretization be given by $s_i = i\Delta s$ and $t_j = t_{j-1} + \Delta t_{j-1}$, $t_0 = 0$ with fixed space increment Δs, and adaptive time step Δt_j, $(i,j) \in I_0^n \times I_0^m$. Then denote the respective function values at the fiber point s_i at time t_j with subscript $_i$ and superscript ^{j}, e.g., $\mathbf{r}_i^j = \mathbf{r}(s_i, t_j)$.

The numerical generation of the correlated aerodynamic force $\mathbf{f}_{cc}^{air}$ utilizes autoregressive moving average (ARMA) processes [2] for the centered, homogeneous, independent, local fluctuation velocity fields $\mathbf{w}_f^{\sigma,\tau}$ along the fiber, whereas the implementation of the uncorrelated force $\mathbf{f}_{uc}^{air}$ is exclusively based on Gaussian white noise $\mathbf{p}$,

$$\lim_{(\Delta s, \Delta t_j)\to\mathbf{0}} (\Delta s \Delta t_j)^{1/2}\, \mathbf{p}_i^j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

ALGORITHM 4.1 (computation of correlated force). *Choose l_T and t_T as characteristic turbulent large scales of the problem. Consider a fixed fiber and a time point that is indicated by the index tuple $(i,j) \in I_0^n \times I_0^m$.*

1. *Determine its corresponding index set N_i^j,*

$$N_i^j = \left\{ (\phi, \tau) \mid \left\| \mathbf{r}_i^j - \mathbf{r}_\phi^\tau - \bar{\mathbf{u}}_i^j \sum_{q=1}^{j-\tau} \Delta t_{j+1-q} \right\|_2 \leq l_\mathrm{T} \wedge \sum_{q=1}^{j-\tau} \Delta t_{j+1-q} \leq t_\mathrm{T} \right\}$$

with feasible tuples $(\phi, \tau) \in (I_0^n \times I_0^{j-1}) \cup (I_0^i \times I_j^j)$.

2. *Compute the centered, homogeneous, local fluctuations* $(\mathbf{w}^{\boldsymbol{\ell}}_f)^j_i$ *for all* $\boldsymbol{\ell} = (\ell_1, \ell_2) \in N^j_i$. *For this purpose, consider a fixed* $\boldsymbol{\ell}$*:*
   (a) *Set the turbulent fine-scale length* $\lambda^{\boldsymbol{\ell}}_{\mathrm{T}} = (20k^{\boldsymbol{\ell}}\nu/\epsilon^{\boldsymbol{\ell}})^{1/2}$.
   (b) *Determine the correlation index set* $(J^{\boldsymbol{\ell}})^j_i$,

$$(J^{\boldsymbol{\ell}})^j_i = \left\{ (\phi, \tau) \mid \left\| \mathbf{r}^j_i - \mathbf{r}^\tau_\phi - \bar{\mathbf{u}}^{\boldsymbol{\ell}} \sum_{q=1}^{j-\tau} \Delta t_{j+1-q} \right\|_2 \leq \lambda^{\boldsymbol{\ell}}_{\mathrm{T}} \right\}$$

with feasible tuples

$$(\phi, \tau) \in \begin{cases} I^{i-1}_{\ell_1} \times I^{\ell_2}_{\ell_2}, & \ell_2 = j,\ \ell_1 < i, \\ (I^n_{\ell_1} \times I^{\ell_2}_{\ell_2}) \cup (I^{i-1}_0 \times I^j_j), & \ell_2 = j-1, \\ (I^n_{\ell_1} \times I^{\ell_2}_{\ell_2}) \cup (I^n_0 \times I^{j-1}_{\ell_2+1}) \cup (I^{i-1}_0 \times I^j_j), & \ell_2 < j-1, \\ \emptyset & \textit{otherwise}. \end{cases}$$

   (c) *If* $(J^{\boldsymbol{\ell}})^j_i \neq \emptyset$, *then*:
      i. *Define a bijective mapping* $\rho : \{1, \ldots, |(J^{\boldsymbol{\ell}})^j_i|\} \to (J^{\boldsymbol{\ell}})^j_i$ *and set* $\rho(0) = (i, j)$.
      ii. *Consider the vectorial ARMA process*

$$(\mathbf{w}^{\boldsymbol{\ell}}_f)^j_i = (\mathbf{w}^{\boldsymbol{\ell}}_f)_{\rho(0)} = \sum_{q=1}^{|(J^{\boldsymbol{\ell}})^j_i|} \mathbf{A}_q (\mathbf{w}^{\boldsymbol{\ell}}_f)_{\rho(q)} + (\boldsymbol{\xi}^{\boldsymbol{\ell}})^j_i \tag{4.1}$$

with unknown coefficients $\mathbf{A}_q \in \mathbb{R}^{3\times3}$ *and noise* $(\boldsymbol{\xi}^{\boldsymbol{\ell}})^j_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ *that is assumed to be independent of* $(\mathbf{w}^{\boldsymbol{\ell}}_f)_{\rho(q)}$.
      iii. *Define* $\mathbf{C}_{(p,q)} := \mathbb{E}[(\mathbf{w}^{\boldsymbol{\ell}}_f)_{\rho(p)} \otimes (\mathbf{w}^{\boldsymbol{\ell}}_f)_{\rho(q)}]$ *for* $p, q = 0, \ldots, |(J^{\boldsymbol{\ell}})^j_i|$ *by means of the correlation tensor* $\boldsymbol{\gamma}^{\boldsymbol{\ell}}_0$ *in the canonical basis representation. Then particularly,* $\mathbf{C}_{(p,p)} = \boldsymbol{\gamma}^{\boldsymbol{\ell}}_0(\mathbf{0})$ *and* $\mathbf{C}_{(p,q)} = \mathbf{C}_{(q,p)}$ *hold.*
      iv. *Approximate the lateral correlation function of* $\boldsymbol{\gamma}^{\boldsymbol{\ell}}_0$ *by* $c^{\boldsymbol{\ell}}_1(z) = 2k^{\boldsymbol{\ell}}/3 - \epsilon^{\boldsymbol{\ell}} z^2/(30\nu)$, *i.e.,* $\boldsymbol{\gamma}^{\boldsymbol{\ell}}_0(\mathbf{z}) = (c_1(z) + z\partial_z c_1(z)/2)\mathbf{I} - \partial_z c_1(z)/(2z)\mathbf{z} \otimes \mathbf{z}$, $z = \|\mathbf{z}\|_2$ [13].
      v. *Compute the coefficients* $\mathbf{A}_q$ *by solving*

$$\sum_{q=1}^{|(J^{\boldsymbol{\ell}})^j_i|} \mathbf{C}_{(p,q)}\ \mathbf{A}_q = \mathbf{C}_{(p,0)}, \qquad p = 0, \ldots, |(J^{\boldsymbol{\ell}})^j_i| - 1. \tag{4.2}$$

      vi. *Calculate the covariance* $\mathbf{K}$ *of the noise term* $(\boldsymbol{\xi}^{\boldsymbol{\ell}})^j_i$ *from*

$$\begin{aligned} \mathbf{K} = {} & \mathbf{C}_{(0,0)} - \sum_{p=1}^{|(J^{\boldsymbol{\ell}})^j_i|} \mathbf{A}_p \mathbf{C}_{(p,p)} \mathbf{A}^T_p \\ & - \sum_{p=1}^{|(J^{\boldsymbol{\ell}})^j_i|-1} \sum_{q=p+1}^{|(J^{\boldsymbol{\ell}})^j_i|} \mathbf{A}_p \mathbf{C}_{(p,q)} \mathbf{A}^T_q - \sum_{p=1}^{|(J^{\boldsymbol{\ell}})^j_i|-1} \sum_{q=p+1}^{|(J^{\boldsymbol{\ell}})^j_i|} \mathbf{A}_q \mathbf{C}_{(q,p)} \mathbf{A}^T_p. \end{aligned}$$

vii. *Generate the correlated noise term* $(\boldsymbol{\xi}^{\boldsymbol{\ell}})_i^j = (\xi_1, \xi_2, \xi_3)$ *according to its covariance* $\mathbf{K} = (K_{pq})_{p,q=1,2,3}$ *and the ansatz*

$$\begin{aligned} &\xi_1 \sim \mathcal{N}(0, K_{11}), \\ &\xi_2 = \alpha\xi_1 + \xi_2', \\ &\xi_3 = \beta_1\xi_1 + \beta_2\xi_2 + \xi_3', \end{aligned} \tag{4.3}$$

where the parameters α, β_1, β_2 *and the independent random numbers* ξ_2', ξ_3' *are prescribed by*

$$\begin{aligned} &\alpha = K_{22}/K_{12} \ \textit{and} \ \sum_{q=1}^{2} K_{pq}\,\beta_q = K_{p3} \ \textit{for}\ p = 2, 3, \\ &\xi_2' \sim \mathcal{N}(0, K_{22} - \alpha^2 K_{11}), \\ &\xi_3' \sim \mathcal{N}(0, K_{33} - \beta_1^2 K_{11} - \beta_2^2 K_{22} - 2\beta_1\beta_2 K_{12}). \end{aligned}$$

viii. *Plug the determined coefficients* $\mathbf{A}_q$ *of* (4.2) *and the correlated noise* $(\boldsymbol{\xi}^{\boldsymbol{\ell}})_i^j$ *of* (4.3) *into the ARMA process* (4.1).

else, $(J^{\boldsymbol{\ell}})_i^j = \emptyset$*:*

Set

$$(\mathbf{w}_f^{\boldsymbol{\ell}})_i^j = \left(\frac{2k^{\boldsymbol{\ell}}}{3}\right)^{1/2} (\boldsymbol{\xi}^{\boldsymbol{\ell}})_i^j \qquad \textit{with}\ (\boldsymbol{\xi}^{\boldsymbol{\ell}})_i^j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4.4}$$

3. *Determine the correlated aerodynamic force*

$$(\mathbf{f}_{cc}^{air})_i^j = \mathbf{f}(\bar{\mathbf{v}}_i^j, \mathbf{t}_i^j) + \mathbf{L}^{air}(\bar{\mathbf{v}}_i^j, \mathbf{t}_i^j, k_i^j)\ |N_i^j|^{-1/2} \sum_{\boldsymbol{\ell}\in N_i^j} (\mathbf{w}_f^{\boldsymbol{\ell}})_i^j. \tag{4.5}$$

ALGORITHM 4.2 (computation of uncorrelated force). *Consider a fixed fiber and a time point that is indicated by the index tuple* $(i, j) \in I_0^n \times I_0^m$. *Set* $\varpi_i^j = \|\bar{\mathbf{v}}_i^j\|_2/(2k_i^j)^{1/2}$, $\omega_i^j = (\bar{v}_{\rm n})_i^j/\bar{v}_{\rm n}^{crit}$ *and let the projectors* $\mathbf{P}$ *depend on space and time discretization. Then, the uncorrelated aerodynamic force is determined by*

$$(\mathbf{f}_{uc}^{air})_i^j = \mathbf{f}(\bar{\mathbf{v}}_i^j, \mathbf{t}_i^j) + \sqrt{\frac{2\pi F_2}{\Delta s \Delta t_j}}\ \frac{(k_i^j)^2}{\epsilon_i^j \sqrt{(\bar{v}_{\rm n})_i^j}}\ \boldsymbol{\phi}_i^j, \tag{4.6}$$

where

$$\boldsymbol{\phi}_i^j = \begin{cases} \nabla_{\mathbf{v}} \mathbf{f}(\bar{\mathbf{v}}_i^j, \mathbf{t}_i^j)\ \mathbf{P}_{(\mathbf{t},\mathbf{n})_i^j}\ \boldsymbol{\xi}_i^j, & \varpi_i^j > 1, \omega_i^j > 1, \\ \nabla_{\mathbf{v}} \mathbf{f}(\bar{\mathbf{v}}_i^j, \mathbf{t}_i^j)\ \Big[(1-\omega_i^j)\sqrt{\omega_i^j}\ \big(\mathbf{P}_{\mathbf{t}_i^j} + (\mathbf{I} - \mathbf{P}_{\mathbf{t}_i^j})/2\big) + \omega_i^j\, \mathbf{P}_{(\mathbf{t},\mathbf{n})_i^j}\Big]\ \boldsymbol{\xi}_i^j, & \\ & \varpi_i^j > 1, \omega_i^j \le 1, \\ \big[(1-\varpi_i^j)\ \big(a_{\mathrm{t}_0}(k_i^j)\ \mathbf{P}_{\mathbf{t}_i^j} + a_{\mathrm{n}_0}(k_i^j)\ \mathbf{P}_{\mathbf{n}_i^j}\big) & \\ \quad + \varpi_i^j\ \nabla_{\mathbf{v}} \mathbf{f}((\varpi_i^j)^{-1}\, \bar{\mathbf{v}}_i^j, \mathbf{t}_i^j)\ \mathbf{P}_{(\mathbf{t},\mathbf{n})_i^j}\big]\ \boldsymbol{\xi}_i^j, & \varpi_i^j \le 1, \omega_i^j > 1, \\ \Big[\big[(1-\varpi_i^j)\, a_{\mathrm{t}_0}(k_i^j)\ \mathbf{I} + \varpi_i^j\ \nabla_{\mathbf{v}} \mathbf{f}((\varpi_i^j)^{-1}\, \bar{\mathbf{v}}_i^j, \mathbf{t}_i^j)\big](1-\omega_i^j)\sqrt{\omega_i^j}\ \mathbf{P}_{\mathbf{t}_i^j} & \\ \quad + (1-\varpi_i^j)\, a_{\mathrm{t}_0}(k_i^j)\, \omega_i^j\ \mathbf{P}_{\mathbf{t}_i^j} & \\ \quad + \big[(1-\varpi_i^j)\, a_{\mathrm{n}_0}(k_i^j)\ \mathbf{I} + \varpi_i^j\ \nabla_{\mathbf{v}} \mathbf{f}((\varpi_i^j)^{-1}\, \bar{\mathbf{v}}_i^j, \mathbf{t}_i^j)\big], & \\ \qquad \Big[(1-\omega_i^j)\sqrt{\omega_i^j}\ (\mathbf{I} - \mathbf{P}_{\mathbf{t}_i^j})/2 + \omega_i^j\ \mathbf{P}_{(\mathbf{t},\mathbf{n})_i^j}\Big] g\Big]\ \boldsymbol{\xi}_i^j, & \varpi_i^j \le 1, \omega_i^j \le 1, \end{cases}$$

and $\boldsymbol{\xi}_i^j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; *i.e., the components* $(\xi_l)_i^j \sim \mathcal{N}(0,1)$, $l = 1,2,3,$ *are independent and normally distributed.*

Regarding memory and computational effort, Algorithm 4.1 is extremely costly. Apart from the two searching procedures in steps 1 and 2(b), it requires in general the solving of $|N|$ linear systems of $3|J|$ equations for each fiber and time point specified by (i,j), step 2(c)v. Thereby, the cardinal numbers $|N|$ and $|J|$ depend not only on the fiber dynamics at (i,j), but also crucially on the spatial and temporal grid size, which should be chosen to be a compromise between computational capacity and desirable accuracy of the correlation structures to be realized. The required $3|N|$ Gaussian deviates for step 2(c)vii are here generated by the Box–Muller method [5]. In comparison to Algorithm 4.1, Algorithm 4.2 is obviously enormously cheaper and faster. Its evaluation is independent of the chosen discretization and needs only three Gaussian deviates per fiber and time point.

In case of large-scale resolution, where $N_i^j = \{(i,j)\}$ and $(J^{\boldsymbol{\ell}})_i^j = \emptyset$ for all (i,j) in Algorithm 4.1, the correlated aerodynamic force $\mathbf{f}_{cc}^{air}$ is obviously approximated numerically by the uncorrelated $\mathbf{f}_{uc}^{air}$, since (4.4), (4.5) correspond to the white noise approach of Algorithm 4.2 with $\Delta s \sim l_{\mathrm{T}}$ and $\Delta t \sim t_{\mathrm{T}}$ in (4.6). But, also for fine-scale resolution, the respective numerical representatives match very well as far as their effects on the fiber dynamics are concerned. To show the statistical coincidence of their influence, we analyze the imposed fiber dynamics by means of a curvature measure in the following. Thereby, we restrict the comparison exemplarily on a fixed appropriate fiber discretization because of the extremely long run-time and the enormous memory demands of Algorithm 4.1.

4.2. Results. Simulating the motion of an inextensible slender fiber swinging freely in a turbulent flow field, we show the similarity of the macroscopic effects on the fiber that are caused by the correlated and uncorrelated force model. For this purpose, we introduce the following curvature measure.

DEFINITION 4.1 (curvature measure). *Let* $\mathbf{r}_i^j$, $(i,j) \in \mathcal{I}_0^n \times \mathcal{I}_0^m$, *be the spatially and temporally discretized fiber line that is imposed by the aerodynamic forces according*

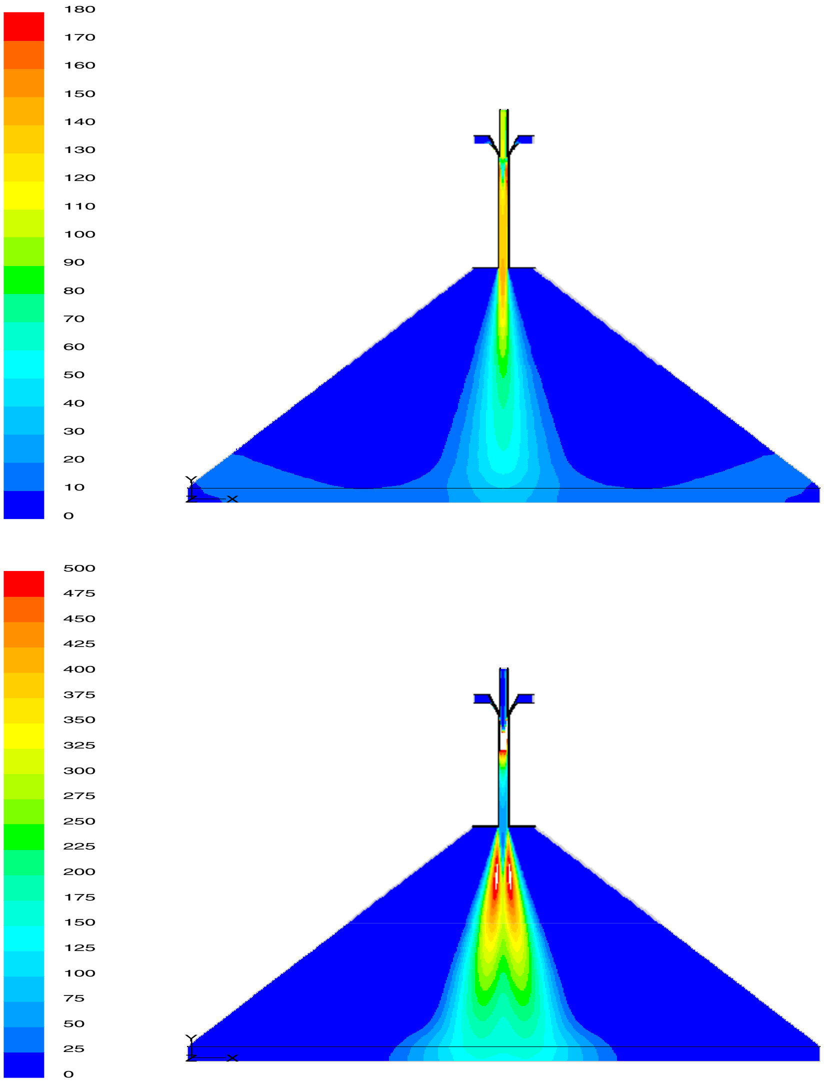


FIG. 4. *k-ϵ simulation results for turbulent flow. Top to bottom: Stationary two-dimensional vertical mean stream $\|\bar{\mathbf{u}}\|_2$, kinetic energy k in SI-units.*

to (1.1), (1.2). *Then, its* curvature measure *at time t_j is defined by*

$$\mathcal{K}^j = \frac{1}{n-1} \sum_{i=1}^{n-1} \|\Delta_{ss} \mathbf{r}_i^j\|_2$$

using the central difference $\Delta_{ss}\mathbf{r}_i^j = (\mathbf{r}_{i+1}^j - 2\mathbf{r}_i^j + \mathbf{r}_{i-1}^j)/\Delta s^2$.

Evaluating the fiber line over a certain time interval gives statistically comparable parameters for $\mathcal{K}$, i.e., its mean μ and its standard deviation σ.

Apart from the similarity, the curvature measure states the significance of the turbulent aerodynamic force for entanglement and loop-forming of the fiber. To illustrate

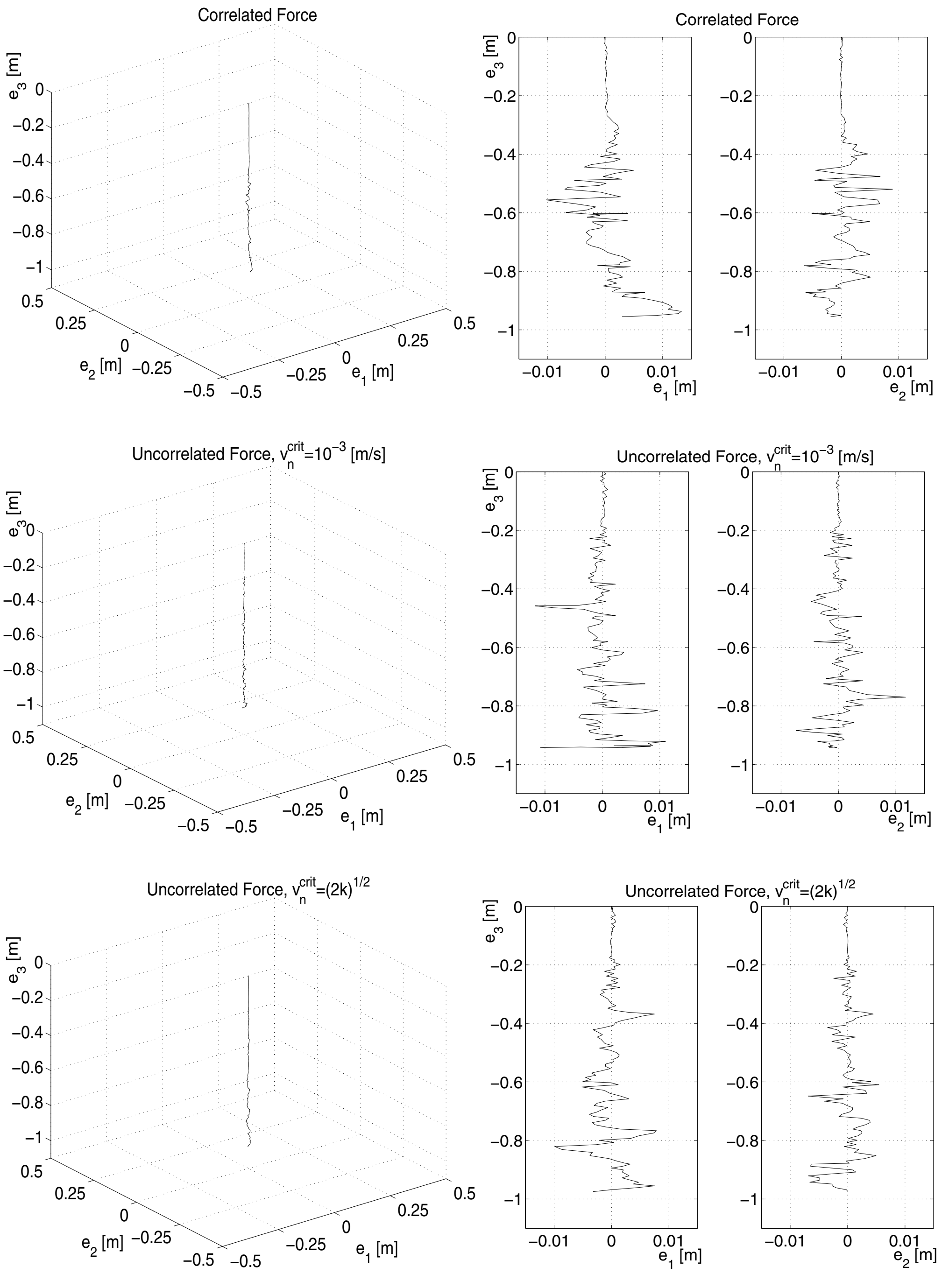


FIG. 5. *From top to bottom: Fiber exposed to* $\mathbf{f}_{cc}^{air}$ *and* $\mathbf{f}_{uc}^{air}$ *with* $\bar{v}_{\mathrm{n}}^{crit} = 10^{-3}$ *m/s, resp.,* $\bar{v}_{\mathrm{n}}^{crit} = (2k)^{1/2}$*. Left: Instantaneous fiber dynamics. Right: Two-dimensional projections zoomed in.*

these effects, we consider a fiber of length $l = 1$ m and material properties according to Table 1 that is initially hanging in the symmetry axis of a stationary, vertically directed two-dimensional mean flow field $\bar{\mathbf{u}}$ (cf. Figure 4). The turbulent fluctuations

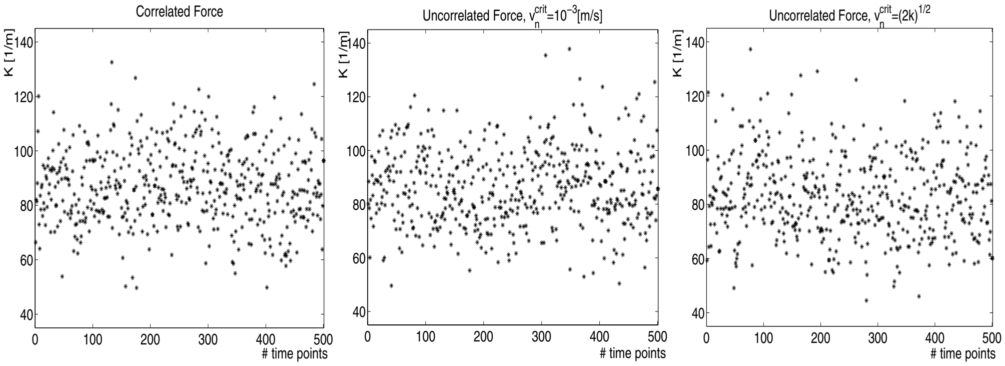


Fig. 6. *Curvature measures $\mathcal{K}$ over 500 time points for fiber exposed to stochastic forces for $5 \cdot 10^{-2}$ s. From left to right: Results for $\mathbf{f}_{cc}^{air}$, $\mathbf{f}_{uc}^{air}$ with $\bar{v}_{\mathrm{n}}^{crit} = 10^{-3}$ m/s, resp., $\bar{v}_{\mathrm{n}}^{crit} = (2k)^{1/2}$.*

are prescribed by the stationary kinetic energy k and dissipation rate ϵ. Then, the resulting deterministic force part $\overline{\mathbf{f}}$ is mainly vertically directed and the stochastic part $\mathbf{f}'$ is determined almost exclusively by the small horizontal fiber oscillations. Hence, if the turbulent influence is neglected, the fiber is not excited out of its position of rest. It has the characteristic curvature properties $\mu = 0$ and $\sigma = 0$ which will prescribe our reference state. The used underlying flow data represent a realistic turbulent stream, as might be expected in the deposition region of a melt-spinning process; see the parameter values in Tables 1 and 2. Note that the illustrated geometry in Figure 4 is distorted in width to stress the flow behavior around the symmetry axis, $\mathbf{e_3}$-axis.

Exposing the fiber to the stochastic force models, we obtain the representatives of a momentary fiber position that are visualized in Figure 5. Apart from the correlated force, we distinguish hereby between the uncorrelated force effects by choosing two variants for $\bar{v}_{\mathrm{n}}^{crit}$, i.e., $\bar{v}_{\mathrm{n}}^{crit} = 10^{-3}$ m/s and $\bar{v}_{\mathrm{n}}^{crit} = (2k)^{1/2}$. At first glance, the behavior of the fibers seems to be straightforward and meaningless due to the chosen draw ratio of meters. But, indeed, all three representatives show similar curvatures, which becomes evident by zooming into the two-dimensional fiber projections; see Figure 5 (right). Near the mounting, they hang down almost straight for the first $2 \cdot 10^{-1}$ m before they start to form loops. The observed oscillations then have a typical range of 10^{-3} up to 10^{-2} m, which corresponds with our asymptotic analysis of section 2.2; see Figure 2. Considering the respective fiber motions for a period of $5 \cdot 10^{-2}$ s, we provide further results by the curvature measures $\mathcal{K}$ that are plotted and statistically evaluated for comparable samples of 500 time points; see Figure 6 and Table 3. Thereby, all temporal evolutions turn out to be normally distributed. The mean curvature measure of the uncorrelated force, $\bar{v}_{\mathrm{n}}^{crit} = 10^{-3}$ m/s, differs less than 1% from that of the correlated force. Also, the standard deviations fit very well, and we obtain differences of only 2%. This shows very good agreement. This choice of $\bar{v}_{\mathrm{n}}^{crit}$ overcomes simply the singularity stemming from the underlying correlated frozen turbulence pattern and therefore yields better approximation properties than the other variant that additionally incorporates the decay of the vortices.

Summing up, the uncorrelated force model is undeniably a good substitute for the correlated one on the macroscopic fiber scale. Leading to a statistically similar fiber behavior, it requires—instead of days—only a few minutes of computational time for the simulation of $5 \cdot 10^{-2}$ s real time motion. Thus, it makes long-time fiber studies possible, which is essential for practical application. Note that for the computation

Table 3
Statistic parameters for the curvature measures $\mathcal{K}$ of Figure 6.

Stochastic force	Correlated	Uncorrelated		Without
		$\bar{v}_{\mathrm{n}}^{crit} = 10^{-3}$[m/s]	$\bar{v}_{\mathrm{n}}^{crit} = (2k)^{1/2}$	
$\mathcal{K}$ [1/m]				
μ	86.93 (100%)	86.33 (−0.69%)	82.99 (−4.53%)	0
σ	13.83 (100%)	14.10 (+2.00%)	14.94 (+8.03%)	0
CPU-time	Days	∼4.5 min	∼4 min	∼1.5 min

of the deterministic reference case, Algorithm 4.2 is not needed. Moreover, due to the absence of stochastic forces, a larger (adaptive) time step can be used. The increase of Δt by one order, up to $\Delta t \sim 10^{-5}$ s, together with the skipping of Algorithm 4.2, leads to the bisection of the CPU-time observed in Table 3. Thus, it takes only 1.5 min CPU-time instead of 4 min as in the turbulent cases. All calculations have been performed on an Intel Xeon processor, 2.8 GHz.

5. Conclusions. In [13], a general aerodynamic force concept is derived on the basis of a stochastic k-ϵ turbulence model for the flow field. The turbulence effects on the dynamics of a long slender elastic fiber are modeled by a correlated Gaussian force and in its asymptotic limit on a macroscopic fiber scale by Gaussian white noise with flow-dependent amplitude. Choosing a specific Taylor drag model, this paper has shown the applicability of the force concept for the handling of the complex fiber-turbulence interactions as they occur in a typical melt-spinning process of nonwoven materials. Moreover, it has stated the very good theoretical and numerical approximation properties of the uncorrelated force. The introduction of the uncorrelated aerodynamic force changes the character of the perturbation term into a localized linear integrator such that the fiber dynamics is described by a system of partial differential equations with additive Gaussian white noise. This enables not only a theoretical analysis but also an efficient numerical realization. Adapting the fiber system with appropriate boundary and initial conditions, the FIber DYnamics Simulation Tool (FIDYST) [8] developed at Fraunhofer ITWM, Kaiserslautern, applies the presented algorithm to simulate the turbulent deposition region of melt-spinning processes with hundreds of individual endless fibers. The simulation results are validated with experimental data. However, note that for this purpose, further aspects have to be taken into account, such as fiber-fiber interactions, sticky fiber bunches, conveyor belt effects, or the affection of the turbulence by higher concentrated fiber curtains.

REFERENCES

[1] G. K. Batchelor, *Slender-body theory for particles of arbitrary cross-section in Stokes flow*, J. Fluid Mech., 44 (1970), pp. 419–440.

[2] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Springer-Verlag, Berlin, 1987.

[3] C. Call and J. Kennedy, *Measurements and simulations of particle dispersion in a turbulent flow*, Internat. J. Multiphase Flow, 18 (1992), pp. 891–903.

[4] R. G. Cox, *The motion of long slender bodies in a viscous fluid. Part* 1. *General theory*, J. Fluid Mech., 44 (1970), pp. 791–810.

[5] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

[6] M. Doi and D. Chen, *Simulation of aggregating coilloids in shear flow*, J. Chem. Phys., 90 (1989), pp. 5271–5279.

[7] T. Götz and A. Unterreiter, *Analysis and numerics of an integral equation model for slender bodies in low-Reynolds-number flow*, J. Integral Equations Appl., 12 (2000), pp. 225–270.

[8] D. Hietel and R. Wegener, *Simulation of spinning and laydown processes*, Technical Textiles, 3 (2005), pp. 145–147.

[9] O. Hinze, *Turbulence*, 2nd ed., McGraw-Hill, New York, 1975.

[10] J. B. Keller and S. I. Rubinow, *Slender-body theory for slow viscous flow*, J. Fluid Mech., 75 (1976), pp. 705–714.

[11] M. E. M. Lee, *Mathematical Models of the Carding Process*, Ph.D. thesis, University of Oxford, Oxford, UK, 2001.

[12] N. Marheineke, *Turbulent Fibers—On the Motion of Long, Flexible Fibers in Turbulent Flows*, Ph.D. thesis, Technische Universität Kaiserslautern, Kaiserslautern, Germany, 2005.

[13] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: General modeling framework*, SIAM J. Appl. Math., 66 (2006), pp. 1703–1726.

[14] R. Mei, *Effect of turbulence on the particle settling velocity in the non-linear drag range*, Internat. J. Multiphase Flow, 20 (1994), pp. 273–284.

[15] J. A. Olsen and R. J. Kerekes, *The motion of fibres in turbulent flow*, J. Fluid Mech., 337 (1998), pp. 47–64.

[16] L. M. Pismen and A. Nir, *On the motion of suspended particles in stationary homogeneous turbulence*, J. Fluid Mech., 84 (1978), pp. 193–206.

[17] F. R. Russel and D. J. Klingenberg, *Dynamic simulation of flexible fibers composed of linked rigid bodies*, J. Chem. Phys., 105 (1997), pp. 2949–2960.

[18] T. H. Shih and J. L. Lumley, *Second-order modelling of particle dispersion in a turbulent flow*, J. Fluid Mech., 163 (1986), pp. 349–363.

[19] G. I. Taylor, *Analysis of the swimming of long and narrow animals*, Proc. Roy. Soc. London Ser. A, 214 (1952), pp. 158–183.

[20] B. Underwood, *Random-walk modeling of turbulent impaction to a smooth wall*, Internat. J. Multiphase Flow, 19 (1993), pp. 485–500.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 24–42

ANISOTROPY RECONSTRUCTION FROM WAVE FRONTS IN TRANSVERSELY ISOTROPIC ACOUSTIC MEDIA[*]

JOYCE R. MCLAUGHLIN[†], DANIEL RENZI[†], AND JEONG-ROCK YOON[‡]

Abstract. This paper considers an inverse problem for a transversely isotropic three-dimensional acoustic medium, where there is one preferred direction called the fiber direction along which the wave propagates fastest and there is no preferred wave propagation direction in the isotropic plane, which is the plane orthogonal to the fiber direction. In this medium the parameters to be recovered are (1) the wave speed for a wave propagating in the direction along the fiber; (2) the wave speed for a wave propagating in any direction which is orthogonal to the fiber direction; and (3) the unit fiber direction itself. So four scalar functions are to be recovered. The data are the positions of four distinct wave fronts as the corresponding waves propagate through the medium. The mathematical relation, which is the Eikonal equation, between the wave front locations and the four unknown functions, is nonlinear. Here it is established, perhaps surprisingly, that corresponding to the given data set, there can be up to four possible solution quadruples. We present and implement an algorithm to compute each of the possible solutions and show our selection criteria to obtain the correct solution. The Eikonal equation, which relates the wave front positions to the unknown functions, is the same equation obtained for the horizontally polarized shear wave (SH wave) which propagates in a linear elastic system.

Key words. elastography, inverse problem, arrival time, anisotropic wave equation, transversely isotropic medium, fiber direction

AMS subject classifications. 35R30, 62P10, 92C55

DOI. 10.1137/060651252

1. Introduction.

Motivated by wave propagation directional dependence in tissue, the goal of this paper is to identify directionally dependent stiffness properties from multiple wave fronts. The wave propagation model is an anisotropic wave equation, where the medium has one preferred direction, which we designate as the *fiber direction*, where it has a faster wave speed and the waves propagating in the plane orthogonal to this preferred direction are slower and exhibit no directional dependence. Our goal is the recovery of the unit fiber direction and the ratio of each of two distinct stiffness coefficients to the density. The square roots of these two ratios define the wave speed in the fiber direction and in the plane orthogonal to the fiber. We show that in three dimensions there can be up to four discrete solution triples of two wave speeds and the fiber direction, from four distinct wave fronts. The fact that there is a discrete set of solutions is a direct result of the nonlinear relations, governed by the Eikonal equation, between wave front directions, wave speeds, and the fiber direction.

Shear stiffness recovery has been of interest for about 15 years, and several experiments are being investigated as follows:

[*] Received by the editors January 30, 2006; accepted for publication (in revised form) May 24, 2007; published electronically October 10, 2007.

http://www.siam.org/journals/siap/68-1/65125.html

[†] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (mclauj@rpi.edu, renzid@rpi.edu). The first author was partially supported by NSF Focus Group grant DMS 0101458, ONR grant N00014-96-1-0349, ONR grant N00014-05-1-0600, NIBIB grant 1R21EB003000-01, and IPAM during Fall 2003. The work of the second author was partially supported by NSF VIGRE grant DMS 9983646 and IPAM during Fall 2003.

[‡] Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (jryoon@clemson.edu). Funding for the work of this author was provided by the NSF Focus Group grant DMS 0101458 with partial support from IPAM during Fall 2003, IMA during Fall 2005, and his start-up grant from Clemson University.

(1) tissue that is compressed as stiff tissue compresses less [4, 17];
(2) single frequency excitation, where stiff tissue exhibits low amplitude and stiffness characteristics can be recovered from amplitude variations [8, 15, 19, 20];
(3) crawling or holographic waves, which are produced with excitations at two nearby frequencies, and where phase wave speed can be recovered [14, 22];
(4) interior radiation force excitation at a single point produced by a single ultrasound beam [16];
(5) interior radiation force excitation produced with two ultrasound beams whose excitation frequency difference is in the KHz range [7];
(6) tissue surface line sources, or *supersonic imaging* that effectively produce line sources orthogonal to the tissue surface and produce propagating waves with identifiable fronts [2, 3]; the propagating front locations can be utilized to recover tissue properties.

In each of the above six cases the goal is to image either (a) shear wave speed which is roughly 3 m/sec in normal isotropic tissue and can more than double in abnormal tissue; or (b) shear stiffness which can increase more than four times in abnormal tissue. The goal is to identify abnormal inclusions, which are tumors.

Here we utilize the supersonic imaging experiment, in which a line source is approximated by a set of interior radiation force pushes, produced by focused ultrasound beams all at the same frequency, and made successively along a line. This effectively induces a conical wave in three dimensions whose angle with the line of the source is determined by how fast the succession of pushes is made and whether or not the pushes begin deep in the tissue and move successively toward the surface or vice versa; see Figure 1.1.

Our goal in this paper is to recover anisotropic tissue properties. Our motivations are (a) that some normal, e.g., muscle, tissue is anisotropic and so mathematical models must include this property; and (b) that it has been conjectured [18, 21] that benign and cancerous tumors may have their own distinguishing anisotropic properties. If indeed the latter conjecture is true, the recovery of anisotropic tumor properties could be of considerable medical importance.

To give some background about what is known in the isotropic case, for contrast with the anisotropic case, we recall that previously we established uniqueness

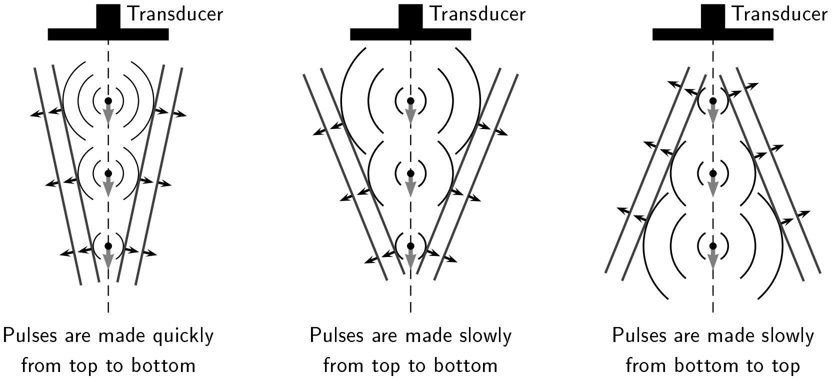


FIG. 1.1. *Illustration of three possible conical wave fronts (two-dimensional view) produced by a succession of interior radiation force pushes. A transducer focuses ultrasound beams to produce interior radiation force pushes and changes focal depth successively, either from top to bottom or vice versa.*

results [10, 11], and the *arrival time algorithm* [10, 12, 13], to reconstruct wave speed in isotropic media. There we showed that the positions of one propagating front established the wave speed uniquely, and that there is at most one pair of the shear stiffness μ and the density ρ, corresponding to a given single displacement data as a function of space and time, provided the medium is initially at rest. In this paper, we establish that four distinct wave fronts in three dimensions yield up to four triples: two distinguishing wave speeds and a fiber direction. We note also that if we are given one of the possible triples and the solution of the anisotropic wave equation (as opposed to only the wave front positions), then also there is at most one density ρ corresponding to that triple.

Our paper is organized as follows. In section 2 we establish that our model has finite propagation speed, that Lipschitz continuous fronts, defined by their arrival times, satisfy an anisotropic Eikonal equation, and we refer to our very recent result that establishes that arrival times are actually Lipschitz continuous; in section 3 we give our analysis that there can be up to four discrete solution triples corresponding to four distinct wave fronts; and in section 4 we show numerical results that includes recoveries of an anisotropic inclusion embedded in an isotropic background.

2. Anisotropic acoustic models. We consider anisotropic models, where the wave speed represented by $\sqrt{c_{44}/\rho}$ in one preferred direction, which we call the fiber direction, $\vec{f}$, is larger than the wave speed $\sqrt{c_{66}/\rho}$ in the plane orthogonal to the fiber direction. In this plane, which we call the isotropic plane, the wave speed is independent of direction. Our language and notation here are consistent with SH-wave propagation in incompressible transversely isotropic linear elastic models, which we will consider in a later paper.

Let Ω be a bounded $\mathcal{C}^2$ open connected subset in $\mathbb{R}^n$ for $n = 2, 3$. Assume

$$\begin{aligned} &\rho \in \mathcal{C}^0(\bar{\Omega}),\ M \in \left[\mathcal{C}^1(\bar{\Omega})\right]^{n\times n} \text{ is a symmetric matrix function, and} \\ &\exists \alpha_0 > 0 \text{ such that } \rho(x) \ge \alpha_0, \quad \vec{v}\cdot M(x)\vec{v} \ge \alpha_0|\vec{v}|^2 \quad \forall x \in \bar{\Omega},\ \forall \vec{v} \in \mathbb{R}^n. \end{aligned} \tag{2.1}$$

Then our anisotropic wave propagation model is

$$\nabla\cdot(M\nabla u) = \rho u_{tt} \quad \text{in } \Omega\times(0,T) \tag{2.2}$$

with homogeneous initial condition, $u(x,0) = u_t(x,0) = 0$ in Ω, and the boundary condition is either Dirichlet or Neumann; $u|_{\partial\Omega\times(0,T)} = g$ or $(\nu\cdot M\nabla u)|_{\partial\Omega\times(0,T)} = h$, where ν is the unit outward normal to $\partial\Omega$. This is an anisotropic extension of the frequently used isotropic elastography model; see [1]. We refer the reader to [6] for techniques to establish existence and uniqueness for the initial-boundary value problem associated with (2.2).

Remark 1. In terms of the SH-wave motivated assumptions mentioned above, the stiffness matrix M is represented by

$$M = c_{66}I + (c_{44} - c_{66})\vec{f}\otimes\vec{f}, \tag{2.3}$$

where $|\vec{f}| = 1$, $c_{44} > c_{66} > 0$ in $\bar{\Omega}$, and I and $\otimes$ denote the identity matrix and tensor product, respectively. Here our assumption that $c_{44} > c_{66}$ is natural since in biological tissue, e.g., muscle tissue, the wave speed is fastest in the direction aligned with the fibers [9].

Since our medium is initially at rest, the wave propagates into the medium from the boundary with a propagating front. In our next two theorems, following [10, 11], we establish that the wave whose propagation is governed by the above model has (1) finite propagation speed, and (2) an arrival time, which we assume to be Lipschitz continuous, that, under this assumption, satisfies the Eikonal equation.

THEOREM 2. *Assume ρ and M satisfy* (2.1). *Let $u \in H^2(\Omega \times (0,T))$ be a solution of* (2.2). *Then for any open ball $B_\epsilon(x_0) \subset \Omega$, u has a finite propagation speed in $B_\epsilon(x_0) \times (0,T)$ with the maximum speed*

$$c = \sup_{x \in B_\epsilon(x_0)} \sqrt{\sigma_M(x)/\rho(x)},$$

where $\sigma_M(x)$ is the largest eigenvalue of $M(x)$.

The proof of the above theorem is along the same lines as that in the isotropic case (Theorem 3.4 in [11]), once we redefine the energy by

$$e(s) := \frac{1}{2} \int_{C_s} \left\{ \rho |u_t|^2 + \nabla u \cdot M \nabla u \right\} dx, \quad C_s := B_{\epsilon - cs}(x_0) \times \{t = t_0 + s\}.$$

So we omit the proof.

As in [10] we define the *arrival time*, $\hat{T}(x)$, of the wave as

$$\hat{T}(x) := \inf\{t \in (0,T) : |u(x,t)| > 0\}, \quad x \in \Omega_{u \neq 0}, \tag{2.4}$$

where $\Omega_{u\neq 0} := \{x \in \Omega\ : u(x,t) \neq 0 \text{ for some } t \in (0,T)\}$, and we assume the solution u of (2.2) is continuous. If $\hat{T} \in \mathcal{C}^1(\Omega)$, then existing unique continuation results would apply to show that the arrival time, $\hat{T}$, satisfies the Eikonal equation given below. Since our target medium is inhomogeneous, we then expect waves originating at more than one point on the boundary to arrive simultaneously at the same interior points of Ω. In this case, $\hat{T}(x)$ could have *kinks* or at least be nondifferentiable there. Hence we assume $\hat{T}(x)$ is Lipschitz continuous and establish the following theorem.

THEOREM 3. *Assume $\rho \in \mathcal{C}^1(\bar{\Omega})$ in addition to* (2.1). *Let $u \in H^2(\Omega \times (0,T)) \cap \mathcal{C}^0(\Omega \times (0,T))$ be a solution of* (2.2) *with $u(x,0) = u_t(x,0) = 0$ in Ω, and either of the following Dirichlet or Neumann boundary conditions: $u|_{\partial\Omega \times (0,T)} = g$ or $(\nu \cdot M\nabla u)|_{\partial\Omega\times(0,T)} = h$. Suppose further that the arrival time $\hat{T} : \Omega_{u\neq 0} \to [0,T]$ is Lipschitz continuous. Then $\hat{T}$ satisfies the Eikonal equation*

$$\rho = \nabla \hat{T} \cdot M \nabla \hat{T} \quad a.e.\ in\ \Omega_{u\neq 0}. \tag{2.5}$$

In particular, when M is given in the form of (2.3), *our Eikonal equation becomes*

$$\frac{1}{|\nabla \hat{T}|^2} = \frac{c_{66}}{\rho} + \left(\frac{c_{44}}{\rho} - \frac{c_{66}}{\rho}\right) \left| \frac{\nabla \hat{T}}{|\nabla \hat{T}|} \cdot \vec{f} \right|^2. \tag{2.6}$$

Proof. Since $\hat{T}$ is Lipschitz continuous, $\nabla \hat{T}$ is well defined almost everywhere. Note that (2.5) is merely a necessary condition for $t = \hat{T}(x)$ to be a characteristic surface with respect to the hyperbolic equation $\rho u_{tt} = \nabla \cdot (M \nabla u)$. If we suppose that $t = \hat{T}(x)$ is a noncharacteristic surface, we can draw a contradiction, as done in Theorem 2.10 in [10], which is based on Theorem 3.6 in [5] and a lemma on page 544 of [6]. See [10] for the details. □

Remark 4. In fact, $\hat{T}$ according to the definition (2.4) may be discontinuous even if the solution u is infinitely smooth. However, in this paper we adopt this definition to make arguments simpler and clearer. Modifying the definition of arrival time by

$$\hat{T}(x) := \inf\{t \in (0,T] : ||u||_{L^2(V\times(0,t))} > 0\ \forall \text{ open } V \subset \Omega \text{ with } x \in V\}, \quad x \in \Omega \setminus \Omega_E,$$

where $\Omega_E := \bigcup\{V \subset \Omega$ is an open set satisfying $||u||_{L^2(V\times(0,T))} = 0\}$, we have recently established that $\hat{T} : \Omega \setminus \Omega_E \to (0, T]$ is actually Lipschitz continuous. This result will be addressed soon.

Note that in the anisotropic case, the wave does not always propagate in the direction orthogonal to the wave front (*group* or *ray* velocity is not always the same as *phase* velocity). Nevertheless, under the assumption that $\hat{T}(x)$ is Lipschitz continuous, the phase wave speed, $c(x)$, in the direction orthogonal to the front, satisfies

$$c(x)|\nabla\hat{T}| = 1, \qquad c(x) = c\left(x, \nabla\hat{T}\right), \tag{2.7}$$

and can be determined by the methods given in [10, 12, 13]. In later sections, we will assume that this speed, $c(x)$, has been determined from $\hat{T}$ so when we solve the inverse problem,

$$\text{find } (c_{66}/\rho, c_{44}/\rho, \vec{f}) \text{ from multiple arrival times,}$$

we will assume we know both $\hat{T}(x)$ and $c(x)$.

Remark 5. In a later paper we will consider a transversely isotropic elastic medium. Note that then (2.6) will be the Eikonal equation with M defined as in (2.3), satisfied by the SH-wave phase $\psi(x)$ in a geometric optics expansion, $\vec{u} = \vec{a}e^{i\omega(t-\psi(x))}$, where $\vec{a} = \vec{a}_0 + \frac{1}{i\omega}\vec{a}_1 + \frac{1}{(i\omega)^2}\vec{a}_2 + \cdots$ is an asymptotic series with $\omega \gg 1$.

3. Reconstruction using four measurements.

Having established the intrinsically nonlinear Eikonal equations (2.6) and (2.7) in section 2, we address the utilization of these equations to recover the three unknown quantities $(c_{66}/\rho, c_{44}/\rho, \vec{f})$ from wave fronts $\hat{T}$. Since $|\vec{f}| = 1$, this means that in three dimensions we have four scalar functions to recover. It is natural then to investigate the inverse problem,

$$\text{find } \left(\frac{c_{66}}{\rho}, \frac{c_{44}}{\rho}, \vec{f}\right) \text{ from four distinct wave fronts } \{\hat{T}_j\}_{j=1}^4. \tag{3.1}$$

Perhaps surprisingly, our analysis establishes that we can have a finite discrete (up to four) set of triples that correspond to given four distinct propagating wave fronts. We make this statement more precise below.

Let $\{\hat{T}_j\}_{j=1}^4$ be four given arrival time data. Define the unit wave normal and the corresponding phase wave speed by $\vec{n}_j := \nabla\hat{T}_j/|\nabla\hat{T}_j|$ and $c_j := 1/|\nabla T_j|$, respectively. Recall c_j can be estimated by solving (2.7) based on the methods given in [10, 12, 13]. Then the Eikonal equation (2.6) becomes

$$c_j^2 = \tilde{c}_{66} + (\tilde{c}_{44} - \tilde{c}_{66})|\vec{f}\cdot\vec{n}_j|^2, \quad j = 1, 2, 3, 4, \tag{3.2}$$

where we define $\tilde{c}_{66} := c_{66}/\rho$ and $\tilde{c}_{44} := c_{44}/\rho$ for convenience. As described in section 2, we are assuming $\tilde{c}_{44} > \tilde{c}_{66}$, which is a reasonable assumption, as the fiber in biological tissue is normally stiffer than the background matrix. Thus $\tilde{c}_{44}$ and $\tilde{c}_{66}$ are the upper and lower bounds of all possible c_j^2, respectively. So we can define $d_j := \sqrt{c_j^2 - \tilde{c}_{66}} \geq 0$ and $\vec{g} := \sqrt{\tilde{c}_{44} - \tilde{c}_{66}}\vec{f} \neq 0$, and from (3.2) we establish linear relations for $\vec{g}$,

$$\vec{g}\cdot\vec{n}_j = \pm d_j, \quad j = 1, 2, 3, 4. \tag{3.3}$$

Then our task is to determine $(\tilde{c}_{66}, \vec{g})$ from the data $\{(\vec{n}_j, c_j)\}_{j=1}^4$. Once we determine $\tilde{c}_{66}$, knowing $\vec{g}$ is equivalent to knowing $\tilde{c}_{44}$ and $\vec{f}$, since $\tilde{c}_{44} = \tilde{c}_{66} + |\vec{g}|^2$ and $\vec{f} = \vec{g}/|\vec{g}|$.

In this section, we will show that $\tilde{c}_{66}$ is a root of a fourth order polynomial $p(x)$ (Theorem 13), and hence we may have four possible $\tilde{c}_{66}$. For each $\tilde{c}_{66}$, we have a *generic* uniqueness to determine $\vec{g}$ (Corollary 20) and an explicit formula for $\vec{g}$ (Theorem 14). So we will have at most four possible solutions $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$. Since $\tilde{c}_{66}$ can be a multiple root of $p(x)$, despite the generic uniqueness, it may look like we have multiple $\tilde{c}_{44}$ and $\vec{f}$ corresponding to a single $\tilde{c}_{66}$ (Theorems 15 and 17). However, to realize this special case, the data $\{(\vec{n}_j, c_j)\}_{j=1}^4$ must satisfy one of a very special set of conditions (3.10)–(3.12) that are unlikely to occur in the actual experiments.

3.1. Coordinate system and data preparation. For convenience, we assume we have a *well prepared* data set, defined below, and fix an appropriate coordinate system, defined as follows.

DEFINITION 6. *We define two concepts for our data and a coordinate system.*

(a) *Data* $\{(\vec{n}_j, c_j)\}_{j=1}^4$ *are called compatible if* $c_1 > c_2 > c_3 > c_4 > 0$ *and all of the following are not vanishing:*

$$\widehat{D}_1 := \det(\vec{n}_2, \vec{n}_3, \vec{n}_4), \quad \widehat{D}_2 := \det(\vec{n}_1, \vec{n}_3, \vec{n}_4),$$
$$\widehat{D}_3 := \det(\vec{n}_1, \vec{n}_2, \vec{n}_4), \quad \widehat{D}_4 := \det(\vec{n}_1, \vec{n}_2, \vec{n}_3),$$

where det *denotes the determinant of a matrix consisting of three vectors. This means that at any given point the normals to any three of the four wave fronts are linearly independent.*

(b) *Data* $\{\vec{n}_j, c_j\}_{j=1}^4$ *are called well prepared if they are compatible and* $\vec{n}_3$, $\vec{n}_4$ *are oriented so that* $\widehat{D}_3 > 0$ *and* $\widehat{D}_4 > 0$.

(c) *For convenience, set the coordinate system* $\{\vec{e}_1, \vec{e}_3, \vec{e}_3\}$ *utilizing* $\vec{n}_1$ *and* $\vec{n}_2$ *by*

$$\vec{e}_1 := \vec{n}_1, \quad \vec{e}_2 := \frac{\vec{n}_2 - (\vec{n}_1 \cdot \vec{n}_2)\vec{n}_1}{|\vec{n}_1 \times \vec{n}_2|}, \quad \vec{e}_3 := \frac{\vec{n}_1 \times \vec{n}_2}{|\vec{n}_1 \times \vec{n}_2|}.$$

Since $-\vec{n}_3$ and $-\vec{n}_4$ also satisfy (3.3), any compatible data can be processed into well prepared data. For well prepared data, we have

$$\vec{n}_1 = \vec{e}_1, \qquad \vec{n}_2 = (\vec{n}_1 \cdot \vec{n}_2)\vec{e}_1 + |\vec{n}_1 \times \vec{n}_2|\vec{e}_2,$$
$$\vec{n}_3 = (\vec{n}_1 \cdot \vec{n}_3)\vec{e}_1 + \frac{(\vec{n}_1 \times \vec{n}_2) \cdot (\vec{n}_1 \times \vec{n}_3)\vec{e}_2}{|\vec{n}_1 \times \vec{n}_2|} + \frac{\widehat{D}_4 \vec{e}_3}{|\vec{n}_1 \times \vec{n}_2|} =: \alpha_3 \vec{e}_1 + \beta_3 \vec{e}_2 + \gamma_3 \vec{e}_3,$$
$$\vec{n}_4 = (\vec{n}_1 \cdot \vec{n}_4)\vec{e}_1 + \frac{(\vec{n}_1 \times \vec{n}_2) \cdot (\vec{n}_1 \times \vec{n}_4)\vec{e}_2}{|\vec{n}_1 \times \vec{n}_2|} + \frac{\widehat{D}_3 \vec{e}_3}{|\vec{n}_1 \times \vec{n}_2|} =: \alpha_4 \vec{e}_1 + \beta_4 \vec{e}_2 + \gamma_4 \vec{e}_3.$$

Here we have $\gamma_3, \gamma_4 > 0$.

3.2. Lemmas based on two or three measurements. Two lemmas using only two or three measurements are presented to show what information can be obtained with the limited data sets.

LEMMA 7 (two measurements). *Using only two data* $\{(\vec{n}_j, c_j)\}_{j=1}^2$, $\tilde{c}_{66}$ *can be any arbitrary number in* $(0, c_2^2]$, *and the first two components of* $\vec{g} = \vec{g}(\tilde{c}_{66})$ *are determined up to four possibilities in terms of* $\tilde{c}_{66}$ *and the measured data*:

$$(\vec{g} \cdot \vec{e}_1, \vec{g} \cdot \vec{e}_2) = \pm \left(d_1, \frac{-d_2 - d_1(\vec{n}_1 \cdot \vec{n}_2)}{|\vec{n}_1 \times \vec{n}_2|} \right) \quad or \quad \pm \left(d_1, \frac{d_2 - d_1(\vec{n}_1 \cdot \vec{n}_2)}{|\vec{n}_1 \times \vec{n}_2|} \right),$$

where $d_1 = \sqrt{c_1^2 - \tilde{c}_{66}}$ *and* $d_2 = \sqrt{c_2^2 - \tilde{c}_{66}}$.

Proof. From (3.3) for $j = 1, 2$ we have $\vec{g} \cdot \vec{e}_1 = \vec{g} \cdot \vec{n}_1 = \pm d_1$ and

$$\pm d_2 = \vec{g} \cdot \vec{n}_2 = \vec{g} \cdot \left[(\vec{n}_1 \cdot \vec{n}_2)\vec{e}_1 + |\vec{n}_1 \times \vec{n}_2| \vec{e}_2 \right] = (\vec{n}_1 \cdot \vec{n}_2)(\vec{g} \cdot \vec{e}_1) + |\vec{n}_1 \times \vec{n}_2|(\vec{g} \cdot \vec{e}_2).$$

Thus we get $\vec{g} \cdot \vec{e}_2 = \frac{\pm d_2 - (\vec{n}_1 \cdot \vec{n}_2)(\vec{g} \cdot \vec{e}_1)}{|\vec{n}_1 \times \vec{n}_2|}$, which completes the proof. □

LEMMA 8 (three measurements). *Using only three data* $\{(\vec{n}_j, c_j)\}_{j=1}^3$, $\tilde{c}_{66}$ *can be any arbitrary number in* $(0, c_3^2]$, *and* $\vec{g} = \vec{g}(\tilde{c}_{66})$ *is determined up to four possibilities in terms of* $\tilde{c}_{66}$ *and the measured data*:

$$\vec{g}_1 = d_1\vec{e}_1 + \eta\vec{e}_2 + \tfrac{1}{\gamma_3}(d_3 - \omega)\vec{e}_3, \qquad \vec{g}_2 = -d_1\vec{e}_1 - \eta\vec{e}_2 + \tfrac{1}{\gamma_3}(d_3 + \omega)\vec{e}_3,$$
$$\vec{g}_3 = d_1\vec{e}_1 + \tilde{\eta}\vec{e}_2 + \tfrac{1}{\gamma_3}(d_3 - \tilde{\omega})\vec{e}_3, \qquad \vec{g}_4 = -d_1\vec{e}_1 - \tilde{\eta}\vec{e}_2 + \tfrac{1}{\gamma_3}(d_3 + \tilde{\omega})\vec{e}_3,$$

where $d_j = \sqrt{c_j^2 - \tilde{c}_{66}}$ *for* $j = 1, 2, 3$, $\eta = \frac{-d_2 - d_1(\vec{n}_1 \cdot \vec{n}_2)}{|\vec{n}_1 \times \vec{n}_2|}$, $\tilde{\eta} = \frac{d_2 - d_1(\vec{n}_1 \cdot \vec{n}_2)}{|\vec{n}_1 \times \vec{n}_2|}$, $\omega = d_1\alpha_3 + \eta\beta_3$, *and* $\tilde{\omega} = d_1\alpha_3 + \tilde{\eta}\beta_3$. *Note that* $\vec{g}_k \cdot \vec{n}_3 = d_3 > 0$, $k = 1, 2, 3, 4$.

Proof. From (3.3) for $j = 3$ we get $\pm d_3 = \vec{g} \cdot \vec{n}_3 = \alpha_3(\vec{g} \cdot \vec{e}_1) + \beta_3(\vec{g} \cdot \vec{e}_2) + \gamma_3(\vec{g} \cdot \vec{e}_3)$. From Lemma 7 we get

$$\vec{g} \cdot \vec{e}_3 = \begin{cases} \frac{1}{\gamma_3}(\pm d_3 - \alpha_3 d_1 - \beta_3\eta) = \frac{1}{\gamma_3}(\pm d_3 - \omega) & \text{if } (\vec{g} \cdot \vec{e}_1, \vec{g} \cdot \vec{e}_2) = (d_1, \eta), \\ \frac{1}{\gamma_3}(\pm d_3 + \alpha_3 d_1 + \beta_3\eta) = \frac{1}{\gamma_3}(\pm d_3 + \omega) & \text{if } (\vec{g} \cdot \vec{e}_1, \vec{g} \cdot \vec{e}_2) = -(d_1, \eta), \\ \frac{1}{\gamma_3}(\pm d_3 - \alpha_3 d_1 - \beta_3\tilde{\eta}) = \frac{1}{\gamma_3}(\pm d_3 - \tilde{\omega}) & \text{if } (\vec{g} \cdot \vec{e}_1, \vec{g} \cdot \vec{e}_2) = (d_1, \tilde{\eta}), \\ \frac{1}{\gamma_3}(\pm d_3 + \alpha_3 d_1 + \beta_3\tilde{\eta}) = \frac{1}{\gamma_3}(\pm d_3 + \tilde{\omega}) & \text{if } (\vec{g} \cdot \vec{e}_1, \vec{g} \cdot \vec{e}_2) = -(d_1, \tilde{\eta}). \end{cases}$$

Thus we have eight possibilities:

$$\begin{pmatrix} \vec{g} \cdot \vec{e}_1 \\ \vec{g} \cdot \vec{e}_2 \\ \vec{g} \cdot \vec{e}_3 \end{pmatrix} = \begin{pmatrix} d_1 \\ \eta \\ \frac{d_3 - \omega}{\gamma_3} \end{pmatrix}, \begin{pmatrix} -d_1 \\ -\eta \\ \frac{d_3 + \omega}{\gamma_3} \end{pmatrix}, \begin{pmatrix} d_1 \\ \tilde{\eta} \\ \frac{d_3 - \tilde{\omega}}{\gamma_3} \end{pmatrix}, \begin{pmatrix} -d_1 \\ -\tilde{\eta} \\ \frac{d_3 + \tilde{\omega}}{\gamma_3} \end{pmatrix},$$
$$\begin{pmatrix} d_1 \\ \eta \\ \frac{-d_3 - \omega}{\gamma_3} \end{pmatrix}, \begin{pmatrix} -d_1 \\ -\eta \\ \frac{-d_3 + \omega}{\gamma_3} \end{pmatrix}, \begin{pmatrix} d_1 \\ \tilde{\eta} \\ \frac{-d_3 - \tilde{\omega}}{\gamma_3} \end{pmatrix}, \begin{pmatrix} -d_1 \\ -\tilde{\eta} \\ \frac{-d_3 + \tilde{\omega}}{\gamma_3} \end{pmatrix}.$$

Since the second line is the same as the first line with opposite sign, which gives the same fiber direction ($\vec{g}$ and $-\vec{g}$) in transversely isotropic media, we select the first line that satisfies $\vec{g} \cdot \vec{n}_3 = d_3 > 0$, and label the four triples in that line as $\vec{g}_1$, $\vec{g}_2$, $\vec{g}_3$, $\vec{g}_4$. □

Remark 9. Note that $(\vec{g}_1 \cdot \vec{n}_k)_{k=1}^3 = (d_1, -d_2, d_3)$, $(\vec{g}_2 \cdot \vec{n}_k)_{k=1}^3 = (-d_1, d_2, d_3)$, $(\vec{g}_3 \cdot \vec{n}_k)_{k=1}^3 = (d_1, d_2, d_3)$, and $(\vec{g}_4 \cdot \vec{n}_k)_{k=1}^3 = (-d_1, -d_2, d_3)$.

3.3. Four measurements.

In the previous subsection, we showed that from three measurements, $\tilde{c}_{66}$ is a continuous parameter that can be anything in $(0, c_3^2]$ and our solution $(\tilde{c}_{66}, \vec{g})$ can be any of four continuous families

$$\{(\tilde{c}_{66}, \vec{g}_k(\tilde{c}_{66})) : \tilde{c}_{66} \in (0, c_3^2]\}_{k=1}^4.$$

But in this subsection we will show that for four measurements the set of possible $\tilde{c}_{66}$ becomes discrete, with a maximum number of at most four, and will provide an explicit formula for $\vec{g}$ corresponding to each $\tilde{c}_{66}$.

LEMMA 10. *For* $\{\vec{g}_k\}_{k=1}^4$ *in Lemma* 8, *we get*

$$\vec{g}_1 \cdot \vec{n}_4 = \frac{d_1\widehat{D}_1 + d_2\widehat{D}_2 + d_3\widehat{D}_3}{\widehat{D}_4}, \quad \vec{g}_2 \cdot \vec{n}_4 = \frac{-d_1\widehat{D}_1 - d_2\widehat{D}_2 + d_3\widehat{D}_3}{\widehat{D}_4},$$
$$\vec{g}_3 \cdot \vec{n}_4 = \frac{d_1\widehat{D}_1 - d_2\widehat{D}_2 + d_3\widehat{D}_3}{\widehat{D}_4}, \quad \vec{g}_4 \cdot \vec{n}_4 = \frac{-d_1\widehat{D}_1 + d_2\widehat{D}_2 + d_3\widehat{D}_3}{\widehat{D}_4}.$$

Proof. Because all of the others are analogous, we will show only the first case. From Lemma 8, we have

$$\vec{g}_1 \cdot \vec{n}_4 = d_1\alpha_4 + \eta\beta_4 + \frac{d_3\gamma_4}{\gamma_3} - \frac{\omega\gamma_4}{\gamma_3} = \frac{d_1\alpha_4\gamma_3 + \eta\beta_4\gamma_3 - \omega\gamma_4}{\gamma_3} + \frac{d_3\widehat{D}_3}{\widehat{D}_4}.$$

Since $d_1\vec{e}_2 - \eta\vec{e}_1 = \frac{d_1\vec{n}_2 + d_2\vec{n}_1}{|\vec{n}_1 \times \vec{n}_2|} = \frac{\gamma_3(d_1\vec{n}_2 + d_2\vec{n}_1)}{\widehat{D}_4}$, we get

$$\begin{aligned}\frac{d_1\alpha_4\gamma_3 + \eta\beta_4\gamma_3 - \omega\gamma_4}{\gamma_3} &= \frac{d_1(\alpha_4\gamma_3 - \alpha_3\gamma_4) - \eta(\beta_3\gamma_4 - \beta_4\gamma_3)}{\gamma_3} \\ &= \frac{(d_1\vec{e}_2 - \eta\vec{e}_1)\cdot(\vec{n}_3 \times \vec{n}_4)}{\gamma_3} = \frac{d_1\widehat{D}_1 + d_2\widehat{D}_2}{\widehat{D}_4},\end{aligned}$$

which completes the proof. □

THEOREM 11. *Let well prepared data $\{\vec{n}_j, c_j\}_{j=1}^4$ be given. Then $\tilde{c}_{66} \in (0, c_4^2]$ is the first function, c_{66}/ρ, in the solution of the inverse problem* (3.1) *if and only if $\tilde{c}_{66}$ satisfies one of the following eight equations*:

$$\pm d_1\widehat{D}_1 \pm d_2\widehat{D}_2 + d_3\widehat{D}_3 \pm d_4\widehat{D}_4 = 0, \qquad d_j := \sqrt{c_j^2 - \tilde{c}_{66}} \geq 0. \tag{3.4}$$

Proof. If $\tilde{c}_{66} \in (0, c_4^2]$ is a solution, then there exist $\vec{g}$ and $\{d_j \geq 0\}_{j=1}^4$ that satisfy (3.3). By Lemma 8, $\vec{g}$ should be one of $\{\vec{g}_k\}_{k=1}^4$, and by Remark 9 we have $(\vec{g}\cdot\vec{n}_1, \vec{g}\cdot\vec{n}_2, \vec{g}\cdot\vec{n}_3, \vec{g}\cdot\vec{n}_4) = (\pm d_1, \pm d_2, d_3, \pm d_4)$. Thus we have $AX = 0$, where

$$A := \begin{pmatrix} \vec{n}_1 & \mp d_1 \\ \vec{n}_2 & \mp d_2 \\ \vec{n}_3 & -d_3 \\ \vec{n}_4 & \mp d_4 \end{pmatrix}, \qquad X := \begin{pmatrix} \vec{g} \\ 1 \end{pmatrix}.$$

For X to be a nontrivial solution, we must get $0 = \det A = \mp d_1\widehat{D}_1 \pm d_2\widehat{D}_2 - d_3\widehat{D}_3 \pm d_4\widehat{D}_4$, which proves the necessity. For sufficiency, from Remark 9 we know that all four of the $\vec{g}_k$ in Lemma 8 already satisfy (3.3) for $j = 1, 2, 3$ for any $\tilde{c}_{66} \in (0, c_3^2]$. In addition, if $\tilde{c}_{66}$ satisfies one of (3.4), then $\tilde{c}_{66} \leq c_4^2$, and one of $\vec{g}_k \cdot \vec{n}_4$ in Lemma 10 satisfies $\vec{g}_k \cdot \vec{n}_4 = \pm d_4$. So these particular $\vec{g}_k$ and $\tilde{c}_{66}$ satisfy (3.3) for $j = 1, 2, 3, 4$. Therefore $\tilde{c}_{66}$ is the first function in a solution of (3.1). □

LEMMA 12. *Let $\Pi(a, b, c, d)$ be the alternating product*

$$\Pi(a, b, c, d) := \prod_{i,j,\ell \in \{0,1\}} \Big((-1)^i a + (-1)^j b + c + (-1)^\ell d \Big).$$

Then we have $\Pi(a, b, c, d) = (A_4 + A_2)^2(A_4 - A_2)^2 + 4(A_1 - A_3)(A_1 A_4^2 - A_3 A_2^2)$, where $A_1 = a^2 + b^2$, $A_2 = a^2 - b^2$, $A_3 = c^2 + d^2$, $A_4 = c^2 - d^2$.

Proof. The proof can be easily shown by tedious calculation. □

THEOREM 13 (determination of $\tilde{c}_{66}$). *$\tilde{c}_{66}$ satisfies one of* (3.4) *if and only if $\tilde{c}_{66} \in (0, c_4^2]$ is a root of $p(x) := \Pi(\hat{d}_1\widehat{D}_1, \hat{d}_2\widehat{D}_2, \hat{d}_3\widehat{D}_3, \hat{d}_4\widehat{D}_4)$, where $\hat{d}_j = \hat{d}_j(x) := \sqrt{c_j^2 - x}$. Here $p(x)$ becomes a fourth order polynomial*

$$p(x) = (l_4 + l_2)^2(l_4 - l_2)^2 + 4(l_1 - l_3)(l_1 l_4^2 - l_3 l_2^2), \qquad l_j := l_j(x) = a_j x - b_j, \tag{3.5}$$

where $a_1 = \widehat{D}_1^2 + \widehat{D}_2^2$, $a_2 = \widehat{D}_1^2 - \widehat{D}_2^2$, $a_3 = \widehat{D}_3^2 + \widehat{D}_4^2$, $a_4 = \widehat{D}_3^2 - \widehat{D}_4^2$, $b_1 = c_1^2\widehat{D}_1^2 + c_2^2\widehat{D}_2^2$, $b_2 = c_1^2\widehat{D}_1^2 - c_2^2\widehat{D}_2^2$, $b_3 = c_3^2\widehat{D}_3^2 + c_4^2\widehat{D}_4^2$, and $b_4 = c_3^2\widehat{D}_3^2 - c_4^2\widehat{D}_4^2$.

Proof. By definition, $p(x) = \Pi(\hat{d}_1\widehat{D}_1, \hat{d}_2\widehat{D}_2, \hat{d}_3\widehat{D}_3, \hat{d}_4\widehat{D}_4)$ is simply a product of the following eight factors:

$$\pm\hat{d}_1\widehat{D}_1 \pm \hat{d}_2\widehat{D}_2 + \hat{d}_3\widehat{D}_3 \pm \hat{d}_4\widehat{D}_4.$$

Since $d_j = \hat{d}_j(\tilde{c}_{66})$, the fact that $\tilde{c}_{66}$ satisfies one of (3.4) is equivalent to finding a root of $p(x)$. Moreover, using Lemma 12 with $A_j = -l_j(x)$, we can easily show that $p(x)$ is the fourth order polynomial given in (3.5). □

Since $p(x)$ is a fourth order polynomial, we have at most four possible $\tilde{c}_{66}$, and each $\tilde{c}_{66}$ satisfies at least one of (3.4), or equivalently, one of the following four equations:

$$(d_1\widehat{D}_1 + d_2\widehat{D}_2 + d_3\widehat{D}_3)^2 = d_4^2\widehat{D}_4^2, \tag{3.6}$$

$$(d_1\widehat{D}_1 + d_2\widehat{D}_2 - d_3\widehat{D}_3)^2 = d_4^2\widehat{D}_4^2, \tag{3.7}$$

$$(d_1\widehat{D}_1 - d_2\widehat{D}_2 + d_3\widehat{D}_3)^2 = d_4^2\widehat{D}_4^2, \tag{3.8}$$

$$(d_1\widehat{D}_1 - d_2\widehat{D}_2 - d_3\widehat{D}_3)^2 = d_4^2\widehat{D}_4^2. \tag{3.9}$$

Each equation corresponds to a product of two equations in (3.4). So we obtain the following theorem, which provides the corresponding $\vec{g}$ (i.e., $\tilde{c}_{44}$ and $\vec{f}$) for each case when $\tilde{c}_{66}$ solves one of the above four equations.

THEOREM 14 (determination of $\vec{g}$). *Let $\tilde{c}_{66} \in (0, c_4^2]$ be a root of $p(x)$ given in* (3.5). *Then $\tilde{c}_{66}$ satisfies at least one of* (3.6)–(3.9), *and for each case the corresponding $\vec{g}$ is determined by*

$$\vec{g} = \begin{cases} \vec{g}_1 & \textit{if and only if } \tilde{c}_{66} \textit{ satisfies } (3.6), \\ \vec{g}_2 & \textit{if and only if } \tilde{c}_{66} \textit{ satisfies } (3.7), \\ \vec{g}_3 & \textit{if and only if } \tilde{c}_{66} \textit{ satisfies } (3.8), \\ \vec{g}_4 & \textit{if and only if } \tilde{c}_{66} \textit{ satisfies } (3.9), \end{cases}$$

where $\{\vec{g}_k\}_{k=1}^4$ are defined as in Lemma 8.

Proof. By Remark 9, (3.3) is already satisfied for $j = 1, 2, 3$. For $j = 4$, i.e., $\vec{g} \cdot \vec{n}_4 = \pm d_4$, it is easily checked by Lemma 10 for each case. □

Later we will show that *generically* only one of (3.6)–(3.9) is satisfied for each $\tilde{c}_{66}$, and hence the maximum number of possible solutions $(\tilde{c}_{66}, \vec{g})$ will be at most four.

3.4. Multiple $\vec{g}$ for a single $\tilde{c}_{66}$. Throughout the rest of the paper we will assume that the data sets are well prepared. We define three special types of data allowing multiple $\vec{g}$ corresponding to a single $\tilde{c}_{66}$:

$$\begin{gathered} |\widehat{D}_1| < |\widehat{D}_2|, \quad \widehat{D}_3 < \widehat{D}_4, \quad c_1^2\widehat{D}_1^2 < c_2^2\widehat{D}_2^2, \quad c_3^2\widehat{D}_3 < c_4^2\widehat{D}_4^2, \\ (c_4^2\widehat{D}_4^2 - c_3^2\widehat{D}_3^2)(\widehat{D}_2^2 - \widehat{D}_1^2) = (c_2^2\widehat{D}_2^2 - c_1^2\widehat{D}_1^2)(\widehat{D}_4^2 - \widehat{D}_3^2), \end{gathered} \tag{3.10}$$

$$\begin{gathered} |\widehat{D}_1| < \widehat{D}_3, \quad |\widehat{D}_2| < \widehat{D}_4, \quad c_1^2\widehat{D}_1^2 < c_3^2\widehat{D}_3^2, \quad c_2^2\widehat{D}_2 < c_4^2\widehat{D}_4^2, \\ (c_4^2\widehat{D}_4^2 - c_2^2\widehat{D}_2^2)(\widehat{D}_3^2 - \widehat{D}_1^2) = (c_3^2\widehat{D}_3^2 - c_1^2\widehat{D}_1^2)(\widehat{D}_4^2 - \widehat{D}_2^2), \end{gathered} \tag{3.11}$$

$$\begin{gathered} |\widehat{D}_1| < \widehat{D}_4, \quad |\widehat{D}_2| < \widehat{D}_3, \quad c_1^2\widehat{D}_1^2 < c_4^2\widehat{D}_4^2, \quad c_2^2\widehat{D}_2 < c_3^2\widehat{D}_3^2, \\ (c_3^2\widehat{D}_3^2 - c_2^2\widehat{D}_2^2)(\widehat{D}_4^2 - \widehat{D}_1^2) = (c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2)(\widehat{D}_3^2 - \widehat{D}_2^2). \end{gathered} \tag{3.12}$$

First, note that we should have $\tilde{c}_{66} < c_4^2$ in order to have more than one $\vec{g}$: the reason is that if $d_4 = \sqrt{c_4^2 - \tilde{c}_{66}} = 0$, then by Theorem 14 at least two of (3.6)–(3.9) are satisfied, implying that at least one of $d_1\widehat{D}_1$, $d_2\widehat{D}_2$, $d_3\widehat{D}_3$ is zero, which is a contradiction. In the following theorem we will see that each of (3.10)–(3.12) actually enforces $\tilde{c}_{66} < c_4^2$.

THEOREM 15 (two $\vec{g}$ for a single $\tilde{c}_{66}$).

(a) *Both* $\vec{g}_1$ *and* $\vec{g}_2$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.10) *and* $\widehat{D}_1\widehat{D}_2 < 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_2\widehat{D}_2 = -d_1\widehat{D}_1$ *and* $d_4\widehat{D}_4 = d_3\widehat{D}_3$.
Both $\vec{g}_3$ *and* $\vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.10) *and* $\widehat{D}_1\widehat{D}_2 > 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_2\widehat{D}_2 = d_1\widehat{D}_1$ *and* $d_4\widehat{D}_4 = d_3\widehat{D}_3$.
For the above two cases,

$$\tilde{c}_{66} = \frac{c_2^2\widehat{D}_2^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_2^2 - \widehat{D}_1^2} = \frac{c_4^2\widehat{D}_4^2 - c_3^2\widehat{D}_3^2}{\widehat{D}_4^2 - \widehat{D}_3^2} \in (0, c_4^2).$$

(b) *Both* $\vec{g}_1$ *and* $\vec{g}_3$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.11) *and* $\widehat{D}_1 < 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_3\widehat{D}_3 = -d_1\widehat{D}_1$ *and* $d_4\widehat{D}_4 = \pm d_2\widehat{D}_2$.
Both $\vec{g}_2$ *and* $\vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.11) *and* $\widehat{D}_1 > 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_3\widehat{D}_3 = d_1\widehat{D}_1$ *and* $d_4\widehat{D}_4 = \pm d_2\widehat{D}_2$.
For the above two cases,

$$\tilde{c}_{66} = \frac{c_3^2\widehat{D}_3^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_3^2 - \widehat{D}_1^2} = \frac{c_4^2\widehat{D}_4^2 - c_2^2\widehat{D}_2^2}{\widehat{D}_4^2 - \widehat{D}_2^2} \in (0, c_4^2).$$

(c) *Both* $\vec{g}_1$ *and* $\vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.12) *and* $\widehat{D}_2 < 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_3\widehat{D}_3 = -d_2\widehat{D}_2$ *and* $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$.
Both $\vec{g}_2$ *and* $\vec{g}_3$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.12) *and* $\widehat{D}_2 > 0$
$\Leftrightarrow$ $\tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_3\widehat{D}_3 = d_2\widehat{D}_2$ *and* $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$.
For the above two cases,

$$\tilde{c}_{66} = \frac{c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2} = \frac{c_3^2\widehat{D}_3^2 - c_2^2\widehat{D}_2^2}{\widehat{D}_3^2 - \widehat{D}_2^2} \in (0, c_4^2).$$

Proof. Because all of the others are analogous, we will show only (c). For $\vec{g}_1$ and $\vec{g}_4$ to be the solutions, from Theorem 14 $\tilde{c}_{66}$ must satisfy (3.6) and (3.9). Since $d_1\widehat{D}_1 \neq 0$, we must have $d_3\widehat{D}_3 = -d_2\widehat{D}_2$, and so $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$. For $\vec{g}_2$ and $\vec{g}_3$ to be the solutions, from Theorem 14 $\tilde{c}_{66}$ must satisfy (3.7) and (3.8). Since $d_1\widehat{D}_1 \neq 0$, we must have $d_3\widehat{D}_3 = d_2\widehat{D}_2$, and so $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$.

Now we will show that the data satisfy (3.12) and $\widehat{D}_2 \lessgtr 0$, respectively, if $\tilde{c}_{66} \in (0, c_4^2)$ solves $d_3\widehat{D}_3 = \mp d_2\widehat{D}_2$ and $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$. First note that $\widehat{D}_2 \lessgtr 0$, respectively, since $d_2, d_3 > 0$. For both cases, we get

$$\widehat{D}_3^2(c_3^2 - \tilde{c}_{66}) = d_3^2\widehat{D}_3^2 = d_2^2\widehat{D}_2^2 = \widehat{D}_2^2(c_2^2 - \tilde{c}_{66}),$$
$$\widehat{D}_4^2(c_4^2 - \tilde{c}_{66}) = d_4^2\widehat{D}_4^2 = d_1^2\widehat{D}_1^2 = \widehat{D}_1^2(c_1^2 - \tilde{c}_{66}),$$

and thus we should have $\tilde{c}_{66} = \frac{c_3^2\widehat{D}_3^2 - c_2^2\widehat{D}_2^2}{\widehat{D}_3^2 - \widehat{D}_2^2} = \frac{c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2}$. So we have

$$(c_3^2\widehat{D}_3^2 - c_2^2\widehat{D}_2^2)(\widehat{D}_4^2 - \widehat{D}_1^2) = (c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2)(\widehat{D}_3^2 - \widehat{D}_2^2).$$

Moreover, since $d_1 > d_2 > d_3 > d_4$, we get $|\widehat{D}_1| < \widehat{D}_4$ and $|\widehat{D}_2| < \widehat{D}_3$. From $\tilde{c}_{66} > 0$, we also get $c_1^2\widehat{D}_1 < c_4^2\widehat{D}_4^2$ and $c_2^2\widehat{D}_2^2 < c_3^2\widehat{D}_3^2$.

Finally, we will show $\tilde{c}_{66} \in (0, c_4^2)$, and it solves $d_3\widehat{D}_3 = \mp d_2\widehat{D}_2$, $d_4\widehat{D}_4 = \pm d_1\widehat{D}_1$ if the data satisfy (3.12) and $\widehat{D}_2 \lesseqgtr 0$, respectively. Set $\tilde{c}_{66} := \frac{c_3^2\widehat{D}_3^2 - c_2^2\widehat{D}_2^2}{\widehat{D}_3^2 - \widehat{D}_2^2} = \frac{c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2} > 0$. Since $c_1^2 > c_4^2$, we get $\tilde{c}_{66} = \frac{c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2} < \frac{c_4^2\widehat{D}_4^2 - c_4^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2} = c_4^2$. Thus $\tilde{c}_{66} \in (0, c_4^2)$ and solves $d_3^2\widehat{D}_3^2 = d_2^2\widehat{D}_2^2$ and $d_4^2\widehat{D}_4^2 = d_1^2\widehat{D}_1^2$. Since $\widehat{D}_2 \lesseqgtr 0$, $\tilde{c}_{66}$ solves $d_3\widehat{D}_3 = \mp d_2\widehat{D}_2$, respectively, and $d_4^2\widehat{D}_4^2 = d_1^2\widehat{D}_1^2$. □

In each case in (3.10)–(3.12), the fourth order polynomial $p(x)$ for $\tilde{c}_{66}$ in Theorem 13 is now further simplified. We will use the following theorem to show the generic uniqueness of $\vec{g}$ in Corollary 20.

THEOREM 16. *In each case in* (3.10)–(3.12), $p(x)$ *in Theorem* 13 *becomes*

$$p(x) = \begin{cases} \left((\widehat{D}_2^2 - \widehat{D}_1^2)x - (c_2^2\widehat{D}_2^2 - c_1^2\widehat{D}_1^2)\right)^2 q_1(x) & \textit{if the data satisfy } (3.10), \\ \left((\widehat{D}_3^2 - \widehat{D}_1^2)x - (c_3^2\widehat{D}_3^2 - c_1^2\widehat{D}_1^2)\right)^2 q_2(x) & \textit{if the data satisfy } (3.11), \\ \left((\widehat{D}_4^2 - \widehat{D}_1^2)x - (c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2)\right)^2 q_3(x) & \textit{if the data satisfy } (3.12), \end{cases}$$

where q_1, q_2, *and* q_3 *are second order polynomials.*

Proof. Because all of the others are analogous, we will show only the case (3.11). First note that

$$\widehat{D}_1^2 = \frac{a_1 + a_2}{2}, \quad \widehat{D}_2^2 = \frac{a_1 - a_2}{2}, \quad \widehat{D}_3^2 = \frac{a_3 + a_4}{2}, \quad \widehat{D}_4^2 = \frac{a_3 - a_4}{2},$$

$$c_1^2\widehat{D}_1^2 = \frac{b_1 + b_2}{2}, \quad c_2^2\widehat{D}_2^2 = \frac{b_1 - b_2}{2}, \quad c_3^2\widehat{D}_3^2 = \frac{b_3 + b_4}{2}, \quad c_4^2\widehat{D}_4^2 = \frac{b_3 - b_4}{2},$$

and define

$$A := \frac{a_4 + a_3 - a_2 - a_1}{2} = \widehat{D}_3^2 - \widehat{D}_1^2 > 0, \quad B := \frac{b_4 + b_3 - b_2 - b_1}{2} = c_3^2\widehat{D}_3^2 - c_1^2\widehat{D}_1^2 > 0.$$

Since we can show $(a_4 - a_2)(b_3 - b_1) = (a_3 - a_1)(b_4 - b_2)$ from (3.11), we also get $A(b_3 - b_1) = B(a_3 - a_1)$ and $A(b_4 - b_2) = B(a_4 - a_2)$. Thus we get

$$A(l_1(x) - l_3(x)) = (a_1 - a_3)(Ax - B), \qquad A(l_2(x) - l_4(x)) = (a_2 - a_4)(Ax - B).$$

Hence the polynomial in (3.5) becomes

$$p(x) = C_2^2(Ax - B)^2(l_4 + l_2)^2 - 4C_1(Ax - B)Q(x),$$

where $C_1 = \frac{a_3 - a_1}{A}$, $C_2 = \frac{a_4 - a_2}{A}$, and

$$\begin{aligned} Q(x) &= l_1 l_4^2 - l_3\left[l_4 - C_2(Ax - B)\right]^2 \\ &= (l_1 - l_3)l_4^2 - l_3(Ax - B)\left[-2l_4C_2 + C_2^2(Ax - B)\right] \\ &= (Ax - B)\left[2C_2l_3l_4 - C_1l_4^2 - C_2^2l_3(Ax - B)\right] \\ &= (Ax - B)\left[2C_2l_3l_4 - C_1l_4^2 + C_2l_3(l_2 - l_4)\right] = (Ax - B)\left[C_2l_3(l_4 + l_2) - C_1l_4^2\right]. \end{aligned}$$

Hence we get $p(x) = (Ax - B)^2 q_2(x)$, where q_2 is a second order polynomial given by $q_2(x) = C_2^2(l_4 + l_2)^2 + 4C_1^2l_4^2 - 4C_1C_2l_3(l_2 + l_4)$. □

For three possible $\vec{g}$ corresponding to a single $\tilde{c}_{66}$, we consider another special type of data that satisfies

$$\begin{gathered}|\widehat{D}_1| < |\widehat{D}_2| < \widehat{D}_3 < \widehat{D}_4, \quad c_1^2\widehat{D}_1^2 < c_2^2\widehat{D}_2^2 < c_3^2\widehat{D}_3 < c_4^2\widehat{D}_4^2, \\ \text{and there exists a single } K := \frac{c_j^2\widehat{D}_j^2 - c_i^2\widehat{D}_i^2}{\widehat{D}_j^2 - \widehat{D}_i^2} \in (0, c_4^2) \quad \forall j > i.\end{gathered} \tag{3.13}$$

Note that (3.13) implies $\{(x_j, y_j) = (\widehat{D}_j^2, c_j^2\widehat{D}_j^2)\}$ are on a single straight line with slope $K > 0$. From Theorem 15, we can easily prove the following theorem, showing exactly when we shall get three $\vec{g}$ corresponding to a single $\tilde{c}_{66}$.

THEOREM 17 (three $\vec{g}$ for a single $\tilde{c}_{66}$).

(a) $\vec{g}_1, \vec{g}_2, \vec{g}_3$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.13), $\widehat{D}_1 < 0$, *and* $\widehat{D}_2 > 0$ $\Leftrightarrow \tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_4\widehat{D}_4 = d_3\widehat{D}_3 = d_2\widehat{D}_2 = -d_1\widehat{D}_1$.

(b) $\vec{g}_1, \vec{g}_2, \vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.13), $\widehat{D}_1 > 0$, *and* $\widehat{D}_2 < 0$ $\Leftrightarrow \tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_4\widehat{D}_4 = d_3\widehat{D}_3 = -d_2\widehat{D}_2 = d_1\widehat{D}_1$.

(c) $\vec{g}_1, \vec{g}_3, \vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.13), $\widehat{D}_1 < 0$, *and* $\widehat{D}_2 < 0$ $\Leftrightarrow \tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_4\widehat{D}_4 = d_3\widehat{D}_3 = -d_2\widehat{D}_2 = -d_1\widehat{D}_1$.

(d) $\vec{g}_2, \vec{g}_3, \vec{g}_4$ *are solutions* $\Leftrightarrow$ *the data satisfy* (3.13), $\widehat{D}_1 > 0$, *and* $\widehat{D}_2 > 0$ $\Leftrightarrow \tilde{c}_{66} \in (0, c_4^2)$ *and it solves* $d_4\widehat{D}_4 = d_3\widehat{D}_3 = d_2\widehat{D}_2 = d_1\widehat{D}_1$.

In each case, we should have $\tilde{c}_{66} = K \in (0, c_4^2)$ *as given in* (3.13).

As in Theorem 16, the fourth order polynomial $p(x)$ for $\tilde{c}_{66}$ in Theorem 13 is also further simplified when the data satisfy (3.13). The following theorem will also be used to show the generic uniqueness of $\vec{g}$ in Corollary 20.

THEOREM 18. *If the data satisfy* (3.13), $p(x)$ *in Theorem* 13 *becomes*

$$p(x) = (x - K)^3 \left(\Pi(\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4)x - \frac{\Pi(c_1\widehat{D}_1, c_2\widehat{D}_2, c_3\widehat{D}_3, c_4\widehat{D}_4)}{K^3} \right),$$

where $K \in (0, c_4^2)$ *is given as in* (3.13).

Proof. Since $a_2, a_4, a_1 - a_3 < 0$ and $K = \frac{b_2}{a_2} = \frac{b_4}{a_4} = \frac{b_1 - b_3}{a_1 - a_3}$, we get

$$l_2(x) = a_2 l(x), \quad l_4(x) = a_4 l(x), \quad l_1(x) - l_3(x) = (a_1 - a_3)l(x),$$

where $l(x) := x - K$. Hence the polynomial in (3.5) becomes $p(x) = [l(x)]^3 Q(x)$, where

$$Q(x) = (a_4^2 - a_2^2)^2 l(x) + 4(a_1 - a_3)(a_4^2 l_1(x) - a_2^2 l_3(x)).$$

Here $Q(x)$ is definitely a linear function, and from

$$\begin{aligned} Q'(0) &= (a_4^2 - a_2^2)^2 + 4(a_1 - a_3)(a_1 a_4^2 - a_3 a_2^2) = \Pi(\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4), \\ Q(0) &= -K(a_4^2 - a_2^2)^2 + 4(a_1 - a_3)(a_2^2 b_3 - a_4^2 b_1) \\ &= -\frac{(b_4^2 - b_2^2)^2 + 4(b_1 - b_3)(b_1 b_4^2 - b_3 b_2^2)}{K^3} = -\frac{\Pi(c_1\widehat{D}_1, c_2\widehat{D}_2, c_3\widehat{D}_3, c_4\widehat{D}_4)}{K^3}, \end{aligned}$$

we conclude $Q(x) = \Pi(\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4)x - \frac{1}{K^3}\Pi(c_1\widehat{D}_1, c_2\widehat{D}_2, c_3\widehat{D}_3, c_4\widehat{D}_4)$. □

Remark 19. From Theorem 17 we can easily prove that the four $\vec{g}_1$, $\vec{g}_2$, $\vec{g}_3$, $\vec{g}_4$ cannot all be solutions at the same time for a single $\tilde{c}_{66}$: If so, then $d_1\widehat{D}_1 = 0$, which is a contradiction.

3.5. Generic uniqueness of $\vec{g}$ for a single $\tilde{c}_{66}$. In this subsection we show that *generically* only one of (3.6)–(3.9) is satisfied for each root $\tilde{c}_{66} \in (0, c_4^2]$ of $p(x)$, and hence generically the maximum number of possible solutions $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$ is at most four.

Corollary 20.

(a) *Let $\tilde{c}_{66} \in (0, c_4^2]$ be a root of $p(x)$ in* (3.5) *and m be its multiplicity. If we denote by $G(\tilde{c}_{66})$ the number of possible $\vec{g}$ corresponding to this $\tilde{c}_{66}$, then* $1 \le G(\tilde{c}_{66}) \le \min(m, 3)$.

(b) *The number of all possible $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$ is less than or equal to the number of* (*multiply counted*) *roots of $p(x)$ in $(0, c_4^2]$, which cannot exceed four.*

(c) *Unless the data satisfy one of the special conditions* (3.10)–(3.12), *there exists only one $\vec{g}$ corresponding to a single $\tilde{c}_{66}$. So in this case, the number of all possible $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$ is exactly the same as the number of (not multiply counted) roots of $p(x)$ in $(0, c_4^2]$, which cannot exceed four.*

Proof. We first prove (a). For any root $\tilde{c}_{66}$, at least one of (3.6)–(3.9) is satisfied, so by Theorem 14 we have $G(\tilde{c}_{66}) \ge 1$. Also Remark 19 says that $G(\tilde{c}_{66}) \le 3$. Hence it suffices to show $G(\tilde{c}_{66}) \le m$ for $m = 1, 2$. For a simple root ($m = 1$), if $G(\tilde{c}_{66}) \ge 2$, then by Theorems 15 and 16 we get $m \ge 2$, which is a contradiction. So we should have $G(\tilde{c}_{66}) \le 1$. For a double root ($m = 2$), if $G(\tilde{c}_{66}) \ge 3$, then by Theorems 17 and 18 we get $m \ge 3$, which is a contradiction. So we should have $G(\tilde{c}_{66}) \le 2$. (b) is straightforward from (a), and so is (c) from Theorems 14 and 15. □

From all the above, we can summarize our algorithm as follows:

1. Make the compatible data to be well prepared.
2. Determine possible $\tilde{c}_{66} \in (0, c_4^2]$ by finding roots of the fourth order polynomial $p(x)$ in (3.5).
3. For each $\tilde{c}_{66} \in (0, c_4^2]$ obtained above, check which one among (3.6)–(3.9) is satisfied.
4. For each case, use Theorem 14 to determine $\vec{g}$ (equivalently, $\tilde{c}_{44}$ and $\vec{f}$).

Remark 21. If we use all of the information about the solution u of (2.2) (as opposed to only the wave front positions $\hat{T}$), which is actually measured in experiments, then we can apply the same arguments of section 5 in [11]. That is, for one of the possible triples $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$ given there corresponds at most one density ρ corresponding to that triple under the Neumann boundary condition (for the Dirichlet boundary condition, ρ needs to be specified on the boundary). Therefore, in this case, we have at most four possibilities in determining four parameters $(\rho, c_{66}, c_{44}, \vec{f})$ from the data set $\{u_j(x,t) \mid x \in \Omega,\ t \in (0,T)\}_{j=1}^4$.

3.5.1. Examples. Here a complete set of examples is presented showing that sometimes there exist no solution, a unique solution, two solutions, three solutions, or four solutions. Here we converted the final solution into the standard coordinate system to represent $\vec{f}$.

Example 22 (no solution). Consider the following well prepared data:

$$\vec{n}_1 = (1, 0, 0), \quad \vec{n}_2 = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \quad \vec{n}_3 = (0, 0, -1), \quad \vec{n}_4 = \left(0, \frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$$

with $c_1^2 = 9$, $c_2^2 = 3$, $c_3^2 = \frac{5}{2}$, and $c_4^2 = 1$. Then the fourth order polynomial for $\tilde{c}_{66}$ is given by $p(x) = \frac{1}{65536}(18145 + 34592x + 21440x^2 + 7168x^3 + 1024x^4)$, which has no root in $(0, c_4^2]$. Hence there exists no solution matching the data.

Example 23 (unique solution). Consider the same $\vec{n}_j$ as in Example 22 with $c_1^2 = 8$, $c_2^2 = 7$, $c_3^2 = 2$, and $c_4^2 = 1$. Then $p(x) = \frac{1}{64}(-188 + 252x + 53x^2 - 18x^3 + x^4)$, which has only one root $\tilde{c}_{66} = \frac{1}{2}(9 - \sqrt{137 - 32\sqrt{6}}) \approx 0.6719$ in $(0, c_4^2]$. Since $\tilde{c}_{66}$ satisfies (3.7), from Theorem 14 we get the unique solution given by

$$\begin{pmatrix} \tilde{c}_{66} \\ \tilde{c}_{44} \\ \vec{f} \end{pmatrix} \approx \begin{pmatrix} 0.6719 \\ 10.0514 \\ (-0.8839, -0.2777, -0.3763) \end{pmatrix}.$$

Example 24 (two solutions). Consider the same $\vec{n}_j$ as in Example 22 with $c_1^2 = 9$, $c_2^2 = 5$, $c_3^2 = 4$, and $c_4^2 = 3$. Since these data satisfy none of (3.10)–(3.12), only one of $\vec{g}$ will correspond to each root of the fourth order polynomial given by

$$p(x) = \tfrac{1}{4096}(481 - 1488x + 1264x^2 - 384x^3 + 64x^4).$$

This polynomial has two roots $\xi_1 \approx 0.5194$ and $\xi_2 \approx 1.2439$ in $(0, c_4^2]$, where $\tilde{c}_{66} = \xi_1$ satisfies (3.7) and $\tilde{c}_{66} = \xi_2$ satisfies (3.8). So by Theorem 14, we get two solutions:

$$\begin{pmatrix} \tilde{c}_{66} \\ \tilde{c}_{44} \\ \vec{f} \end{pmatrix} \approx \begin{pmatrix} 0.5194 \\ 12.4873 \\ (-0.8418, -0.0235, -0.5393) \end{pmatrix}, \quad \begin{pmatrix} 1.2439 \\ 42.2907 \\ (0.4347, -0.8625, -0.2591) \end{pmatrix}.$$

Example 25 (three solutions). Consider the same $\vec{n}_j$ as in Example 22 with $c_1^2 = 3$, $c_2^2 = \frac{5}{3}$, $c_3^2 = \frac{4}{3}$, and $c_4^2 = \frac{5}{4}$. Since these data satisfy (3.12) and $\widehat{D}_2 > 0$, by Theorem 15(c) we get two solutions $\vec{g}_2$ and $\vec{g}_3$ for $\tilde{c}_{66} = \frac{c_4^2\widehat{D}_4^2 - c_1^2\widehat{D}_1^2}{\widehat{D}_4^2 - \widehat{D}_1^2} = \frac{2}{3}$. Moreover, we have $p(x) = \frac{1}{576}(3x-2)^2\left[x - \left(\sqrt{2} - \frac{1}{6}\right)\right]\left[x + \left(\sqrt{2} + \frac{1}{6}\right)\right]$, which has another root $\tilde{c}_{66} = \sqrt{2} - \frac{1}{6} \approx 1.2476$ in $(0, c_4^2]$ satisfying (3.6). So we get three solutions:

$$\begin{pmatrix} \tilde{c}_{66} \\ \tilde{c}_{44} \\ \vec{f} \end{pmatrix} \approx \begin{pmatrix} 0.6667 \\ 3.6795 \\ (-0.8800, 0.0653, -0.4704) \end{pmatrix}, \quad \begin{pmatrix} 1.2476 \\ 3.2525 \\ (0.9349, -0.2883, -0.2069) \end{pmatrix},$$

$$\begin{pmatrix} 0.6667 \\ 12.3205 \\ (0.4475, -0.8617, -0.2392) \end{pmatrix}.$$

Example 26 (four solutions). Consider the following well prepared data:

$$\vec{n}_1 = (0,0,1), \quad \vec{n}_2 = \left(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad \vec{n}_3 = \left(-\frac{\sqrt{3}}{2}, \frac{1}{2}, 0\right), \quad \vec{n}_4 = (-1,0,0)$$

with $c_1^2 = 11$, $c_2^2 = 10$, $c_3^2 = 9.9$, and $c_4^2 = 9.8$. Since these data satisfy none of (3.10)–(3.12), only one of $\vec{g}$ corresponds to each root of

$$p(x) = \frac{1}{160000}(14055561 - 6614520x + 1150900x^2 - 88000x^3 + 2500x^4).$$

This polynomial has four roots $\xi_1 \approx 6.2722$, $\xi_2 \approx 9.4936$, $\xi_3 \approx 9.6576$, and $\xi_4 \approx 9.7766$ in $(0, c_4^2]$, where $\tilde{c}_{66} = \xi_1$ satisfies (3.8), and $\tilde{c}_{66} = \xi_2$ satisfies (3.7), and $\tilde{c}_{66} = \xi_3$ and

$\tilde{c}_{66} = \xi_4$ satisfy (3.9). From Theorem 14, we get four solutions:

$$\begin{pmatrix} \tilde{c}_{66} \\ \tilde{c}_{44} \\ \vec{f} \end{pmatrix} \approx \begin{pmatrix} 6.2722 \\ 14.8371 \\ (-0.6418, 0.1900, 0.7430) \end{pmatrix}, \quad \begin{pmatrix} 9.4936 \\ 16.2959 \\ (0.2122, 0.8564, -0.4706) \end{pmatrix},$$

$$\begin{pmatrix} 9.6576 \\ 11.2520 \\ (-0.2989, 0.2622, -0.9176) \end{pmatrix}, \quad \begin{pmatrix} 9.7766 \\ 11.2149 \\ (-0.1275, 0.3649, -0.9223) \end{pmatrix}.$$

4. Numerical implementation. Here we indicate the success of the approach of using four data sets to solve the inverse problem. That is, find the triple $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$ from four propagating fronts, where the four normals and corresponding (estimated) wave speeds $\{(\vec{n}_j, c_j)\}_{j=1}^{4}$ are compatible; that is, the wave speeds are all different and any three normals are linearly independent (see Definition 6(a)).

Furthermore, since we develop our theory under the assumption that the medium properties may not be symmetric about the image plane, we calculate the three-dimensional wave front in the neighborhood of the image plane. Our supersonic excitations are assumed to be slightly out of the image plane to easily achieve the linear independence mentioned above, and we expect that this configuration could be realizable with a full planar array of transducers for three-dimensional imaging or three lines of closely spaced transducers in a so-called $2\frac{1}{2}$-dimensional imaging setting (see Figure 4.1(a)). For this synthetic data experiment we calculate the wave fronts using a first order anisotropic Eikonal solver based on fast marching methods with code developed at Rensselaer.

The successive supersonic imaging pushes to create the approximate line sources are made at a sweeping speed faster than the background shear wave speed and indicated by the multiple of the background shear wave speed (Mach number); hence the label *supersonic* (see [3]). The background wave speed is indicated in each of the labeled figures and also is given in our text description below. In our examples, the

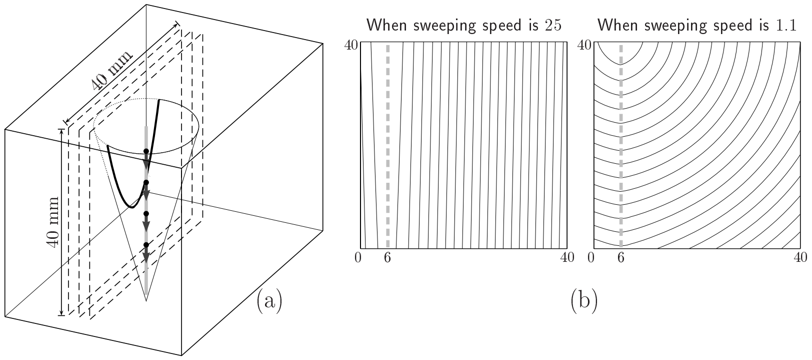


FIG. 4.1. (a) *Configuration: Data are collected on three consecutive image planes (dashed lines) by either a full planar array or three parallel, closely spaced linear arrays. Supersonic excitations are slightly off the imaging planes (gray line). A generated conical wave front yields parabolic intersections with each image plane. The shapes of parabola depend on the location and the sweeping speed of supersonic excitations.* (b) *Observed conical wave fronts on the central image plane, when the supersonic excitation line is* 8*mm away from the central image plane and* 6*mm away from the left side.*

pushes are either made slowly from top to bottom (1.1 sweeping speed), slowly from bottom to top (-1.1 sweeping speed), fast from top to bottom (25 sweeping speed), or fast from bottom to top (-25 sweeping speed). Each set of pushes produces a conical wave front in three dimensions whose intersection with the image plane is generally a parabola but looks like a straight line for high sweeping speeds like ± 25. See Figure 4.1(b).

We show two numerical reconstructions. For the first we have uniform anisotropy, where the fiber direction is out of the image plane; see Figure 4.2. The uniform anisotropic cube is 40 mm on a side with two excitation lines for the pushes, each being 6 mm from the outside edge; note that the excitation lines are at different distances from the image plane with one 8 mm from the image plane and the other 12 mm from the image plane. We take separately the two sweeping speeds, ± 1.1, yielding four propagating fronts. It is assumed that $\sqrt{\tilde{c}_{66}} = 1$, $\sqrt{\tilde{c}_{44}} = 2$. Setting up the three orthogonal coordinates with the x coordinate out of the plane, we show our results for $\sqrt{\tilde{c}_{44}}$, $\sqrt{\tilde{c}_{66}}$, the wave speeds along and across the fiber direction, respectively, and the squares of the fiber direction coordinates f_x^2, f_y^2, f_z^2. In addition we exhibit $\sqrt{\tilde{c}_{66}}$, $\sqrt{\tilde{c}_{44}}$, and $\{c_j\}_{j=1}^4$ along the line $z = 25$, $0 < y < 40$.

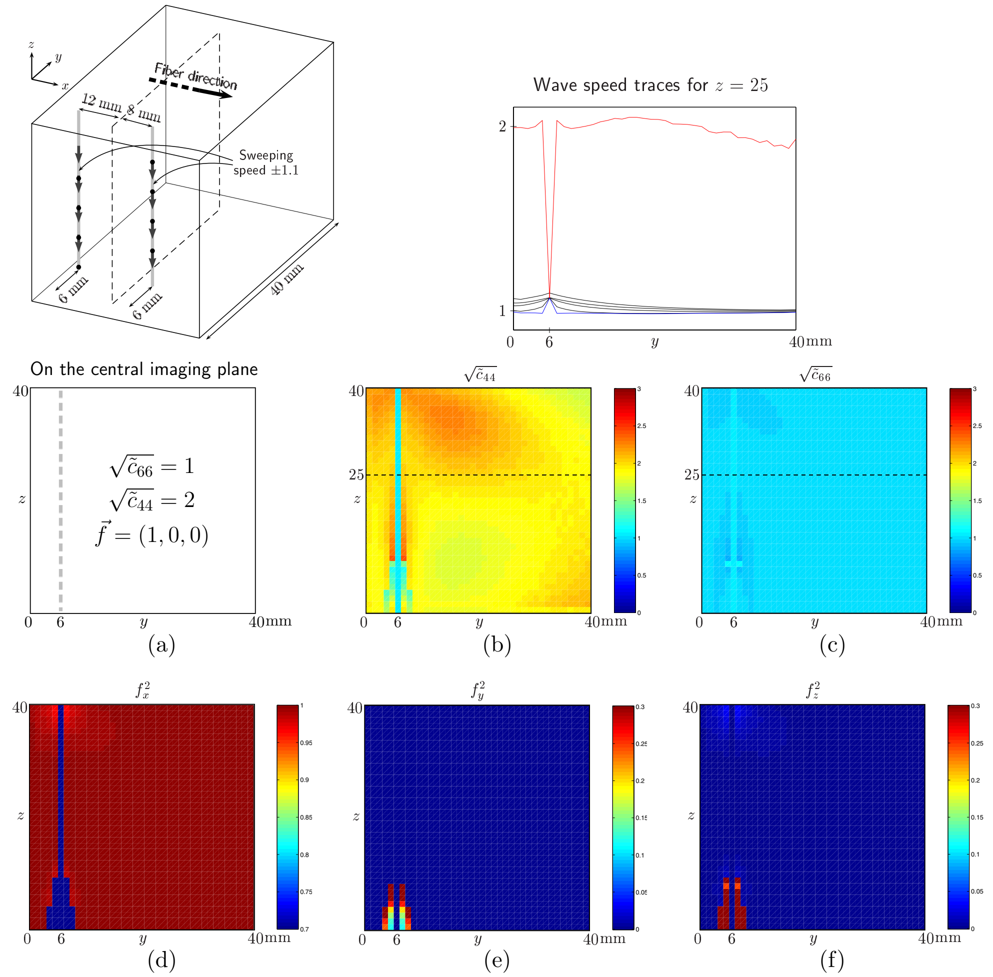
As we have seen in section 3, sometimes the fourth order polynomial for $\tilde{c}_{66}$ may have multiple roots in $(0, c_4^2]$, which is the source of our nonuniqueness. In this case, we have chosen to select the largest possible root in $(0, c_4^2]$, as in all of our simulations that choice consistently gave the correct recovery. Note that there are artifacts near the projections of the excitation lines, $y = 6$, onto the image plane because the fourth order polynomial is not well defined there (there four wave normals, $\vec{n}_j$, are on one plane perpendicular to our image plane, i.e., $\hat{D}_j = 0$, which yields $p(x) \equiv 0$). But otherwise the recovery is quite acceptable.

For our second simulation the excitation lines are in the same locations, but along one line we take the sweeping speeds ± 1.1, 25, and along the second line the sweeping speed is 25. Here the fiber is again out of the plane but only in the anisotropic cube inclusion with 10mm on each side. The anisotropic inclusion is embedded in an isotropic medium; see Figure 4.3. Again the recovery is quite acceptable; note that in all images of the material properties we observe anisotropic cube edge effects, except in the image for $\sqrt{\tilde{c}_{66}}$.

Here, also in the first simulation, the points where all four wave speeds $\{c_j\}_{j=1}^4$ are so close (using some threshold) are considered as isotropic points. We established a threshold, $\delta = 0.01$ for the first simulation and 0.04 for the second simulation, and consider the points isotropic when $\max\{c_j\}_{j=1}^4 - \min\{c_j\}_{j=1}^4 \leq \delta \max\{c_j\}_{j=1}^4$. For isotropic points, we assign a zero vector to the fiber direction and set $\sqrt{\tilde{c}_{44}} = \sqrt{\tilde{c}_{66}}$. As mentioned before, the fourth order polynomial is not well defined on the excitation lines, which stems from the fact that all four wave speeds are so close there. Because those points are considered as isotropic points, they are buried in the isotropic background in the second simulation, while in the first simulation the *isotropic* excitation line stands out in the anisotropic background. Compare the graphs near $y = 6$ in Figures 4.2 and 4.3.

5. Conclusion.

Here we address the following question: How do we obtain anisotropic medium properties from a set of wave fronts? Our target application is tissue shear stiffness imaging and we assume the medium is three-dimensional. There is a fiber direction along which the wave speed, $\sqrt{\tilde{c}_{44}}$, is faster than in the plane orthogonal to the fiber where the wave speed is $\sqrt{\tilde{c}_{66}}$ and directionally independent in that plane. We show that from four wave fronts, where any three normals at each



(a) (b) (c)

(d) (e) (f)

FIG. 4.2. *Top left: Orientation of fibers, central image plane (neighboring planes are omitted) and supersonic line sources.* (a) *Target wave speeds for uniform anisotropy, and fiber direction. Gray dashed line depicts the line source projections in the image plane.* (b) *Reconstructed wave speed along the fiber.* (c) *Reconstructed wave speed across the fiber. Top right: Graph shows the wave speed traces for $z = 25$, $0 < y < 40$, dashed lines in* (b) *and* (c)*: along the fiber (top dotted line), across the fiber (bottom dotted line), estimated wave speeds c_j in the directions $\vec{n}_j$ orthogonal to the four wave fronts (middle solid lines).* (d)–(f) *Squares of the fiber direction components.*

point are linearly independent, we can have up to four distinct triples $(\tilde{c}_{66}, \tilde{c}_{44}, \vec{f})$, where $\vec{f}$ is the unit fiber direction. We exhibit examples to show that multiple solutions can occur and show numerical reconstructions with synthetic data. The multiple solutions are a result of the nonlinearity in the Eikonal equation.

From our work to obtain reconstructions we have observed the importance of (1) having well-separated normals to the wave fronts, and that necessitates some normals having out of image plane components; (2) the need for multiple image planes to capture all three components of the normals; and (3) the fact that in a high contrast subregion embedded in a constant medium, initially well-separated normals may align themselves (the angle between their normals becomes smaller) at some points, and at other points the angle may become larger. This angle change may

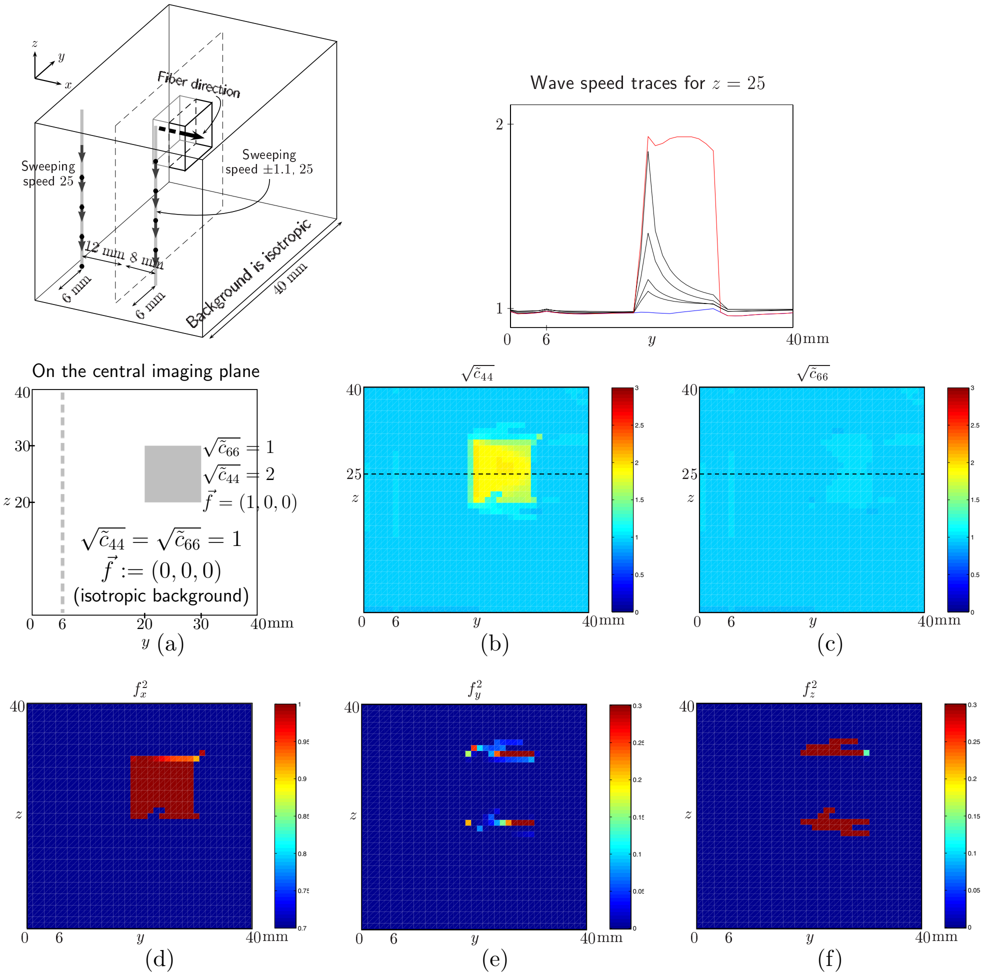


FIG. 4.3. *Top left: Orientation of fibers in the anisotropic cube inclusion and supersonic line sources.* (a) *Target wave speeds and fiber direction in the background and in the anisotropic cube.* (b) *Reconstructed wave speed along the fiber.* (c) *Reconstructed wave speed across the fiber. Top right: Graph shows the wave speed traces for $z = 25$, $0 < y < 40$, dashed lines in* (b) *and* (c)*: along the fiber (top dotted line), across the fiber (bottom dotted line), estimated wave speeds c_j in the directions $\vec n_j$ orthogonal to the four wave fronts (middle solid lines).* (d)–(f) *Squares of the fiber direction components.*

occur also at points beyond that subregion. The degree of this angle change depends on the wave speed contrast, size of inclusion, and the initial incident directions. This indicates important features in experimental design when wave fronts are used to image anisotropic properties of the kind modeled in this paper.

Acknowledgments. We have benefitted from discussions with Antoinette Maniatty, Maarten de Hoop, William Symes, Lizabeth Rachele, and Gunther Uhlmann.

REFERENCES

[1] J. Bercoff, S. Chaffai, M. Tanter, L. Sandrin, S. Catheline, M. Fink, J.-L. Gennisson, and M. Meunier, *In vivo breast tumor detection using transient elastography*, Ultrasound in Med. and Biol., 29 (2003), pp. 1387–1396.

[2] J. Bercoff, M. Tanter, S. Chaffai, and M. Fink, *Ultrafast imaging of beamformed shear waves induced by the acoustic radiation force: Application to transient elastography*, in Proceedings of the 2002 IEEE Ultrasonics Symposium, Vol. 2, IEEE, Piscataway, NJ, 2002, pp. 1899–1902.

[3] J. Bercoff, M. Tanter, and M. Fink, *Supersonic shear imaging: A new technique for soft tissue elasticity mapping*, IEEE Trans. Ultrasonics Ferroelectrics Frequency Control, 51 (2004), pp. 396–409.

[4] M. M. Doyley, P. M. Meaney, and J. C. Bamber, *Evaluation of an iterative reconstruction method for quantitative elasticity*, Phys. Med. Biol., 45 (2000), pp. 1521–1540.

[5] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, *Uniqueness and stability in the Cauchy problem for Maxwell and elasticity systems*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar 14, North–Holland, Amsterdam, 2002, pp. 329–349.

[6] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[7] F. Fatemi and J. F. Greenleaf, *Ultrasound-simulated vibro-acoustic spectography*, Science, 280 (1998), pp. 82–85.

[8] L. Gao, K. J. Parker, and S. K. Alam, *Sonoelasticity imaging: Theory and experimental verification*, J. Acoust. Soc. Amer., 97 (1995), pp. 3875–3880.

[9] J.-L. Gennisson, S. Catheline, S. Chaffai, and M. Fink, *Transient elastography in anisotropic medium: Application to the measurement of slow and fast shear wave speeds in muscles*, J. Acoust. Soc. Amer., 114 (2003), pp. 536–541.

[10] L. Ji, J. R. McLaughlin, D. Renzi, and J.-R. Yoon, *Interior elastodynamics inverse problems: Shear wave speed reconstruction in transient elastography*, Inverse Problems, 19 (2003), pp. S1–S29.

[11] J. R. McLaughlin and J.-R. Yoon, *Unique identifiability of elastic parameters from time dependent interior displacement measurement*, Inverse Problems, 20 (2004), pp. 25–45.

[12] J. R. McLaughlin and D. Renzi, *Shear wave speed recovery in transient elastography and supersonic imaging using propagating fronts*, Inverse Problems, 22 (2006), pp. 681–706.

[13] J. R. McLaughlin and D. Renzi, *Using level set based inversion of arrival times to recover shear wave speed in transient elastography and supersonic imaging*, Inverse Problems, 22 (2006), pp. 707–725.

[14] J. R. McLaughlin, D. Renzi, K. J. Parker, and Z. Wu, *Shear wave speed recovery using moving interference patterns obtained in sonoelastography experiments*, J. Acoust. Soc. Amer., 121 (2007), pp. 2438–2446.

[15] R. Muthpillai, D. J. Lomas, P. J. Rossman, J. F. Greenleaf, A. Manduca, and R. I. Ehman, *Magnetic resonance elastography by direct visualization of propagating acoustic strain waves*, Science, 269 (1995), pp. 1854–1857.

[16] K. R. Nightingale, J. L. Palmeri, R. W. Nightingale, and G. E. Trahey, *On the feasibility of remote palpation using acoustic radiation force*, J. Acoust. Soc. Amer., 110 (2001), pp. 625–634.

[17] J. Ophir, I. Cespedes, H. Ponnekanti, Y. Yazdi, and X. Li, *Elastography: A quantitative method for imaging the elasticity of biological tissues*, Ultrasonic Imaging, 13 (1991), pp. 111–134.

[18] R. Sinkus, J. Lorenzen, D. Schrader, M. Lorenzen, M. Dargatz, and D. Holz, *High-resolution tensor MR elastography for breast tumor detection*, Phys. Med. Biol., 4 (2000), pp. 1649–1664.

[19] R. Sinkus, M. Tanter, S. Catheline, J. Lorenzen, C. Kuhl, E. Sondermann, and M. Fink, *Imaging anisotropic and viscous properties of breast tissue by magnetic resonance-elastography*, Magnetic Resonance in Medicine, 53 (2005), pp. 372–387.

[20] L. S. Taylor, B. C. Porter, D. J. Rubens, and K. J. Parker, *Three-dimensional sonoelastography: Principles and practices*, Phys. Med. Biol., 45 (2000), pp. 1477–1494.

[21] J. Weaver, M. Doyley, E. Van Houten, M. Hood, X. C. Qin, F. Kennedy, S. Poplack, and K. Paulsen, *Evidence of the anisotropic nature of the mechanical properties of breast tissue*, Med. Phys., 29 (2002), p. 1291.

[22] Z. Wu, L. S. Taylor, D. J. Rubens, and K. J. Parker, *Sonoelastographic imaging of interference patterns for estimation of the shear velocity of homogeneous biomaterials*, Phys. Med. Biol., 49 (2004), pp. 911–922.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 43–58

THE ATOMIC MIX APPROXIMATION FOR CHARGED PARTICLE TRANSPORT[*]

EDWARD W. LARSEN[†] AND LIANG LIANG[†]

Abstract. The classic atomic mix approximation for particle transport in a stochastic spatial medium is accurate when the material chunks in the medium are small compared to a mean free path. In this paper, we show that for charged particle transport in a stochastic medium, the atomic mix approximation is accurate when the chunk sizes are small compared to a *transport* mean free path. For charged particle transport, the transport mean free path is generally several orders of magnitude larger than the mean free path. Therefore, the result obtained in this paper greatly extends the known range of applicability of the atomic mix approximation. Numerical results are given that validate the asymptotic theory, and an application of the theory to a practical problem in radiation oncology is discussed.

Key words. particle transport, random media, asymptotic analysis, homogenization

AMS subject classifications. 82C70, 78A48, 78M35, 78M40

DOI. 10.1137/060663015

1. Introduction.

The *atomic mix* approximation is a classic technique in physics and chemistry that has also been used for many years in the particle (radiation) transport community [1]. The underlying (particle transport) problems are described by a linear Boltzmann equation [1, 2, 3, 4], applied to a heterogeneous spatial medium consisting of randomly located "chunks" of two or more materials. If the chunk diameters are small compared to a typical mean free path and the chunks are distributed in a statistically uniform way throughout the system, the atomic mix approximation applies and the highly space-dependent cross sections can be replaced by their volume averages. The resulting approximate "atomic mix" Boltzmann equation, with volume-averaged cross sections, then accurately determines the radiation flux. The approximate atomic mix problem is much simpler than the original problem because (i) it is not necessary to know the detailed structure of the physical system (it is only necessary to know the cross sections and volume fractions of the constituent parts), and (ii) the atomic mix Boltzmann equation with volume-averaged cross sections is much easier to solve than the original Boltzmann equation, with highly space-dependent cross sections.

Recently, Dumas and Golse have proved that the atomic mix approximation is an asymptotic limit of the Boltzmann equation, for stochastic physical systems in which the chunk sizes are small compared to a typical mean free path [5]. (This is the physical regime in which the atomic mix approximation is commonly understood to hold.) More recently, Larsen [6] and Larsen, Vasques, and Vilhena [7] have shown—by a formal asymptotic analysis—that for 1-D (one-dimensional) diffusive stochastic systems the atomic mix approximation is valid when the chunk sizes are *comparable* to a mean free path.

[*] Received by the editors June 15, 2006; accepted for publication (in revised form) May 29, 2007; published electronically October 10, 2007. Part of this work was performed under U.S. Government Contract DE-AC52-06NA25396 for Los Alamos National Laboratory, which is operated by the Los Alamos National Security (LANS) LLC for the U.S. Department of Energy.

http://www.siam.org/journals/siap/68-1/66301

[†] Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, MI 48109-2104 (edlarsen@umich.edu, lliang@umich.edu).

In the present paper, we consider 3-D charged particle transport problems that are dominated by "soft" collision processes in which the particles experience very small changes in direction and energy. We show that for such problems the atomic mix approximation is a formal asymptotic limit of the linear Boltzmann equation in a stochastic medium in which the chunk sizes are small compared to a *transport* mean free path. For charged particle transport, the transport mean free path is usually several orders of magnitude greater than a mean free path. Therefore, the result obtained in this paper greatly extends the known range of applicability of the atomic mix approximation.

Our theoretical approach employs two different asymptotic limits. First, we use an asymptotic approximation developed by Pomraning [8]—valid when the mean free path is small and particles experience very small changes in the direction of flight and energy in a collision—to approximate the soft collision operator by its Fokker–Planck limit [9]. This reduces the original linear Boltzmann equation to a Boltzmann–Fokker–Planck (BFP) equation [10, 11]. Then we apply a generalization of the asymptotic analysis of Dumas and Golse to the BFP equation to show that when chunk sizes are small compared to a transport mean free path, the BFP equation limits to its atomic mix approximation. The resulting atomic mix BFP equation is identical to the equation obtained by (i) formally replacing the original Boltzmann equation by its atomic mix approximation and (ii) applying Pomraning's Fokker–Planck approximation to the resulting atomic mix soft collision operator.

Therefore, the atomic mix BFP equation is an asymptotic limit of *both* the original linear Boltzmann equation *and* its atomic mix approximation. This implies that, for charged particle transport problems in a stochastic medium in which (i) soft collisions dominate hard collisions and (ii) a typical chunk size within the medium is small compared to a transport mean free path, the atomic mix model is an asymptotic approximation to the linear Boltzmann equation.

The remainder of this paper is organized as follows. In section 2 we introduce the original Boltzmann equation and present our formal asymptotic analysis. To validate the predictions of the asymptotic theory, we present in section 3 the results of realistic Monte Carlo simulations of electron beams penetrating random binary systems of water and air. In work presented elsewhere [12], we have used the results in this paper to develop a practical computer model of the human lung, in order to assess the accuracy of certain treatment planning techniques in radiation oncology. This application of the asymptotic theory is discussed in the concluding section 4 of the present paper.

2. Asymptotic analysis.

We consider the following particle transport problem:

$$\begin{aligned}
&\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)+\Sigma(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\Sigma(\boldsymbol{x},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\Psi(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE',\quad \boldsymbol{x}\in V, \qquad (2.1a)
\end{aligned}$$

$$\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)=\Psi^b(\boldsymbol{x},\boldsymbol{\Omega},E),\quad \boldsymbol{x}\in\partial V,\ \ \boldsymbol{\Omega}\cdot\boldsymbol{n}<0. \qquad (2.1b)$$

Our notation is standard:

$$\boldsymbol{x}=(x,y,z)=\text{ position}, \qquad (2.2a)$$

$$\boldsymbol{\Omega}=(\sqrt{1-\mu^2}\cos\gamma,\sqrt{1-\mu^2}\sin\gamma,\mu)=\text{ direction of flight}, \qquad (2.2b)$$

$$E=\text{ energy}, \qquad (2.2c)$$

and

$$\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) = \text{ angular flux (intensity)}, \tag{2.3a}$$

$$\Sigma(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E) = \text{ differential scattering cross section}, \tag{2.3b}$$

$$\begin{aligned}\Sigma(\boldsymbol{x},E) &= \int_0^\infty \int_{4\pi} \Sigma(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\,d\Omega' dE' \\ &= \text{ scattering cross section.}\end{aligned} \tag{2.3c}$$

Problem (2.1) describes a particle transport process within a physical system V. The process is driven by a specified incident angular flux Ψ^b on the outer surface ∂V of V. V is spatially heterogeneous, consisting of a large number of chunks of two or more materials. Also, the scattering process in V is dominated by soft collisions (in which particles experience very small changes in direction of flight and energy), but rare hard collisions (in which the changes in direction of flight and energy are not small) can also occur. To separate these two types of scattering events, we use $\mu_0 = \boldsymbol{\Omega}'\cdot\boldsymbol{\Omega}$ = scattering cosine and write

$$\Sigma(\boldsymbol{x},\mu_0,E'\to E) = \Sigma_h(\boldsymbol{x},\mu_0,E'\to E) + \Sigma_r(\boldsymbol{x},\mu_0,E'\to E), \tag{2.4}$$

where Σ_h = differential scattering cross section for hard collisions and Σ_r = differential scattering cross section for soft collisions (also called the "restricted" differential scattering cross section). We define

$$\begin{aligned}\Sigma_h(\boldsymbol{x},E) &= \int_0^\infty \int_{4\pi} \Sigma(\boldsymbol{x},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E\to E')\,d\Omega' dE' \\ &= \text{ hard scattering cross section,}\end{aligned} \tag{2.5a}$$

$$\begin{aligned}\Sigma_r(\boldsymbol{x},E) &= \int_0^\infty \int_{4\pi} \Sigma(\boldsymbol{x},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E\to E')\,d\Omega' dE' \\ &= \text{ soft (restricted) scattering cross section,}\end{aligned} \tag{2.5b}$$

and clearly,

$$\Sigma(\boldsymbol{x},E) = \Sigma_h(\boldsymbol{x},E) + \Sigma_r(\boldsymbol{x},E). \tag{2.5c}$$

We also define the *phase function* for hard collisions:

$$p_h(\boldsymbol{x},\mu_0,E'\to E) = \frac{\Sigma_h(\boldsymbol{x},\mu_0,E'\to E)}{\Sigma_h(\boldsymbol{x},E')}\ , \tag{2.6a}$$

which by (2.5a) and (2.6a) satisfies

$$\int_0^\infty \int_{4\pi} p_h(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\,d\Omega dE = 1. \tag{2.6b}$$

Introducing (2.4)–(2.6) into (2.1a), we obtain

$$\begin{aligned}\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) &+ \Sigma_h(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) \\ &= \int_0^\infty \int_{4\pi} \Sigma_h(\boldsymbol{x},E')p_h(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\Psi(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE' \\ &\qquad + L_r\Psi(\boldsymbol{x},\boldsymbol{\Omega},E),\end{aligned} \tag{2.7a}$$

where L_r is the *restricted collision operator:*

$$\begin{aligned} L_r\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) = \int_0^\infty \int_{4\pi} &\Sigma_r(\boldsymbol{x},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\Psi(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE' \\ &-\Sigma_r(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E). \end{aligned} \tag{2.7b}$$

To this point we have made no approximations; (2.7) are equivalent to the original Boltzmann equation (2.1a).

Now we make our first approximation. Because soft collisions generate very small changes in direction of flight and energy, the restricted (soft) differential scattering cross section $\Sigma_t(\boldsymbol{x},\mu_0,E'\to E)$ is very highly peaked near $\mu_0\approx 1$ and $E'\approx E$. In this situation, Pomraning [8] has shown that L_r is asymptotically approximated by the Fokker–Planck operator:

$$\begin{aligned} L_r\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) &\approx L_{FP}\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) \\ &= \frac{\Sigma_{r,tr}(\boldsymbol{x},E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) \\ &\quad + \frac{\partial}{\partial E}S_r(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E), \end{aligned} \tag{2.8}$$

where

$$\begin{aligned} \Sigma_{r,tr}(\boldsymbol{x},E) &= \int_0^\infty\int_{4\pi}(1-\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega})\Sigma_r(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E\to E')\,d\Omega' dE' \\ &= \text{ restricted transport cross section,} \end{aligned} \tag{2.9a}$$

$$\begin{aligned} S_r(\boldsymbol{x},E) &= \int_0^\infty\int_{4\pi}(E-E')\Sigma_r(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E\to E')\,d\Omega' dE' \\ &= \text{ restricted stopping power.} \end{aligned} \tag{2.9b}$$

Introducing (2.8)–(2.9) into (2.7a) and using the boundary condition (2.1b), we obtain the following BFP problem:

$$\begin{aligned} \boldsymbol{\Omega}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) &+ \Sigma_h(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) \\ &= \int_0^\infty\int_{4\pi}\Sigma_h(\boldsymbol{x},E')p_h(\boldsymbol{x},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\Psi(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE' \\ &\quad + \frac{\Sigma_{r,tr}(\boldsymbol{x},E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) \\ &\quad + \frac{\partial}{\partial E}S_r(\boldsymbol{x},E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E), \quad \boldsymbol{x}\in V, \end{aligned} \tag{2.10a}$$

$$\Psi(\boldsymbol{x},\boldsymbol{\Omega},E) = \Psi^b(\boldsymbol{x},\boldsymbol{\Omega},E), \quad \boldsymbol{x}\in\partial V, \ \ \boldsymbol{\Omega}\cdot\boldsymbol{n}<0. \tag{2.10b}$$

The BFP equation is well known in the literature [10, 11]. It provides an accurate way to simulate transport problems in which soft collisions dominate but rare hard collisions can also occur. The advantage of the BFP equation (2.10a) over the Boltzmann equation (2.1a) is that the BFP equation contains neither a large scattering cross section nor a highly peaked differential scattering cross section. (Hard collisions, which produce large changes of direction and energy loss, have a relatively smooth differential scattering cross section.) Also, because the soft differential scattering cross section is highly peaked near $\mu_0 \approx 1$, (2.9a) and (2.5b) imply

$$\Sigma_{r,tr}(\boldsymbol{x}, E) \ll \Sigma_r(\boldsymbol{x}, E) < \Sigma(\boldsymbol{x}, E). \tag{2.11}$$

Here Σ^{-1} is the mean free path (the mean distance between collisions), and $\Sigma_{r,tr}^{-1}$ is the (restricted) transport mean free path (the mean distance a particle must travel for its direction of flight to change by an $O(1)$ amount). Equation (2.11) implies that the restricted transport mean free path is much larger than the mean free path.

The restricted stopping power $S_r(\boldsymbol{x}, E)$ has units of MeV/cm and the interpretation

$$\begin{aligned} S_r(\boldsymbol{x}, E)ds = {} & \text{the energy loss that a particle at } (\boldsymbol{x}, E) \text{ experiences} \\ & \text{through soft collisions while traveling a distance } ds. \end{aligned} \tag{2.12a}$$

Therefore, the function

$$T(\boldsymbol{x}, E) = \frac{S_r(\boldsymbol{x}, E)}{E} \tag{2.12b}$$

has units of cm^{-1} and the interpretation

$$\begin{aligned} T(\boldsymbol{x}, E)ds = {} & \text{the fractional energy loss that a particle at } (\boldsymbol{x}, E) \text{ experiences} \\ & \text{through soft collisions while traveling a distance } ds. \end{aligned} \tag{2.12c}$$

Now we write the functions $\Sigma_h(\boldsymbol{x}, E)$, $\Sigma_{r,tr}(\boldsymbol{x}, E)$, and $T(\boldsymbol{x}, E)$ in a useful dimensionless form. These functions are highly space-dependent, due to the assumption that the physical system V consists of a large number of "chunks" of two or more materials. We define a characteristic length λ_{ch} by

$$\lambda_{ch} = \text{ typical width of a chunk in } V, \tag{2.13}$$

and we introduce the dimensionless spatial variable

$$\boldsymbol{y} = \frac{\boldsymbol{x}}{\lambda_{ch}} \ . \tag{2.14}$$

In terms of $\boldsymbol{y}$, a typical chunk width is $O(1)$.

We also define the characteristic lengths λ_h, $\lambda_{r,tr}$, and λ_r by

$$\frac{1}{\lambda_h} = \text{ typical value of } \Sigma_h(\boldsymbol{x}, E), \tag{2.15a}$$

$$\frac{1}{\lambda_{r,tr}} = \text{ typical value of } \Sigma_{r,tr}(\boldsymbol{x}, E), \tag{2.15b}$$

$$\frac{1}{\lambda_r} = \text{ typical value of } T(\boldsymbol{x}, E). \tag{2.15c}$$

These characteristic lengths have straightforward physical interpretations: λ_h is the typical distance a particle must travel to experience a hard collision; $\lambda_{r,tr}$ is the typical distance a particle must travel for its direction of flight to be altered through soft collisions only by an $O(1)$ amount; and λ_r is the typical distance that a particle must travel to lose an $O(1)$ fraction of its energy through soft collisions. We make the following assumptions:

$$\text{A1}: \qquad \frac{\lambda_h}{\lambda_{r,tr}} = O(1), \quad \frac{\lambda_r}{\lambda_{r,tr}} = O(1), \qquad \varepsilon \equiv \frac{\lambda_{ch}}{\lambda_{r,tr}} \ll 1. \tag{2.16}$$

Thus, λ_h, $\lambda_{r,tr}$, and λ_r are comparable to each other and large compared to λ_{ch}. Physically, this implies that a typical chunk width is small (O(ε)) compared to the distances over which the effect of hard collisions, soft angular deflections, and soft energy loss are $O(1)$. Alternatively, when a particle travels across a typical chunk, the effects of hard collisions, soft angular deflections, and soft energy loss are small ($O(\varepsilon)$).

Since λ_h, $\lambda_{r,tr}$, and λ_r are comparable, the dimensionless functions

$$\sigma_h(\boldsymbol{y}, E) \equiv \lambda_{r,tr}\Sigma_h(\lambda_{ch}\boldsymbol{y}, E) = \lambda_{r,tr}\Sigma_h(\boldsymbol{x}, E), \tag{2.17a}$$

$$\sigma_{r,tr}(\boldsymbol{y}, E) \equiv \lambda_{r,tr}\Sigma_{r,tr}(\lambda_{ch}\boldsymbol{y}, E) = \lambda_{r,tr}\Sigma_{r,tr}(\boldsymbol{x}, E), \tag{2.17b}$$

$$t(\boldsymbol{y}, E) \equiv \lambda_{r,tr}T(\lambda_{ch}\boldsymbol{y}, E) = \lambda_{r,tr}T(\boldsymbol{x}, E) = \lambda_{r,tr}\frac{S(\boldsymbol{x}, E)}{E} \tag{2.17c}$$

are $O(1)$ in magnitude and vary by $O(1)$ amounts when $\boldsymbol{y}$ varies by an $O(1)$ amount. Introducing (2.17) into the BFP equation (2.10a), we obtain

$$\begin{aligned}
&\boldsymbol{\Omega} \cdot \boldsymbol{\nabla}\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E) + \frac{1}{\lambda_{r,tr}}\sigma_h(\boldsymbol{y}, E)\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E) \\
&\quad = \frac{1}{\lambda_{r,tr}}\int_0^\infty \int_{4\pi} \sigma_h(\boldsymbol{y}, E')p_h(\boldsymbol{y}, \boldsymbol{\Omega} \cdot \boldsymbol{\Omega}', E' \to E)\Psi(\boldsymbol{x}, \boldsymbol{\Omega}', E')\, d\Omega' dE' \\
&\qquad + \frac{\sigma_{r,tr}(\boldsymbol{y}, E)}{2\lambda_{r,tr}}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu} + \frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E) \\
&\qquad + \frac{1}{\lambda_{r,tr}}\frac{\partial}{\partial E}Et(\boldsymbol{y}, E)\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E).
\end{aligned} \tag{2.18}$$

Now we express Ψ in terms of dimensionless spatial variables. Two fundamental length scales are present in (2.18): λ_{ch} = a typical chunk width, and $\lambda_{r,tr}$ = a typical restricted transport cross section. In (2.14), we defined the "fast" dimensionless spatial variable $\boldsymbol{y} = \boldsymbol{x}/\lambda_{ch}$, which describes $O(1)$ variations in the problem data that take place over the length scale of a chunk width. We now define the "slow" dimensionless spatial variable

$$\boldsymbol{z} = \frac{\boldsymbol{x}}{\lambda_{r,tr}}, \tag{2.19}$$

which describes $O(1)$ variations in Ψ that take place on the length scale of a restricted transport mean free path, and we assume that Ψ is a function of both $\boldsymbol{y}$ and $\boldsymbol{z}$:

$$\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E) = \psi(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Omega}, E). \tag{2.20a}$$

Then

$$\boldsymbol{\Omega} \cdot \boldsymbol{\nabla}\Psi(\boldsymbol{x}, \boldsymbol{\Omega}, E) = \boldsymbol{\Omega} \cdot \boldsymbol{\nabla}_y\psi(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Omega}, E)\frac{1}{\lambda_{ch}} + \boldsymbol{\Omega} \cdot \boldsymbol{\nabla}_z\psi(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Omega}, E)\frac{1}{\lambda_{r,tr}}\ . \tag{2.20b}$$

Introducing (2.20) into (2.18), multiplying by $\lambda_{r,tr}$, and using ε defined in (2.16), we obtain

$$\begin{aligned}
&\frac{1}{\varepsilon}\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_y\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)+\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_z\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)+\sigma_h(\boldsymbol{y},E)\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\sigma_h(\boldsymbol{y},E')p_h(\boldsymbol{y},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega}',E')\,d\Omega' dE'\\
&\qquad+\frac{\sigma_{r,tr}(\boldsymbol{y},E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}Et(\boldsymbol{y},E)\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E).
\end{aligned}\tag{2.21}$$

Equation (2.21) is mathematically equivalent to the BFP equation (2.10).

Now we derive the leading-order term in a formal asymptotic solution of (2.21) for $\varepsilon \ll 1$. To do this, we assume that $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent spatial variables, we introduce the ansatz

$$\psi(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)=\sum_{n=0}^{\infty}\varepsilon^n\psi_n(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)\tag{2.22}$$

into (2.21), and we equate the coefficients of different powers of ε.

The first $[O(1/\varepsilon)]$ equation in the resulting asymptotic hierarchy is

$$\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_y\psi_0(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)=0,\tag{2.23}$$

which requires that ψ_0 be independent of the fast spatial variable $\boldsymbol{y}$ in the direction of flight $\boldsymbol{\Omega}$. The general solution of (2.23) is

$$\psi_0(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)=\tilde{\psi}_0[\boldsymbol{y}-(\boldsymbol{y}\cdot\boldsymbol{\Omega})\boldsymbol{\Omega},\boldsymbol{z},\boldsymbol{\Omega},E],\tag{2.24}$$

where $\tilde{\psi}_0$ is arbitrary. Thus, the leading-order term ψ_0 in the asymptotic expansion can exhibit fast spatial dependence in directions orthogonal to $\boldsymbol{\Omega}$. A less general solution of (2.23) is

$$\psi_0(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)=\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E),\tag{2.25}$$

which has *no* fast spatial variation. In the following, we assume that (2.25) holds, rather than the more general (2.24), and we systematically derive an equation for $\hat{\psi}_0$ in (2.25). To accomplish this we make an additional assumption A2, stated below, which places a condition on the randomness of the media and is consistent with (2.25). If A2 is not satisfied, then there may be circumstances in which the more general (2.24) might hold. We discuss this in more detail in the final paragraph of this section.

Assuming that ψ_0 is given by (2.25), where $\hat{\psi}_0$ is undetermined, the next $(O(1))$ equation in the asymptotic hierarchy becomes

$$\begin{aligned}
&\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_y\psi_1(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\Omega},E)+\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_z\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)+\sigma_h(\boldsymbol{y},E)\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\sigma_h(\boldsymbol{y},E')p_h(\boldsymbol{y},\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega}',E')\,d\Omega' dE'\\
&\qquad+\frac{\sigma_{r,tr}(\boldsymbol{y},E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}Et(\boldsymbol{y},E)\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E).
\end{aligned}\tag{2.26}$$

Integrating this equation over the line $\boldsymbol{y}+s\boldsymbol{\Omega}$, $-R \le s \le R$, using

$$\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_y\psi_1(\boldsymbol{y}+s\boldsymbol{\Omega},\boldsymbol{z},\boldsymbol{\Omega},E) = \frac{d}{ds}\psi_1(\boldsymbol{y}+s\boldsymbol{\Omega},\boldsymbol{z},\boldsymbol{\Omega},E),$$

and then dividing by $2R$, we obtain

$$\begin{aligned}
&\frac{1}{2R}\left[\psi_1(\boldsymbol{y}+R\boldsymbol{\Omega},\boldsymbol{z},\boldsymbol{\Omega},E)-\psi_1(\boldsymbol{y}-R\boldsymbol{\Omega},\boldsymbol{z},\boldsymbol{\Omega},E)\right]\\
&\quad+\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_z\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)+\left[\frac{1}{2R}\int_{-R}^{R}\sigma_h(\boldsymbol{y}+s\boldsymbol{\Omega},E)\,ds\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\left[\frac{1}{2R}\int_{-R}^{R}\sigma_hp_h(\boldsymbol{y}+s\boldsymbol{\Omega},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\,ds\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega}',E')\,d\Omega'dE'\\
&\qquad+\frac{1}{2}\left[\frac{1}{2R}\int_{-R}^{R}\sigma_{r,tr}(\boldsymbol{y}+s\boldsymbol{\Omega},E)\,ds\right]\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}E\left[\frac{1}{2R}\int_{-R}^{R}t(\boldsymbol{y}+s\boldsymbol{\Omega},E)\,ds\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E).
\end{aligned}\tag{2.27}$$

Now we let $R\to\infty$. Using the assumption that ψ_1 is bounded, and introducing the notation

$$\langle f\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y}) = \lim_{R\to\infty}\frac{1}{2R}\int_{-R}^{R}f(\boldsymbol{y}+s\boldsymbol{\Omega})\,ds,\tag{2.28}$$

which denotes averaging over the infinite line passing through the point $\boldsymbol{y}$ in the direction $\boldsymbol{\Omega}$, we obtain

$$\begin{aligned}
&\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_z\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)+\langle\sigma_h\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y},E)\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\langle\sigma_hp_h\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega}',E')\,d\Omega'dE'\\
&\qquad+\frac{1}{2}\langle\sigma_{r,tr}\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y},E)\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}E\langle t\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y},E)\,\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E).
\end{aligned}\tag{2.29}$$

Now we make another fundamental assumption:

(2.30) A2 : *Each line-averaged quantity in* (2.29) *equals its volume average.*

More explicitly, if we define the volume average

$$\overline{\sigma}_h(E) = \lim_{R\to\infty}\frac{3}{4\pi R^2}\int_{|\boldsymbol{y}|\le R}\sigma_h(\boldsymbol{y},E)\,dy,\tag{2.31}$$

then we assume

$$\langle\sigma_h\rangle_{\boldsymbol{\Omega}}(\boldsymbol{y},E) = \overline{\sigma}_h(E),\tag{2.32}$$

and similarly for the other line averages in (2.29). Assumption A2 places constraints on the uniformity and isotropicity of the randomness in the physical system. These constraints are not true for all heterogeneous systems. For example, they are not true for certain $\boldsymbol{\Omega}$ for crystalline systems in which the spatial heterogeneities are small but spatially periodic. However, for the randomized media considered in this paper, it is reasonable to assume that A2 is satisfied. Equation (2.29) then becomes

$$\begin{aligned}
&\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_z\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)+\overline{\sigma}_h(E)\,\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\overline{\sigma_h p_h}\,(\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\,\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega}',E')\,d\Omega' dE'\\
&\qquad+\frac{\overline{\sigma}_{r,tr}(E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}E\overline{t}(E)\,\hat{\psi}_0(\boldsymbol{z},\boldsymbol{\Omega},E).
\end{aligned}\tag{2.33}$$

Finally, we convert (2.33) back to the original independent variables. We define

$$\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)=\hat{\psi}_0\left(\frac{\boldsymbol{x}}{\lambda_{r,tr}},\boldsymbol{\Omega},E\right),\tag{2.34}$$

and we use (2.17) to obtain

$$\frac{1}{\lambda_{r,tr}}\,\overline{\sigma}_h(E)=\overline{\Sigma}_h(E)=\text{ volume average of }\Sigma_h(\boldsymbol{x},E),\tag{2.35a}$$

$$\begin{aligned}
\frac{1}{\lambda_{r,tr}}\,\overline{\sigma_h p_h}(\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)&=\overline{\Sigma}_h(\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\\
&=\text{ volume average of }\Sigma_h(\boldsymbol{x},\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E),
\end{aligned}\tag{2.35b}$$

$$\frac{1}{\lambda_{r,tr}}\,\overline{\sigma}_{r,tr}(E)=\overline{\Sigma}_{r,tr}(E)=\text{ volume average of }\Sigma_{r,tr}(\boldsymbol{x},E),\tag{2.35c}$$

$$\frac{1}{\lambda_{r,tr}}\,E\overline{t}(E)=\overline{S}_r(E)=\text{ volume average of }S_r(\boldsymbol{x},E).\tag{2.35d}$$

Introducing (2.34) and (2.35) into (2.33), we obtain that the solution $\psi(\boldsymbol{x},\boldsymbol{\Omega},E)$ of (2.1) has the asymptotic approximation

$$\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)=\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)+O(\varepsilon),\tag{2.36}$$

where $\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)$ satisfies the BFP problem:

$$\begin{aligned}
&\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}_x\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)+\overline{\Sigma}_h(E)\,\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)\\
&\quad=\int_0^\infty\int_{4\pi}\overline{\Sigma}_h\,(\boldsymbol{\Omega}'\cdot\boldsymbol{\Omega},E'\to E)\,\Psi_0(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE'\\
&\qquad+\frac{\overline{\Sigma}_{r,tr}(E)}{2}\left[\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial^2}{\partial\gamma^2}\right]\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)\\
&\qquad+\frac{\partial}{\partial E}\overline{S}_r(E)\,\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E),
\end{aligned}\tag{2.37a}$$

$$\Psi_0(\boldsymbol{x},\boldsymbol{\Omega},E)=\Psi^b(\boldsymbol{x},\boldsymbol{\Omega},E),\quad \boldsymbol{x}\in\partial V,\ \ \boldsymbol{\Omega}\cdot\boldsymbol{n}<0.\tag{2.37b}$$

Equations (2.37) are the atomic mix approximation to the BFP equation (2.10), with the boundary condition (2.1b). We note that (2.37b) is the same as the original boundary condition (2.1b). This is because we tacitly assumed that the prescribed incident flux Ψ^b varies spatially on the slow spatial scale of $\lambda_{r,tr}$, not on the fast spatial scale of λ_{ch}.

To summarize: we have formally shown that by separating the hard and soft collision operators, and requiring the soft collision operator to be sufficiently peaked for $\mu_0 \approx 1$ and $E \approx E'$, one can asymptotically approximate the Boltzmann equations (2.1) by the BFP equations (2.10). Also, if (A1) the characteristic lengths defined by (2.13) and (2.15) satisfy (2.16), and (A2) the average of the problem data over any line equals the volume average, then the BFP equations (2.10) are asymptotically approximated by the atomic mix BFP equations (2.37).

Now let us formally consider the atomic mix model of (2.1):

$$\begin{aligned} &\boldsymbol{\Omega}\cdot\boldsymbol{\nabla}\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)+\overline{\Sigma}(E)\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)\\ &\quad=\int_0^\infty\int_{4\pi}\overline{\Sigma}(\boldsymbol{\Omega}\cdot\boldsymbol{\Omega}',E'\to E)\Psi(\boldsymbol{x},\boldsymbol{\Omega}',E')\,d\Omega' dE',\quad \boldsymbol{x}\in V, \end{aligned} \tag{2.38a}$$

$$\Psi(\boldsymbol{x},\boldsymbol{\Omega},E)=\Psi^b(\boldsymbol{x},\boldsymbol{\Omega},E),\quad \boldsymbol{x}\in\partial V,\ \boldsymbol{\Omega}\cdot\boldsymbol{n}<0. \tag{2.38b}$$

If we split the scattering operator in (2.38a) into the hard and soft collision operators and asymptotically apply the Fokker–Planck approximation, just as was done to (2.1a), we will obtain exactly (2.37a).

Therefore, under the assumptions described above, (2.37) is an asymptotic limit of *both* the Boltzmann equations (2.1) *and* their atomic mix version, (2.38). This implies that the atomic mix (2.38) asymptotically approximates the original Boltzmann equations (2.1). This is the main result of this paper.

We note that (2.33) is consistent with the assumption (2.25) that the leading-order term in the asymptotic expansion is independent of the fast spatial variable $\boldsymbol{y}$. (None of the cross sections in (2.33) depend on $\boldsymbol{y}$. This follows from assumption A2, (2.31), that any line average of each cross section must equal its volume average.) If for a specified problem assumption A2 is not valid, then the cross sections in (2.28) depend on $\boldsymbol{y}$, and it seems inevitable that $\hat{\psi}_0$ will also depend on $\boldsymbol{y}$. In this case, the much more complicated (2.24) should be used as the solution of (2.23). We will not consider this here because the application that we intend for the preceding asymptotic theory satisfies A2.

3. Numerical results. To test the asymptotic theory, we have devised and run a set of computer experiments using the Monte Carlo code PENELOPE [13]. We consider a sequence of 6.0 cm deep targets consisting of small droplets of water randomly mixed in air. The volume fraction occupied by the water droplets is 0.201. In the first set of experiments, a circular (radius = 1.0 cm) 2.0 MeV monodirectional electron beam is normally incident on these targets. The electrons enter the targets and deposit energy, slowing down to 0.1 MeV, at which point their remaining energy is deposited locally. The asymptotic theory predicts that as the water droplets decrease in size, the dose deposited by the electron beam will limit to the dose deposited in the homogenized (atomic mix) target.

To predict the size of the water droplets for which the atomic mix approximation becomes accurate, let us consider Figure 1, which provides data for electrons with energies between 0.1 MeV and 10.0 MeV in water. (All the data shown in Figure 1 was taken from PENELOPE.) In this figure, the quantities $\Sigma_{r,tr}^{-1}(E)$ = transport mean free path, $T^{-1}(E) = E/S(E)$ = *restricted range* (the distance a particle with energy E would travel while slowing down through soft collisions to zero energy if its stopping power while slowing down were equal to $S(E)$), $\Sigma_h^{-1}(E)$ = hard mean

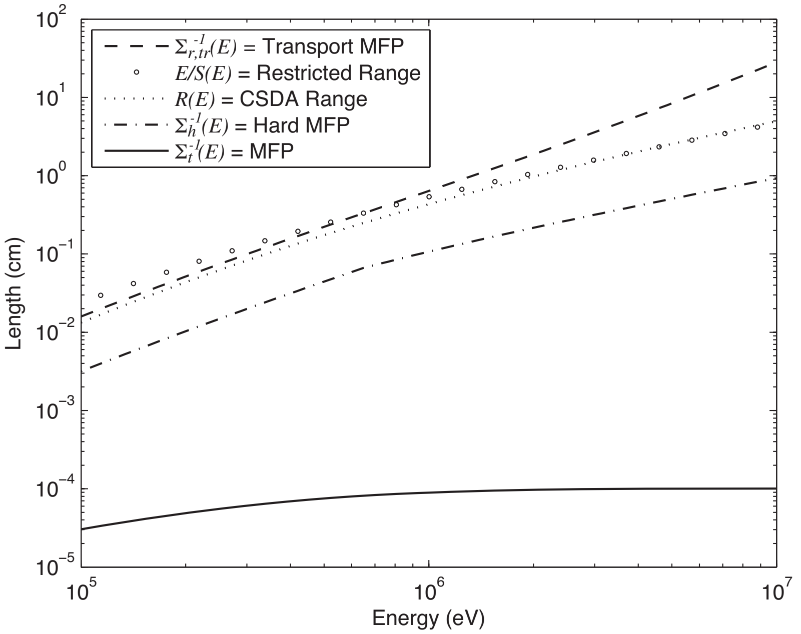


FIG. 1. *Transport MFP, restricted range, CSDA range, hard MFP, and MFP of electrons in water.*

free path, and $\Sigma_t^{-1}(E)$ = mean free path are plotted for water at normal density (1.0 gm/cm^{-1}). Also plotted in Figure 1 is the *continuous slowing down approximation (CSDA) range* $R(E)$, defined by

$$R(E) = \int_0^E \frac{dE'}{S(E')}; \tag{3.1}$$

this is the pathlength that a particle with energy E will travel while slowing down through soft collisions to zero energy. $R(E)$ is included in Figure 1 because it provides an alternate way to measure the distance in which O(1) changes in energy occur. Figure 1 shows that $R(E)$ and $T^{-1}(E)$ are comparable, and that there is a clear separation between the electron mean free path and the more "macroscopic" transport and hard mean free paths.

The asymptotic theory predicts that the atomic mix approximation is accurate if the chunk sizes lie below the transport mean free path (MFP), restricted range, and hard MFP curves in Figure 1. Interpreting this literally would place an upper limit on the chunk size of about $\lambda_{ch} = 3 \times 10^{-3}$ cm. However, the definition of the hard mean free path is somewhat arbitrary. (This definition depends on parameters chosen to run the PENELOPE simulation; the distinction between hard and soft collisions is not well defined.) Also, for most of the energy range of the electrons, the transport MFP, CSDA range, and hard MFP lie well above $\lambda_{ch} = .003$ cm. A less conservative upper bound on the chunk size is $\lambda_{ch} = .01$ cm, which is just below the minimum values of $\Sigma_{r,tr}^{-1}(E)$ and $T^{-1}(E)$.

For the given electron beam, we ran a series of simulations in which the (uniform) diameter of the droplets (λ_{ch}) was set to $\lambda_{ch} = 0.1, 0.05$, and 0.01 cm. We also

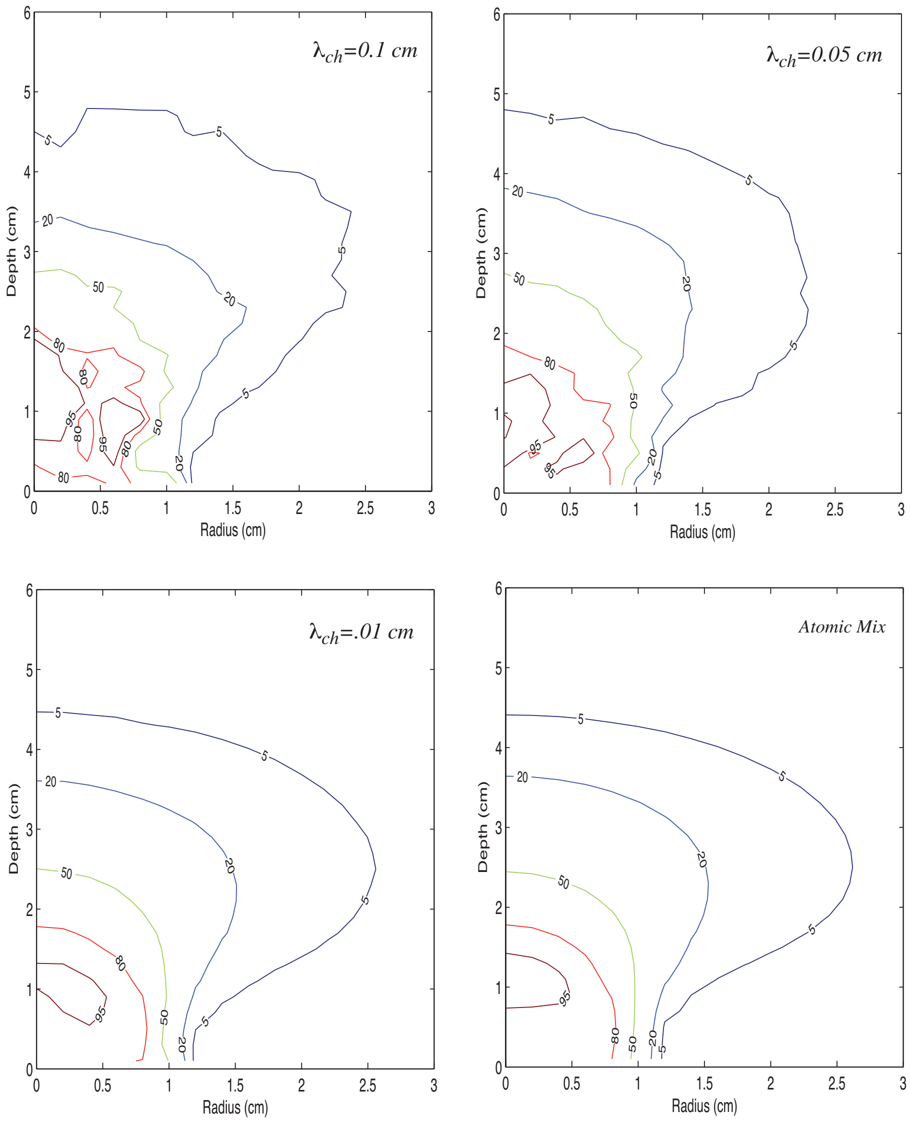


FIG. 2. 2.0 *MeV electron beam contour plots.*

simulated the homogenized (atomic mix) target. The results of these experiments are shown in Figure 2, which depicts isodose contours, normalized so that the maximum dose is unity.

This figure shows that as the chunk sizes decrease from $\lambda_{ch} = 0.1$ cm to 0.01 cm, the dose contours limit very well to the atomic mix result. The only significant difference between the $\lambda_{ch} = 0.01$ cm and atomic mix plots occurs on the 95% isodose contour, within 1.0 cm of the boundary. All of our electron beam simulations indicate such a phenomenon. The flux gradients are steep in these boundary locations, and $\lambda_{ch} = 0.01$ cm is not quite small enough for the atomic mix approximation to be valid there.

In the second set of simulations, a circular (radius = 1.0 cm) 3.4 MeV photon beam is normally incident on the same targets as above. (The energy of the photon beam was selected so that the electrons produced by Compton scattering would have roughly the same energy range as in the electron beam experiments.) These simulations have two Boltzmann equations: one for the photons and one for the electrons. The source term for the photon Boltzmann equation is the prescribed incident photon boundary flux. The source term for the electron Boltzmann equation is a volumetric term, proportional to the rate at which photons Compton scatter. The photon mean free path is on the order of cm, so the atomic mix approximation of these problems easily applies to the photon Boltzmann equation. The relevant electron data is shown in Figure 1, so as before, we predict that the atomic mix approximation for the electron Boltzmann equation should be accurate when the chunk sizes are about 0.01 cm or less.

The results (contour plots) of the photon beam experiments are depicted in Figure 3. The four plots in this figure exhibit the same trends as in Figure 2; as the chunk sizes decrease to about 0.01 cm, the contour plots converge to the contour plot for the atomic mix approximation. The only noteworthy difference is that the boundary effects seen in Figure 2 are not significant in Figure 3. This is likely because, in the photon beam experiments, electrons are produced by the Compton scattering of photons, and hence the electron "source" for the Figure 3 problems is much more spatially distributed than in Figure 2.

Figures 2 and 3 are typical. (We ran multiple realizations for each plot shown in the figures.) They indicate that when the chunk sizes become sufficiently small, the atomic mix limit is attained, and that the asymptotic theory well predicts the chunk sizes ($\lambda_{ch} \approx 0.01$ cm) for which the atomic mix approximation becomes accurate. Figure 1 shows that this size is about two orders of magnitude greater than a typical electron mean free path.

4. Discussion. We have presented a formal asymptotic theory and accompanying numerical results showing that the atomic mix approximation for charged particle transport is valid for physical systems in which (i) soft collisions dominate, (ii) a typical chunk size is small compared to a transport mean free path, and (iii) the average of any cross section over a line equals its volume average. The asymptotic theory uses earlier work by Pomraning [1] and Dumas and Golse [5]. The Monte Carlo results, generated by PENELOPE [13], are consistent with the asymptotic theory.

This work was motivated by the problem of theoretically assessing certain treatment planning procedures used in radiation oncology (radiation cancer therapy). In this field of medicine, carefully sculpted beams of high-energy photons and electrons coverage inside a patient, with the intent of sterilizing a malignant tumor [14, 15, 16]. (Photon beams also produce electrons, through Compton scattering, and these electrons deposit all the dose.) To model radiation beams penetrating the lung, standard computer codes model the lung as a union of homogenized subvolumes of about 75% air and 25% tissue (water), each subvolume having its own density, which is obtained from computerized tomography (CT) scans [17, 18, 19]. (Thus, the atomic mix approximation is used in each subvolume.) The proper resolution (size) of the subvolumes is a matter of debate.

However, the lung is an extraordinarily complex organ, with a complicated hierarchy of structures ranging from the principal bronchi (about 2 cm in diameter) to the alveoli (about 10^{-4} cm in diameter) [20, 21]. The mean free path of photons is on the order of several cm, so the atomic mix model for the entire lung is acceptable for

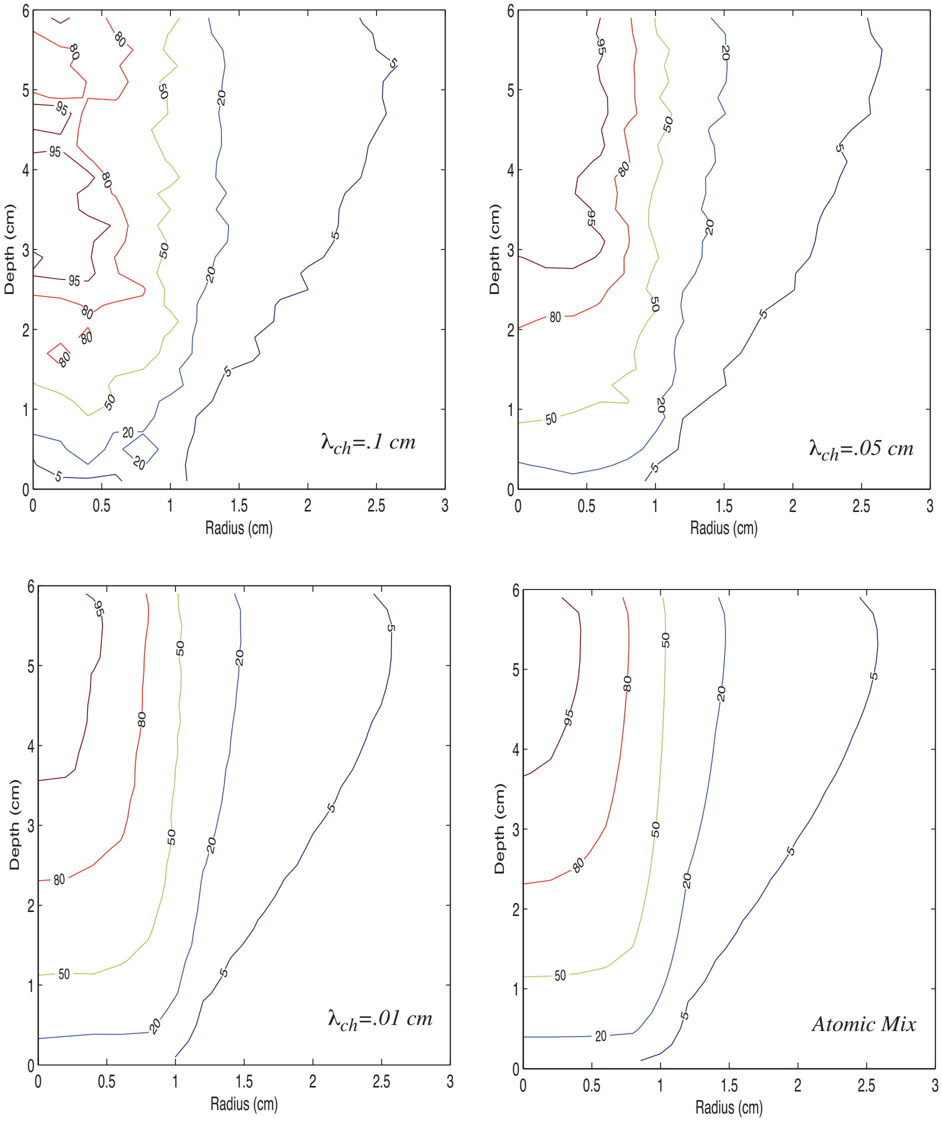


FIG. 3. 3.4 *MeV photon beam contour plots.*

photons. The results of this paper show that for random 75% air–25% water systems with uniformly sized "chunks" of water, the atomic mix approximation (i) is accurate with chunk sizes less than about 0.01 cm in diameter, but (ii) may be inaccurate when the chunk sizes are larger. Since the human lung contains structures with a large hierarchy of sizes, this result cannot be directly used to predict an optimal CT resolution for treatment planning.

However, the work in this paper suggests a way to examine this problem: first, construct a lung model in which all structures larger than a specified critical size are explicitly included and all structures less than the critical size are approximated by atomic mix [12]. (The critical size is chosen as large as possible so that it does not significantly affect Monte Carlo calculations of dose.) Then, compare Monte Carlo simulations of this first model to Monte Carlo simulations of a second CT model, obtained by homogenizing the first model over the user-prescribed CT subvolumes. The differences in dose for the two models are due to the partitioning of the lung into subvolumes. These differences will diminish as the subvolumes become smaller.

We have used PENELOPE to run realistic Monte Carlo simulations on our lung models and have found that $\lambda_{cr} = 0.05$ cm is an acceptable critical value. A detailed description of our lung model and preliminary results of comparisons with CT resolutions are given in [12]. We have found that for large CT resolutions, significant errors in predicted dose can occur, especially for narrow beams that pass through one or more "large" structures.

To conclude, we note from Figure 1 that a typical electron mean free path in tissue is about 10^{-4} cm. If it were necessary to explicitly model all lung structures comparable to or greater than this size, then all structures within the lung would have to be treated explicitly; this would be nearly an impossible computational task. Therefore, the atomic mix result developed in this paper is a crucial theoretical element in the strategy of using Monte Carlo techniques to assess the accuracy of existing treatment planning methods for the lung.

We plan to continue this work—in particular, to more fully assess the accuracy of different CT resolutions—and to report our results in future publications.

Acknowledgments. The authors gratefully thank Indrin Chetty for his encouragement and support.

REFERENCES

[1] G. C. Pomraning, *Linear Kinetic Theory and Particle Transport in Stochastic Mixtures*, World Scientific Press, Singapore, 1991.

[2] S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1960.

[3] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, Van Nostrand Reinhold, New York, 1970.

[4] T. M. Jenkins, W. R. Nelson, and A. Rindi, eds., *Monte Carlo Transport of Electrons and Photons*, Plenum Press, New York, 1988.

[5] L. Dumas and F. Golse, *Homogenization of transport equations*, SIAM J. Appl. Math., 60 (2000), p. 1447–1470.

[6] E. W. Larsen, *Asymptotic derivation of the atomic-mix diffusion model for* 1*-D random media*, Trans. Amer. Nucl. Soc., 89 (2003), pp. 296–297.

[7] E. W. Larsen, R. Vasques, and M. T. Vilhena, *Particle transport in the* 1*-D diffusive atomic mix limit*, in Proceedings of the Conference on Mathematics and Computation, Supercomputing, Reactor Physics and Nuclear and Biological Applications, Avignon, France, 2005. American Nuclear Society, LaGrange Park, IL, 2005 (CD-ROM).

[8] G. C. Pomraning, *The Fokker-Planck operator as an asymptotic limit*, Math. Models Methods Appl. Sci., 2 (1992), pp. 21–36.

[9] S. Chandrasekhar, *Stochastic problems in physics and astronomy*, Rev. Mod. Phys., 15 (1943), pp. 1–89.

[10] K. Przybylski and J. Ligou, *Numerical analysis of the Boltzmann equation including Fokker-Planck Terms*, Nuclear Sci. Engrg., 81 (1982), pp. 92–109.

[11] M. Caro and J. Ligou, *Treatment of scattering anisotropy of neutrons through the Boltzmann-Fokker-Planck equation*, Nuclear Sci. Engrg., 83 (1983), pp. 242–252.

[12] L. Liang, E. W. Larsen, and I. J. Chetty, *An anatomically-realistic lung model for Monte Carlo-based dosimetry*, Med. Phys., 34 (2007), pp. 1013–1025.

[13] F. Salvat, J. M. Fernandez-Varea, and J. Sempau, *PENELOPE*-2006: *A Code System for Monte Carlo Simulation of Electron and Photon Transport*, Nuclear Energy Agency Report 6222, European Organization for Economic Cooperation and Development (OECD), 2006; available online at htp://www.nea.fr/html/dbprog/peneloperef.html.

[14] A. S. Lichter and T. S. Lawrence, *Recent advances in radiation oncology*, New England J. Medicine, 332 (1995), pp. 371–379.

[15] D. W. O. Rogers and A. F. Bielajew, *Monte Carlo techniques of electron and photon transport for radiation dosimetry*, in The Dosimetry of Ionizing Radiation, K. R. Kase, B. E. Bjarngard, and F. H. Attix, eds., Academic Press, New York, 1990, vol. 3, Chapter 5, pp. 427–539.

[16] E. W. Larsen, *The nature of transport calculations used in radiation oncology*, Transport Theory Statist. Phys., 26 (1997), pp. 739–763.

[17] M. R. Arnfield, C. H. Siantar, J. Siebers, P. Garmon, L. Cox, and R. Mohan, *The impact of electron transport on the accuracy of computed dose*, Med. Phys., 27 (2000), pp. 1266–1274.

[18] S. J. Frank, K. M. Forster, C. W. Stevens, J. D. Cox, R. Komaki, A. Liao, S. Tucher, X. Wang, R. E. Steadham, C. Brooks, and G. Starkschall, *Treatment planning for lung cancer: Traditional homogeneous point-dose prescription compared with heterogeneity-corrected dose-volume prescription*, Int. J. Radiat. Oncol. Biol. Phys., 56 (2003), pp. 1308–1318.

[19] N. Papanikolaou, J. J. Battista, A. L. Boyer, C. Kappas, E. Klein, T. R. Mackie, M. Sharpe, and J. Van Dyk, *Tissue Inhomogeneity Corrections for Megavoltage Photon Beams*, AAPM Report 85, Medical Physics Publishing, Madison, WI, 2004.

[20] E. R. Weibel, *Morphometry of the Human Lung*, Springer, Berlin, 1963.

[21] K. Horsfield and G. Cumming, *Morphology of the bronchial tree in man*, J. Appl. Physiol., 24 (1968), pp. 373–383.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 59–79

STOCHASTIC DYNAMICS OF LONG SUPPLY CHAINS WITH RANDOM BREAKDOWNS*

P. DEGOND† AND C. RINGHOFER‡

Abstract. We analyze the stochastic large time behavior of long supply chains via a traffic flow random particle model. As items travel on a virtual road from one production stage to the next, random breakdowns of the processors at each stage are modeled via a Markov process. The result is a conservation law for the expectation of the part density which holds on time scales which are large compared to the mean up and down times of the processors.

Key words. supply chains, traffic flow models, mean field theories, Boltzmann equation, fluid limits

AMS subject classifications. 65N35, 65N05

DOI. 10.1137/060674302

1. Introduction.

Traffic flow models for supply chains model the flow of items through the chain as conservation laws for an item density ρ, depending on time and a stage variable x. Stage $x = 0$ denotes the raw material, stage $x = 1$ denotes the finished product, and the interval $[0, 1]$ models the intermediate stages of the production process and plays the role of the "road" in traffic flow theory. Traffic models have been used to model supply chains in [1, 2, 13, 5, 8] and, more recently, to optimize them in [6, 7, 9].

In previous work [3] we developed a traffic flow model for a chain of suppliers with a given capacity and throughput time. It is of the form

$$\partial_t \rho(x,t) + \partial_x F(x,t) = 0, \quad F(x,t) = \min\{\mu(x), V(x)\rho\}. \tag{1.1}$$

Here x denotes a continuous supplier index, i.e., the stage of the process. $\rho(x,t)$ denotes the density of parts in the supply chain. To compute the number of parts, i.e., the work in progress (WIP) $W_{ab}(t)$ in a certain subset of processors, corresponding to an interval (a,b) at a given time t, we have to compute $W_{ab}(t) = \int_a^b \rho(x,t)\ dx$. As long as the processors run below capacity, the movement of parts is given by the velocity V. So $\frac{dx}{V(x)}$ is proportional to the throughput time of the processor occupying the infinitesimal interval dx. The processors are assumed to have a finite capacity, meaning that they cannot process more that $\mu(x)dt$ parts in any infinitesimal time interval dt. So the variables in (1.1) have units of parts/stage for ρ, parts/time for μ, and stage/time for V. We prescribe a general, time-dependent influx of the form

$$F(0,t) = \lambda(t) \tag{1.2}$$

for the conservation law (1.1).

*Received by the editors November 7, 2006; accepted for publication (in revised form) May 30, 2007; published electronically October 10, 2007. This work was supported by the NSF under grant DMS-0604986.

http://www.siam.org/journals/siap/68-1/67430.html

†MIP, Laboratoire CNRS (UMR 5640), Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex 04, France (degond@mip.ups-tlse.fr).

‡Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (ringhofer@asu.edu).

Equation (1.1) is derived rigorously in [3] from a discrete recursion for the times each part arrives at each processor and from a limiting process for the number of parts and the number of processors $M \to \infty$. However, this recursion relation is completely deterministic, and the supply chain is therefore assumed to work like an automaton. The goal of this paper is to include random behavior of the processors, i.e., random breakdowns and random repair times, into the model. We model the breakdown of processors by setting the capacity $\mu(x)$ to zero. Thus, the model we consider consists of (1.1), where $\mu = \mu(x,t)$ is a time-dependent random variable. To be more precise, we assume $\mu(x,t)$ to be piecewise constant in space and of the form

$$\mu(x,t) = \sum_{m=0}^{M-1} \mu_m(t)\chi_{[\gamma_m,\gamma_{m+1})}(x), \tag{1.3}$$

where $0 = \gamma_0 < \cdots < \gamma_M = 1$ denotes a partition of the stage interval $[0,1]$, corresponding to M processors, and the functions $\mu_m(t),\ m = 0,\dots,M-1$, take on values of either $\mu_m(t) = 0$ or $\mu_m = c_m$, where c_m denotes the capacity of the processor in the case when it is running. We assume that the on/off switches are exponentially distributed in time; that is, we assume mean up and down times τ_m^{up} and τ_m^{down} and generate the random signal $\mu_m(t)$ by the following algorithm:

- Assuming that at time t processor m has just switched from the off state to the on state, choose Δt_{up}^m and Δt_{down}^m randomly from the distributions $d\mathcal{P}[\Delta t_{up}^m = s] = \frac{1}{\tau_m^{up}}\exp(-\frac{s}{\tau_m^{up}})ds$ and $d\mathcal{P}[\Delta t_{down}^m = s] = \frac{1}{\tau_m^{down}}\exp(-\frac{s}{\tau_m^{down}})ds$.
- Set $\mu_m(s) = c_m$ for $t < s < t + \Delta t_{up}^m$ and $\mu_m(s) = 0$ for $t + \Delta t_{up}^m < s < t + \Delta t_{up}^m + \Delta t_{down}^m$.
- At $t = t + \Delta t_{up}^m + \Delta t_{down}^m$ the processor is turned on again and we repeat the above process.

In this way we generate M random time-dependent signals which produce the random capacity $\mu(x,t)$ according to (1.3). For each realization of this process, we solve one realization of the conservation law (1.1), thereby modeling the random breakdown of elements in the chain. To illustrate this, Figure 1.1 shows one realization of one of the signals, namely $\mu_1(t)$, switching between $\mu_1 = c_1$ and $\mu_1 = 0$, and one realization of the solution of the corresponding conservation law. Note that the conservation law (1.1) exhibits, despite its simple form, a rather interesting feature. Since the flux function F is uniformly bounded from above by $\mu(x,t)$, it will necessarily become discontinuous if the flux coming from the left exceeds this value. This can be the case if $\mu(x,t)$ is discontinuous in the stage variable x, which will certainly occur if $\mu(x,t)$ is generated randomly by the algorithm above. Since mass has to be conserved, the discontinuity in the fluxes has to be compensated by δ-functions in the density ρ. The temporary buildup of these δ-functions is what is observed in Figure 1.1.

The goal of this paper is to derive an evolution equation for the expectation $\langle\rho(x,t)\rangle$ of the density ρ given by the stochastic process above. This provides us with a rather inexpensive way to estimate the behavior of long supply chains, with random breakdowns of individual processors, by solving directly one rather simple conservation law for the expectation. The main result of the present paper is that the expectation $\langle\rho(x,t)\rangle$ satisfies an initial boundary value problem for a conservation law of the form

(1.4)

$$\text{(a) } \partial_t\langle\rho(x,t)\rangle + \partial_x F_E(\bar\tau, C, V, \langle\rho\rangle) = 0, \quad F_E(\bar\tau, C, V, \langle\rho\rangle) = \bar\tau C\left[1 - \exp\left(-\frac{V\langle\rho}{C}\right)\right],$$

$$\text{(b) } F_E|_{x=0} = \lambda(t), \quad \langle\rho(x,0)\rangle = 0,$$

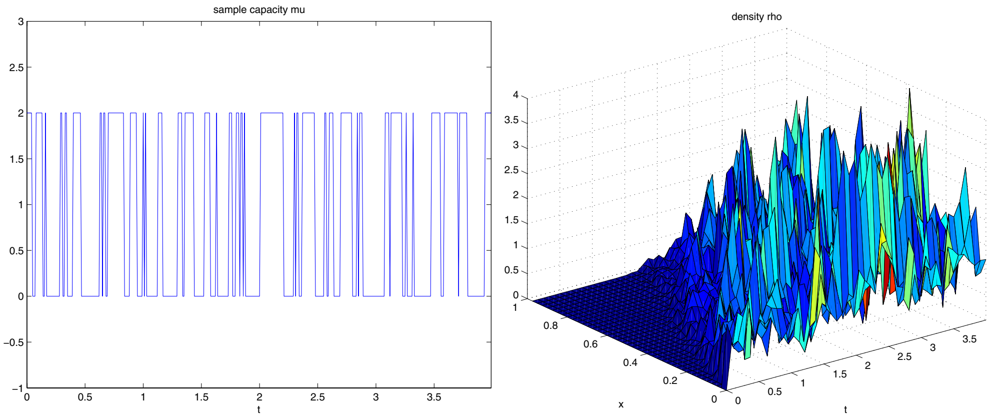


FIG. 1.1. *Left panel: One realization of the random capacity $\mu_1(t)$ of the first processor. Right panel: Density ρ from one realization of the conservation law* (1.1) *with random capacities.*

where the piecewise constant functions $\bar{\tau}$ and C are given by

$$\bar{\tau}(x) = \sum_{m=0}^{M-1} \chi_{[\gamma_m,\gamma_{m+1})}(x) \frac{\tau_m^{up}}{\tau_m^{up} + \tau_m^{down}}, \quad C(x) = \sum_{m=0}^{M-1} \chi_{[\gamma_m,\gamma_{m+1})}(x) c_m. \tag{1.5}$$

The result is derived in a limiting regime for large time scales and many parts and processors. So, it holds when the behavior of the chain, given by the stochastic version of (1.1), is considered on a time scale where a large number of parts arrive and the on/off switches of the processors occur very frequently. Similar models have been used on a heuristic basis, in the context of clearing functions, in [11, 14]. Our result basically states two facts as follows:

- For a large number of parts, the function $\min\{\mu, V\rho\}$ is, under the expectation, replaced by the function $\mu[1 - \exp(-\frac{V\rho}{\mu})]$, which has the same limiting behavior for large and small densities (the limits $\rho \to 0$ and $\rho \to \infty$).
- The effect of the random on/off switches can be incorporated into the model by replacing μ by the on-capacity c and multiplying the whole flux function by the average time $\frac{\tau^{up}}{\tau^{up}+\tau^{down}}$ the processor is on.

This paper is organized as follows. We prove the validity of the limiting equation (1.4) in a somewhat roundabout way. We first discretize one realization of (1.1) by a particle method, using the Lagrangian formulation of (1.1). Then we take the appropriate limits. So, section 2 is devoted to the formulation of the particle method. The limiting behavior is derived in section 3. In section 4 we verify our results numerically and demonstrate the basic premise of the method, namely, that we can accurately model the large time behavior of long chains with the mean field equation (1.4). The proofs of section 3 are given in the appendix.

2. Particle formulation. As mentioned in the introduction, we will derive the main result of this paper, the conservation law (1.4) for the expectation $\langle\rho\rangle$, from a particle discretization of (1.1) in Lagrangian coordinates. Since we are going to employ a mean field theory approach to the particle model in the next section, it is essential that the particle formulation of (1.1) is invariant under permutations of the particles. This will require some special considerations, and therefore we derive first a particle formulation of the deterministic problem, i.e., for one fixed realization of the random capacities $\mu(x, t)$. In section 2.2 we will then generalize this formulation to the random case.

2.1. The deterministic case. First, we reformulate problem (1.1) in Lagrangian coordinates. The transformation from Eulerian to Lagrangian coordinates is given in the usual manner by

$$\text{(2.1)} \qquad \text{(a) } \rho(x,t) = \int \delta(x - \xi(y,t))\ dy, \quad \text{(b) } \rho(\xi(y,t),t) = -\frac{1}{\partial_y \xi(y,t)},$$

where $\xi(y,t)$ denotes the position of a particle with continuous index y at time t. The derivative $\partial_y \xi = -\frac{1}{\rho}$ denotes the specific volume of the flow, i.e., the infinitesimal distance between two neighboring particles, and the minus sign indicates that we number the particles, at least initially, in order of their arrival, i.e., $y_1 < y_2 \Rightarrow \xi(y_1) > \xi(y_2) \Rightarrow \partial_y \xi < 0$ holds. Note, that (2.1)(b) holds only in the absence of caustics, that is, as long as the particles stay ordered and do not overtake one another, whereas (2.1)(a) also holds in the presence of caustics. Using the transformation (2.1), we see the conservation law (1.1) becomes

$$\text{(2.2)} \qquad \partial_t \xi(y,t) = v(\xi,t) = \min\left\{\frac{\mu(\xi,t)}{\rho(\xi,t)}, V(\xi)\right\}$$

and reduces to a parameterized ordinary differential equation (ODE) for the trajectories $\xi(y,t)$. We consider a particle discretization of one realization of the stochastic version of the conservation law (1.1) by replacing $\rho(x,t)$ by the measure corresponding to N particles

$$\text{(2.3)} \qquad \rho(x,t) \approx \Delta y \sum_{n=1}^{N} \delta(x - \xi_n(t))$$

(where we choose the symbol Δy for the particle weight so as to be notationally consistent with (2.1)) and solve the system of ODEs

$$\text{(2.4)} \qquad \partial_t \xi_n(t) = v_n, \quad v_n \approx \min\left\{\frac{\mu(\xi_n)}{\rho(\xi_n)}, V(\xi_n)\right\}.$$

The following three aspects are still missing in the consistent formulation of the particle method (2.4):

- We have to decide on an appropriate weight Δy of each particle.
- We have to define initial conditions for the trajectories ξ_n to reproduce the boundary condition (1.2) of the conservation law.
- We still have to define how to compute the density $\rho(x,t)$ at $x = \xi_n$ from the particle ensemble.

To address the first issue, we assume that we start from an empty system. In this case, the total mass over all time is given by the integral over the influx $\Lambda = \int_0^\infty \lambda(t)\ dt$, which we assume to be finite. To match this total mass Λ to the total mass in (2.3), we set $\Delta y = \frac{\Lambda}{N}$.

To address the second issue, we note that the flux F of the Lagrangian formulation (2.2) is given by

$$F(x,t) = \int \delta(x - \xi(y,t)) v(\xi)\ dy \ \Rightarrow\ \lambda(t) = F(0,t) = \int \delta(\xi(y,t)) v(\xi)\ dy;$$

defining the initial condition $\xi(y, a(y)) = 0$ for the particles in the Lagrangian formulation (2.2) implies

$$\text{(2.5)} \qquad \lambda(t) = \int \delta(t - a(y))\ dy \ \Rightarrow\ a^{-1}(t) = \int_0^t \lambda(s)\ ds.$$

Thus, the arrival times $a(y)$ have to be chosen as the functional inverse of the monotone antiderivative of the influx function $\lambda(t)$, i.e., $a(y) = t \iff y = \int_0^t \lambda(s)\, ds$ has to hold. In the deterministic case, treated in [3], this implies that the arrivals $a(y)$ satisfy the ODE

$$\frac{da(y)}{dy} = \frac{1}{\lambda(a(y))}, \quad a(0) = 0.$$

2.2. The random case. We now consider the stochastic process for the computation of the capacity variables $\mu_m(t)$ in (1.3). The assumption of an exponential distribution of the up and down times τ_m^{up} and τ_m^{down} implies a Markov process. This means that at each infinitesimal time, we can decide whether to switch the processor from on to off and back with a constant frequency $\omega_m(\mu) = \frac{1}{\tau_m^{up/down}}$. Thus, the evolution of $\mu_m(t)$ can be expressed by the process

$$\text{(2.6)} \quad \text{(a) } \mu_m(t + \Delta t) = (1 - r_m)\mu_m(t) + r_m(c_m - \mu_m(t)), \quad r_m = 0 \text{ or } r_m = 1,$$

$$\text{(b) } \mathcal{P}[r_m = 1] = \Delta t \omega_m(\mu_m(t)), \quad \mathcal{P}[r_m = 0] = 1 - \Delta t \omega_m(\mu_m(t)).$$

This means that at each infinitesimal time step Δt, we flip a coin and decide whether to switch, based on the probability $\Delta t \omega_m$, with the frequency ω_m given by

$$\omega_m(0) = \frac{1}{\tau_m^{down}}, \quad \omega_m(c_m) = \frac{1}{\tau_m^{up}}.$$

It is a standard exercise in the analysis of Monte Carlo methods (cf. [10]) that this algorithm results in exponentially distributed up and down times with means τ_m^{up} and τ_m^{down}.

Remark. Note that the assumption of exponentially distributed up and down times $\tau_m^{up}, \tau_m^{down}$ allows us to formulate the on/off switches as the Markov process (2.6). For a general probability distribution, the decision whether to turn the processor on or off at each time step depends on the time it has been in its present state (i.e., its history), leading to a model that is nonlocal in time. The case of a general distribution of up and down times will be the subject of a subsequent paper.

The motion of the particles ξ_n is now discretized in time, which leads to the following time-discrete version of (2.4):

$$\text{(2.7)} \qquad \xi_n(t + \Delta t) = \xi_n(t) + \Delta t v_n(\overrightarrow{\xi}(t), \overrightarrow{\mu}(t)),$$

where the velocities v_n depend on the whole particle ensemble $\overrightarrow{\xi} = (\xi_1, \dots, \xi_N)$ and, in addition, on the random capacity vector $\overrightarrow{\mu}(t) = (\mu_1, \dots, \mu_M)$. We still have to define a way to compute the density $\rho(\xi_n, t)$ from the particle ensemble $\overrightarrow{\xi}$. As stated before, the density ρ is given, in terms of the particle formulation, as the inverse of the specific volume, i.e., the distance of two neighboring particles. Formulating this in a way that is invariant under permutation of the particle index, we set

$$\text{(2.8)} \qquad \frac{1}{\rho(\xi_n, t)} = \min \left\{ \frac{\xi_k - \xi_n}{\Delta y} : \ \xi_k > \xi_n \right\}.$$

Note that if the particles stay in descending order, this reduces to $\frac{1}{\rho} = \frac{\xi_{n-1} - \xi_n}{\Delta y}$, which would be just the difference approximation to (2.1). The significance of the formula

(2.8) lies in the fact that it is also valid if particles pass each other and the descending order is destroyed. Therefore, we choose the velocities $v_n(\overrightarrow{\xi}, \overrightarrow{\mu})$ as

$$(2.9) \qquad v_n(\overrightarrow{\xi}(t), \overrightarrow{\mu}(t)) = \min\left\{V(\xi_n),\ \frac{\mu(\xi_n,t)}{\Delta y}(\xi_k - \xi_n) :\ \xi_k > \xi_n\right\},$$
$$\mu(\xi_n, t) = \sum_m \chi_m(\xi_n)\mu_m(t).$$

Turning to the boundary condition, we replace the influx density $\lambda(t)$ in (2.5) by a measure of the form

$$(2.10) \qquad \lambda(t) \approx \Delta y \sum_{n=1}^{N} \delta(t - a_n) = \frac{\Lambda}{N} \sum_{n=1}^{N} \delta(t - a_n).$$

The goal here is again to formulate (2.10) in such a way that the resulting particle method is invariant under permutations of the particle index n. We do so by randomizing (2.10), and we choose identically distributed random arrival times for each particle, according to the probability distribution $\frac{\lambda}{\Lambda}$. So, we have

$$(2.11) \qquad \xi_n(a_n) = 0, \quad d\mathcal{P}[a_n = t] = \frac{\lambda(t)}{\Lambda}dt, \quad \Lambda = \int_0^\infty \lambda(t)\ dt.$$

Equations (2.6)–(2.7), together with the definitions (2.9) and the initial condition (2.11), give a complete set of rules to advance the particle positions $\overrightarrow{\xi}$ and the capacities $\overrightarrow{\mu}$ from one time step to the next, and these rules are independent under permutations of the particle index. We will reformulate the system once more, to essentially replace the boundary condition (1.2) by an initial condition. This is really a technicality, and the reason for it is that, in the next section, we will derive equations for the probability density of the particle ensemble. To this end it is notationally more convenient to deal with a fixed number of particles in the system, instead of particles which enter at random times a_n. So, instead of imposing the condition $\xi_n(a_n) = 0$, we move the particles with an arbitrary, constant, and deterministic velocity—say $V(0)$—as long as $\xi_n(t) < 0$ holds, and start them out at $\xi_n(0) = -V(0)a_n$. Obviously $\xi_n(t) = V(0)(t - a_n)$ will hold for $\xi_n < 0$ and the particle will arrive at $\xi_n = 0$ at the correct time.

So, in summary, the stochastic particle system, which will be analyzed in the next section, is of the following form.

Start out at $t = 0$ with

$$(2.12) \qquad \text{(a) } \mu_m(0) = c_m,\ m = 0, \dots, M-1,$$
$$\text{(b) } \xi_n(0) = -V(0)a_n,\ n = 1, \dots, N, \quad d\mathcal{P}[a_n = t] = \frac{\lambda(t)}{\Lambda}dt.$$

To move particle positions $\overrightarrow{\xi}$ and capacities $\overrightarrow{\mu}$ for one time step Δt, compute

$$(2.13) \quad \text{(a) } \mu_m(t + \Delta t) = (1 - r_m)\mu_m(t) + r_m(c_m - \mu_m(t)), \quad r_m = 0 \textit{ or } r_m = 1,$$
$$\text{(b) } \mathcal{P}[r_m = 1] = \Delta t\omega_m(\mu_m(t)), \quad \mathcal{P}[r_m = 0] = 1 - \Delta t\omega_m(\mu_m(t)),$$

$$\text{(c) } \xi_{\mathrm{n}}(\mathrm{t}+\Delta\mathrm{t}) = \xi_n(t) + \Delta t v_n(\overrightarrow{\xi}(t), \overrightarrow{\mu}(t)),$$

$$\text{(d) } v_n(\overrightarrow{\xi}(t), \overrightarrow{\mu}(t)) = \begin{pmatrix} & V(0) & \xi_n < 0 \\ \min\{V(\xi_n), & \frac{\mu(\xi_n,t)}{\Delta y}(\xi_k - \xi_n): \quad \xi_k > \xi_n\} & \xi_n \geq 0 \end{pmatrix},$$

$$\text{(e) } \mu(\xi_n, t) = \sum_m \chi_m(\xi_n)\mu_m(t).$$

Remark. We assume in (2.12) for simplicity that all the processors in the beginning are on.

3. The evolution of the expectation.

3. The evolution of the expectation. This section is devoted to the derivation of main result (1.4) from the particle model (2.12)–(2.13). There are three steps involved. In section 3.1 we derive a high dimensional Boltzmann-type equation for the joint probability density of the particle positions $\overrightarrow{\xi}$ and the capacity variables $\overrightarrow{\mu}$ of the previous section. In section 3.2, we then reduce the dimensionality of the problem by employing a type of mean field theory for the conditional probability of the particle positions for a given realization of the capacities. In section 3.3 we compute averages over time scales which are much longer than the mean on/off switching times τ_m^{up} and τ_m^{down}. At leading order when the particle number tends to infinity, this procedure leads to an evolution equation for the probability $p(x,t)$ that an arbitrary particle in (2.13) is at position x at time t. Up to a multiplicative constant, p can be identified with the expectation $\langle\rho\rangle$ in (1.4).

3.1. The probability distribution. We now derive the evolution equation for the probability distribution

$$F(X, Z, t)dXZ = d\mathcal{P}[\overrightarrow{\xi}(t) = X, \overrightarrow{\mu}(t) = Z],$$

where the particle ensemble $\overrightarrow{\xi} = (\xi_1, \ldots, \xi_N)$ is at the N-dimensional position $X = (x_1, \ldots, x_N)$, while the processor capacities $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_m)$ are in the state $Z = (z_1, \ldots, z_m)$. We have the following theorem.

THEOREM 3.1. *Let the evolution of particles $\overrightarrow{\xi}$ and capacities $\overrightarrow{\mu}$ be given by* (2.12) *and* (2.13)*. Then, in the limit $\Delta t \to 0$ the joint probability distribution $F(X, Z, t)$ satisfies the initial value problem for the Boltzmann equation*

$$\text{(3.1)} \qquad \text{(a) } \partial_t F + \sum_n \partial_{x_n}[v_n(X, Z, t)F] = \int Q(Z, Z')F(X, Z', t)\; dZ',$$

$$\text{(b) } F(X, Z, 0) = (\Lambda V(0))^{-N}\left[\prod_{n=1}^{N} \lambda\left(-\frac{x_n}{V(0)}\right)\right]\left[\prod_{m=0}^{M-1} \delta(z_m - c_m)\right],$$

with the integral kernel Q given by

$$\text{(3.2)} \qquad \text{(a) } Q(Z, Z') = \sum_m q_m(z_m, z'_m)\prod_{k\neq m} \delta(z'_k - z_k),$$

$$\text{(b) } q_m(z_m, z'_m) = \omega_m(z'_m)[\delta(c_m - z'_m - z_m) - \delta(z'_m - z_m)].$$

Proof. The proof of Theorem 3.1 consists of summing up over all the possibilities of choosing the random variables r_m in (2.13), and it is an exercise in multidimensional Taylor expansion. It is deferred to the appendix.

In terms of the probability distribution $F(X, Z, t)$, the expectation $\langle \rho(x, t) \rangle$ in (1.4) is then given by

$$\langle \rho(x,t) \rangle = \Delta y \sum_{n=1}^{N} \int \delta(x - \xi_n(t)) F(\overrightarrow{\xi}, Z, t) \; d\overrightarrow{\xi} dZ = \Delta y \sum_{n=1}^{N} p_n(x,t), \tag{3.3}$$

with $p_n(x,t) = \int F(\xi_1, \ldots, \xi_{n-1}, x, \xi_{n+1}, \ldots, \xi_N, Z, t) \; d\xi_1, \ldots, \xi_{n-1}, \xi_{n+1}, \ldots, \xi_N \; dZ$ being the probability density that particle number n is at position x at time t. The density $F(X, Z, t)$ is of course of too high a dimension to be of practical use, and the goal of the next two sections is therefore to reduce the dimensionality of the problem.

3.2. Molecular chaos and mean field theory. Since the joint probability F in (3.1) depends on the capacity variables Z as well, the usual assumption of statistical independence (cf. [4]) has to be slightly modified. We first define the probability for the capacity variables Z as $G(Z,t) = \int F(X, Z, t) \; dX$. Integrating out X in (3.1) gives the initial value problem

$$\text{(a) } \partial_t G = \int Q(Z, Z') G(Z', t) \; dZ', \quad \text{(b) } G(Z, 0) = \prod_{m=0}^{M-1} \delta(z_m - c_m) \tag{3.4}$$

for G. Note that we obtain a closed equation for G, which is an expression of the fact that the capacities $\overrightarrow{\mu}$ evolve independently of the particles. Moreover, the individual capacities μ_m evolve independently of each other. This can be seen by the fact that (3.4) has a solution of the form $G(Z,t) = \prod_{m=0}^{M-1} g_m(z_m, t)$, where the individual probability densities $g_m(z_m, t)$, for the state z_m of the processor m at time t, satisfy the Boltzmann equation

$$\partial_t g_m(z,t) = \int q_m(z, z') g_m(z', t) \; dz', \tag{3.5}$$

with the kernels q_m given by (3.2)(b). We now define the conditional probability density $F^c(X, Z, t) dX = d\mathcal{P}[\overrightarrow{\xi} = X \mid \overrightarrow{\mu} = Z]$, which is the probability of the particle ensemble $\overrightarrow{\xi}$ for a given realization of the $\overrightarrow{\mu}$. The conditional probability density is defined by

$$F^c(X, Z, t) = \frac{F(X, Z, t)}{G(Z, t)}, \quad G(Z, t) = \int F(X, Z, t) \; dX. \tag{3.6}$$

Note that the definition (3.6) implies that $F^c(X, Z, t) dX$ is a probability measure for every fixed Z, i.e., $\int F^c(X, Z, t) \; dX = 1 \; \forall Z$ holds. Using the definition of F^c, (3.1) becomes

$$\partial_t [G F^c] + \sum_n \partial_{x_n} [v_n(X, Z, t) G F^c] = \int Q(Z, Z') F^c(X, Z', t) G(Z', t) \; dZ'. \tag{3.7}$$

The standard molecular chaos assumption employed in particle physics (cf. [4]) now takes the form that *for a given fixed realization of the* $\overrightarrow{\mu}$ the different ξ_n are independently and identically distributed, i.e., that

$$F^c(X, Z, t) = \prod_n f^c(x_n, Z, t), \qquad \int f^c(x, Z, t) \; dx = 1 \; \forall Z, t$$

holds. The molecular chaos assumption implies therefore the ansatz

$$F(X,Z,t) = G(Z,t)\prod_n f^c(x_n,Z,t)$$

for the joint probability F in (3.1). $f^c(x,Z,t)$ is the conditional probability density that any of the identical particles is at position x at time t for a given state Z of the processors. To obtain an evolution equation for $f^c(x,Z,t)$, we integrate (3.7) with respect to the variables $x_2 \dots x_N$ and obtain

$$(3.8)\partial_t[Gf^c(x_1,Z,t)] + \partial_{x_1}\left[Gf^c(x_1,Z,t)\int v_1(X,Z)\prod_{n=2}^{N} f^c(x_n,Z,t)dx_2\dots x_N\right]$$
$$= \int dZ'Q(Z,Z')f^c(x_1,Z')G(Z').$$

To close (3.8) we have to compute the average mean field velocity $u(x_1,Z,f^c)$, given by

$$u(x_1,Z,f^c) = \int v_1(X,Z)\prod_{n=2}^{N} f^c(x_n,Z,t)dx_2\dots x_N,$$

asymptotically for large N. To this end, we recall from (2.13)(d) that for $x_1 \in [\gamma_m,\gamma_{m+1})$, the interval corresponding to processor number m, the velocity $v_1(X,Z)$ is given by

$$v_1(X,Z) = \min\left\{V(x_1),\ \frac{z_m}{\Delta y}(x_k - x_1):\ x_k > x_1\right\}.$$

Theorem 3.2 gives the asymptotic form of the mean field velocity $u(x_1,Z,f^c)$ in the limit for a large number of independent particles ($N \to \infty$).

THEOREM 3.2. *For a given probability measure $f(x)$ and for given constants V and z,*

$$(3.9)\qquad \lim_{N\to\infty \Delta y\to 0}\int \min\left\{V, \frac{z}{\Delta y}(x_k - x_1):\ x_k > x_1\right\}\prod_{n=2}^{N} f(x_n)\ dx_2\dots x_N$$
$$= \frac{z}{\Lambda f(x_1)}\left[1 - \exp\left(-\frac{\Lambda V f(x_1)}{z}\right)\right]$$

holds where $\Lambda = N\Delta y$ is fixed.

Proof. The proof is deferred to the appendix.

Thus, in the mth cell $[\gamma_m,\gamma_{m+1})$, we have that the average mean field velocity is asymptotically given by (3.9) and can therefore be expressed by the piecewise constant function

$$(3.10)\qquad \text{(a) } u(x_1,Z,f^c) = \sum_{m=0}^{M-1} u_m(x_1,z_m,f^c)\chi_{[\gamma_m,\gamma_{m+1})}(x_1),$$
$$\text{(b) } u_m(x_1,z_m,f^c) = \frac{z_m}{\Lambda f^c}\left[1 - \exp\left(-\frac{\Lambda V(x_1)f^c}{z_m}\right)\right].$$

Therefore (3.8) reduces, under the molecular chaos assumption of many independently distributed particles, to the mean field Boltzmann equation

$$(3.11)\qquad \partial_t[Gf^c] + \partial_{x_1}[u(x_1,Z,f^c)Gf^c] = \int dZ'Q(Z,Z')f^c(x_1,Z')G(Z'),$$

with the mean field velocity u given by (3.10). The molecular chaos ansatz is compatible with the initial condition (3.1)(b) for the probability density $F(X,Z,t)$. Using (3.1)(b), we obtain the initial conditions

$$f^c(x_1,Z,0)=\frac{1}{\Lambda V(0)}\lambda\left(-\frac{x_1}{V(0)}\right),\quad G(Z,0)=\prod_{m=0}^{M-1}\delta(z_m-c_m) \tag{3.12}$$

for the evolution equations (3.11) and (3.4). Note that G still independently satisfies (3.4). This is essential, since it guarantees that $f^c(x_1,Z,t)dx_1$ is a probability measure, i.e., $\int f^c(x_1,Z,t)\ dx_1=1\ \forall Z,t$ holds.

The probability density $p_n(x,t)$ of particle number n being at position x at time t in (3.3) is, under the molecular chaos assumption, of the form

$$p_n(x,t)=p(x,t)=\int f^c(x,Z,t)G(Z,t)\ dZ\ \forall n.$$

Since all the p_n's are now identical, the expectation $\langle\rho\rangle$ is, according to (3.3), given by

$$\langle\rho(x,t)\rangle=\Lambda p(x,t).$$

So, the expectation $\langle\rho\rangle$ can be identified with the probability p up to the multiplicative constant Λ, giving the total mass in the system. To obtain the evolution equation (1.4) for the expectation $\langle\rho\rangle$ and the probability density p, we still have to average out somehow the dependence of the conditional probability density f^c on the processor-state variables $Z=(z_0,\dots,z_{M-1})$. The evolution equation for p—and also for $\langle\rho\rangle$—is obtained by integrating out the Z-variable in (3.11). This gives

$$\partial_t p(x,t)+\partial_x[pU(x,t,p)]=0,\quad pU(x,t,p)=\int u(x,Z,f^c)G(Z,t)f^c(x,Z,t)\ dZ. \tag{3.13}$$

Because of the initial condition (3.12) for the conditional probability density f^c, the conservation law (3.13) is subject to the initial condition

$$p(x,0)=\frac{1}{\Lambda V(0)}\lambda\left(-\frac{x}{V(0)}\right). \tag{3.14}$$

Note that f^c still depends on all the capacity variables Z. Therefore (3.13) has to be closed somehow by expressing f^c in terms of p. In section 3.3 this closure is achieved by considering a large time regime.

3.3. The large time regime. First, we note that in the setting of section 2.2 the capacities μ_m can assume only two discrete values, namely $\mu_m=0$ and $\mu_m=c_m$. Therefore the probabilities $g_m(z_m,t)$ in (3.5) are concentrated on these values, and we have an exact solution of (3.5) given by

$$g_m(z,t)=g_m^0(t)\delta(z)+g_m^1(t)\delta(z-c_m).$$

Inserting this into (3.5) and using the form of the integral kernels q_m in (3.2)(b) gives

$$\begin{aligned}&\delta(z)\partial_t g_m^0(t)+\delta(z-c_m)\partial_t g_m^1(t)=q_m(z,0)g_m^0(t)+q_m(z,c_m)g_m^1(t)\\&\quad=[-\omega_m(0)\delta(z)+\omega_m(0)\delta(c_m-z)]g_m^0(t)+[-\omega_m(c_m)\delta(c_m-z)+\omega_m(c_m)\delta(z)]g_m^1(t).\end{aligned}$$

Comparing the coefficients of $\delta(z)$ and $\delta(z-c_m)$, we obtain that $g_m^{0,1}(t)$ are given as solutions of the ODE system

(3.15)
$$\partial_t g_m^0 = -\omega_m(0)g_m^0(t) + \omega_m(c_m)g_m^1(t), \quad \partial_t g_m^1 = \omega_m(0)g_m^0(t) - \omega_m(c_m)g_m^1(t),$$

which preserves the property of $g_m(z,t)dz$ being a probability measure, i.e., $g_m^0 + g_m^1 = 1 \ \forall t$ holds. We now consider a regime where the on/off switches of the processors occur very frequently compared to the overall time scale, i.e., $\tau_m^{up/down} \ll 1$, $\omega_m = \frac{1}{\tau_m} \gg 1$. Thus we rescale the mean up and down times $\tau_m^{up/down}$ as well as the frequencies ω_m in (3.2) by $\tau_m^{up/down} \to \varepsilon\tau_m^{up/down}$ and $\omega_m \to \frac{1}{\varepsilon}\omega_m$. Rescaling the collision kernel Q in (3.11) correspondingly gives the system

$$\partial_t[Gf^c] + \partial_x[u(x,Z,f^c)Gf^c] = \frac{1}{\varepsilon}\int Q(Z,Z')f^c(x,Z',t)G(Z',t)\,dZ', \tag{3.16}$$

with the rescaled collision kernel Q given, according to (3.2), by

$$\text{(a) } Q(Z,Z') = \sum_m q_m(z_m,z_m')\prod_{k\neq m}\delta(z_k'-z_k), \tag{3.17}$$
$$\text{(b) } q_m(z_m,z_m') = \omega_m(z_m')[\delta(c_m - z_m' - z_m) - \delta(z_m' - z_m)].$$

From the above derivation, we have that the probability density $G(Z,t)$ of the processor status factors into M independent densities, supported on $z_m = 0$ and $z_m = c_m$, satisfying the rescaled version of (3.15). Thus, we have

$$\text{(a) } G(Z,t) = \prod_{m=1}^M g_m(z_m,t), \quad g_m(z,t) = g_m^0(t)\delta(z) + g_m^1(t)\delta(z-c_m), \tag{3.18}$$
$$\text{(b) } \varepsilon\partial_t g_m^0 = -\omega_m(0)g_m^0(t) + \omega_m(c_m)g_m^1(t), \quad \varepsilon\partial_t g_m^1 = \omega_m(0)g_m^0(t) - \omega_m(c_m)g_m^1(t),$$

where the small parameter ε denotes the ratio of $\tau^{up/down}$ to the overall time scale. The ODE system (3.18)(b) has two distinct eigenvalues, namely zero and $-\frac{\omega_m(0)+\omega_m(c_m)}{2\varepsilon}$. This, together with the condition that $g_m^0 + g_m^1 = 1 \ \forall t$ holds, implies that the $g_m^{0,1}$ will converge exponentially on an $O(\frac{t}{\varepsilon})$ time scale towards their steady state

$$g_m^0(\infty) = \frac{\omega_m(c_m)}{\omega_m(0)+\omega_m(c_m)}, \quad g_m^1(\infty) = \frac{\omega_m(0)}{\omega_m(0)+\omega_m(c_m)}. \tag{3.19}$$

Therefore, we can, up to exponentially small terms, replace $G(Z,t)$ by $G(Z,\infty)$ in (3.16). Note that $G(Z,\infty)$ is a steady state of (3.4) and therefore satisfies

$$\int dZ'\ Q(Z,Z')G(Z',\infty) = 0 \ \forall Z. \tag{3.20}$$

Replacing $G(Z,t)$ by $G(Z,\infty)$ in (3.16), we obtain

$$\text{(a) } G(Z,\infty)\{\partial_t[f^c] + \partial_x[u(x,Z,f^c)f^c]\} = \frac{1}{\varepsilon}\mathbf{Q_G}[f^c], \tag{3.21}$$
$$\text{(b) } \mathbf{Q_G}[f^c] = \int dZ' Q(Z,Z')f^c(x,Z')G(Z',\infty).$$

Expanding the conditional probability density f^c formally in powers of ε gives that, in zeroth order, $\mathbf{Q_G}[f^c] = 0$ holds. Note that, because of (3.20), functions f^c which are independent of Z are automatically in the kernel of the collision operator $\mathbf{Q_G}$ in (3.21). Theorem 3.3 states that the kernel of the collision operator consists essentially of only such functions.

THEOREM 3.3. *Any element of the kernel of the collision operator* $\mathbf{Q_G}$ *defined in* (3.21)(b) *is constant on the vertices of the hypercube* $\prod_{m=0}^{M-1}[0, c_m]$. *So* $\mathbf{Q_G}[f] = 0$ *implies that*

$$f(z_1 \dots c_m \dots z_M) - f(z_1 \dots 0 \dots z_M) = 0 \ \forall m, \ \forall Z = (z_0, \dots, z_{M-1}) \in \prod_m \{0, c_m\} \tag{3.22}$$

holds.

Proof. The proof is deferred to the appendix.

Theorem 3.3 allows us to compute the macroscopic velocity $U(x, t, p)$ in the evolution equation for the probability p in (3.13) in terms of p itself. Since in zeroth order $\mathbf{Q_G}[f^c] = 0$ has to hold, $f^c(x, Z, t)$ has to be constant on the hypercube vertices $Z \in \prod_{m=0}^{M-1}\{0, c_m\}$. In (3.13) we have to compute the flux as

$$\begin{aligned} &\text{(a) } pU(x, t, p) = \int u(x, Z, f^c) G(Z, \infty) f^c(x, Z, t) \ dZ, \\ &\text{(b) } p(x, t) = \int f^c(x, Z, t) \ dZ. \end{aligned} \tag{3.23}$$

Because of (3.18), $G(Z, \infty)$ is concentrated on the vertices $\prod_{m=0}^{M-1}\{0, c_m\}$, where f^c is constant. Therefore the integral in (3.23)(a) factors, and we obtain

$$U(x, t, p) = U(x, p) = \int u(x, Z, p) G(Z, \infty) \ dZ. \tag{3.24}$$

The derivation above actually computes the zero order term in a Chapman–Enskog procedure for the Boltzmann equation (3.21). The next term would produce a diffusive $O(\varepsilon)$ correction. However, this diffusive correction represents really only a small correction since the mean velocity U of the zero order term is nonzero; i.e., we are still in a primarily hyperbolic instead of a diffusive regime. We now have, in the large time limit, closed (3.13) for the probability density $p(x, t)$ that any of the identical particles is at position x at time t. Since this density is up to the multiplicative constant Λ identical to the expectation $\langle \rho \rangle$, i.e., $\langle \rho \rangle = \Lambda p$ holds, we also obtain a closed form equation for the expectation. This equation is of the form

$$\partial_t \langle \rho(x, t) \rangle + \partial_x \left[\langle \rho \rangle U \left(x, \frac{\langle \rho \rangle}{\Lambda} \right) \right] = 0. \tag{3.25}$$

Using the initial condition (3.14) for the probability density p, we obtain that the conservation law (3.25) is subject to the initial condition

$$\langle \rho(x, 0) \rangle = \frac{1}{V(0)} \lambda \left(-\frac{x}{V(0)} \right). \tag{3.26}$$

We have now assembled all the ingredients for the main result announced in (1.4). Computing $U(x, p)$ in (3.24), using the form (3.10) of the mean field velocity $u(x, Z, f^c)$, we have that in the interval $[\gamma_m, \gamma_{m+1})$, corresponding to the mth processor,

$$U(x, p) = \int \frac{z_m}{\Lambda p} \left[1 - \exp\left(-\frac{\Lambda p V(x)}{z_m} \right) \right] G(Z, \infty) \ dZ, \quad x \in [\gamma_m, \gamma_{m+1}),$$

holds. Using the fact that $G(Z,\infty)$ factors into a product of the g_m, and integrating out all variables, except z_m, we obtain

$$U(x,p) = \int \frac{z_m}{\Lambda p}\left[1 - \exp\left(-\frac{\Lambda p V(x)}{z_m}\right)\right] g_m(z_m,\infty)\, dz_m, \quad x \in [\gamma_m, \gamma_{m+1}).$$

Using the formulas (3.18) and (3.19) for $g_m(z,\infty)$, we integrate with respect to z_m and replace p by $\frac{\langle\rho\rangle}{\Lambda}$, obtaining for the velocity $U(x,\frac{\langle\rho\rangle}{\Lambda})$ in (3.25)

(3.27)
$$U\left(x,\frac{\langle\rho\rangle}{\Lambda}\right) = \frac{\omega_m(0)}{\omega_m(0)+\omega_m(1)}\frac{c_m}{\langle\rho\rangle}\left[1-\exp\left(-\frac{\langle\rho\rangle V(x)}{c_m}\right)\right] \text{ for } x \in [\gamma_m,\gamma_{m+1}),$$

which yields the flux function $F_E = \langle\rho\rangle U(x,\frac{\langle\rho\rangle}{\Lambda})$ in (1.4)(a), since $\frac{\omega_m(0)}{\omega_m(0)+\omega_m(1)} = \frac{\tau_m^{up}}{\tau_m^{up}+\tau_m^{down}}$ holds. Note that the ratio $\frac{\tau_m^{up}}{\tau_m^{up}+\tau_m^{down}}$ is not affected by the rescaling of the mean up and down times $\tau_m^{up/down}$ used in this section. Finally, we remove the technicality of formulating the conservation law as a pure initial value problem, which was used solely to keep the total mass constant in time and to define probability densities. For $x<0$ the velocities of the particles defined in section 2.2 are constantly equal to $V(0)$, and therefore also $U(x,p)=V(0)$ for $x<0$ will hold. The resulting one way wave equation can be solved exactly and, using the initial condition (3.26), we have

$$\langle\rho(x,t)\rangle = \langle\rho(x-tV(0),0)\rangle = \frac{1}{V(0)}\lambda\left(t-\frac{x}{V(0)}\right) \text{ for } x<0.$$

Because of flux continuity,

$$F_E|_{x=0+} = \langle\rho\rangle U\left(x,\frac{\langle\rho\rangle}{\Lambda}\right)|_{x=0+} = V(0)\langle\rho(0-,t)\rangle = \lambda(t)$$

has to hold, which yields the boundary condition (1.4)(b). The boundary condition (1.4)(b) has to be interpreted in the following way. The flux function at influx $F_E|_{x=0+} = \langle\rho\rangle U|_{x=0+}$ is, because of (3.27), bounded from above by the quantity $\frac{\tau_0^{up}c_0}{\tau_0^{up}+\tau_0^{down}}$. If the influx $\lambda(t)$ exceeds this value, as is possible in the transient regime, this results in a flux discontinuity, and correspondingly in a δ-function concentration of the expected density $\langle\rho\rangle$ at the influx boundary at $x=0$.

Remarks.

- Equation (1.4)(a) says that the whole flux (and not just the capacity c_m) is multiplied with the effective up-time $\frac{\tau_m^{up}}{\tau_m^{up}+\tau_m^{down}}$ of processor m. This is reasonable, since even for an empty system $\langle\rho\rangle \ll 1$, the flow will be slowed by shutting down the processors.
- A somewhat puzzling fact is that (1.4) does not reduce to the deterministic conservation law (1.1) in the limit $\tau^{down} \to 0$, i.e., in the case when the processors are always on. The explanation is that the derivation of (1.4) is based on the assumption of molecular chaos for the individual particles, and this assumption is apparently not valid for the deterministic system.
- More precisely, it is the frequent on and off switches *of many or all of the processors* which create the conditions of molecular chaos and the resulting approximate statistical independence of the states. Numerical experiments (not included here) have shown that the presented theory is not applicable;

cf. the case when only one processor switches and the other processors are simply in the on state all the time. (So $\tau_m^{down} = 0$ holds for all but one of the processors.)

- Note that we have made the somewhat arbitrary choice of first applying the molecular chaos assumption in section 3.2 and then carrying out the asymptotics for large time scales; we could have reversed the order by carrying out the large time scale asymptotics on the full equation (3.1) and then making the assumption of molecular chaos on the manifold of the slow dynamics, i.e., on the kernel of the collision operator on the right-hand side of (3.1).

4. Numerical experiments. We now turn to numerically verifying the validity of the approximate conservation law (1.4) for the expectation $\langle\rho(x,t)\rangle$. We do so by comparing the average over realizations of the numerical solution of (1.1) to the solution of (1.4). So, we first generate M random signals $\mu_m(t), \mu = 0, \dots, M-1$, as depicted in Figure 1.1, and compute the corresponding time-dependent capacity function $\mu(x,t)$ according to (1.3). For a given realization of $\mu(x,t)$, the conservation law (1.1) is then solved by a standard Godunov method (see [3] for details and cf. [12] for details on the Godunov method). This process is repeated many times for different realizations, and one approximation to the expectation $\langle\rho(x,t)\rangle$ is obtained by calculating averages over different realizations. We compare this approximation to the direct solution of the conservation law (1.4), also obtained by a Godunov method. It should be pointed out that the Godunov method reduces to simple upwinding in all cases, since the velocities always stay nonnegative. We employ only a first order Godunov scheme since the individual realizations will develop δ-function concentrations as soon as the processors are turned off and μ becomes zero in certain intervals. The convergence of the first order Godunov method in this case is analyzed and documented in [3], whereas the convergence properties of higher order methods are not so obvious. In all the example below we consider a chain of 40 processors ($M = 40$) which are located in the stage interval $x \in [0,1]$. For simplicity, we assume that all processors have identical throughput times. This allows us to choose a constant velocity $V(x) = 1$ in (1.1) and (1.4) by choosing an appropriate time scale. Thus, $T = 1$ is the throughput time of a part through the whole chain if all processors run below capacity, and $T = \frac{1}{40}$ is the throughput time of an individual processor in this case. We set $\tau_m^{up} = \tau_m^{down} = \frac{1}{20}\ \forall m$. The processors run, on average, only half the time, and we are in the large time scale regime of section 3.3 since $\varepsilon = \frac{1}{20} \ll 1$ holds.

Experiment 1. In the first experiment we consider $M = 40$ identical processors with a peak capacity $c_m = 2,\ m = 0, \dots, 39$. Thus, the flux function F_E in (1.4) is bounded by the effective capacity $\frac{c\tau^{up}}{\tau^{up}+\tau^{down}} = 1$. We start with an empty system $\rho(x,0) = 0$ and use a constant influx $\lambda(t) = 0.5$, well below the effective capacity. Figure 4.1 shows the expectation $\langle\rho(x,t)\rangle$ computed by averaging 200 realizations of (1.1) and by solving (1.4). Note that we obtain a good *quantitative* agreement in the size of the steady state distribution as well as in the transient behavior, i.e., the velocity of the wave propagating from $x = 0$ to $x = 1$. From (1.4) we deduce that the steady state density $\langle\rho(x,\infty)\rangle$ is given by the equation $F_E(\frac{1}{2}, 2, 1, \langle\rho(x,\infty)\rangle) = \lambda$ or $1 - e^{-\langle\rho(x,\infty)\rangle/2} = 0.5$.

Experiment 2. In the second experiment we keep the setup of the first experiment but introduce a bottleneck in processors 11–20. The peak capacities c_m are shown in Figure 4.2. Note that, because of our choice of mean up and down times τ_m^{up} and τ_m^{down}, the effective capacities are half the peak capacities shown in Figure 4.2. We choose a constant influx $\lambda = 0.27$. Again, Figure 4.3 shows the expectation $\langle\rho(x,t)\rangle$

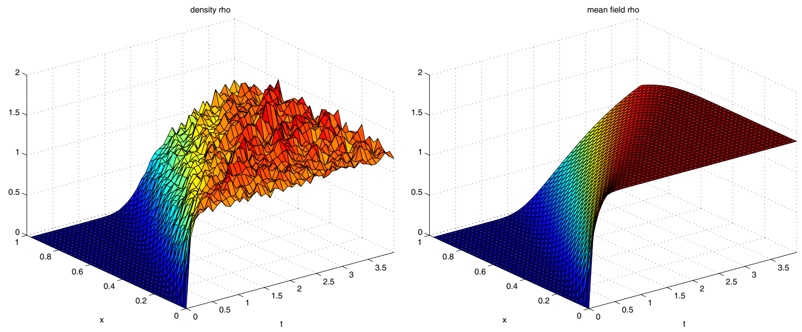


FIG. 4.1. *Experiment* 1. *Left panel: Density ρ from the deterministic conservation law* (1.1) *with random capacities and constant influx $\lambda = 0.5$. Averaged over* 200 *realizations. Right panel: Expectation $\langle \rho \rangle$ of the density ρ according to the mean field model* (1.4) *with constant influx $\lambda = 0.5$.*

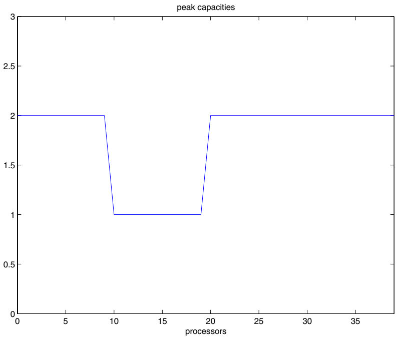


FIG. 4.2. *Experiment* 2. *Peak capacities c_m for $M = 40$ processors.*

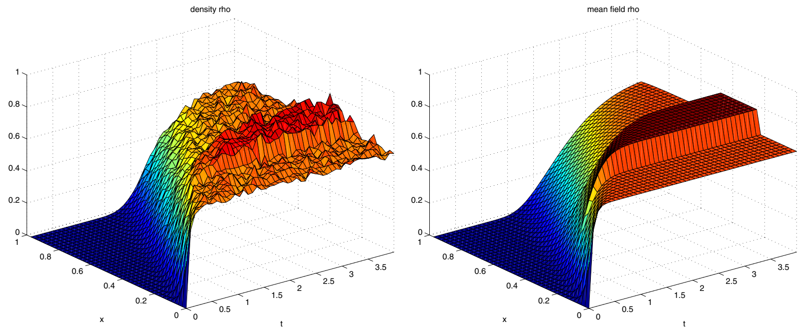


FIG. 4.3. *Experiment* 2. *Left panel: Density ρ from the deterministic conservation law* (1.1) *with random capacities and constant influx $\lambda = 0.27$,* 40 *cells, and bottleneck in cells* 11–20. *Averaged over* 200 *realizations. Right panel: Expectation $\langle \rho \rangle$ of the density ρ according to the mean field model* (1.4) *with constant influx $\lambda = 0.27$,* 40 *cells, and bottleneck in cells* 11–20.

computed by averaging 200 realizations of (1.1) and by solving (1.4).

Experiment 3. The promise of conservation law models for supply chains lies in their ability to provide a relatively inexpensive way to model the *transient* behavior of supply chains *far from steady state regimes*. Therefore, we perform the third experiment for a regime which is truly far from equilibrium. We keep the setup from

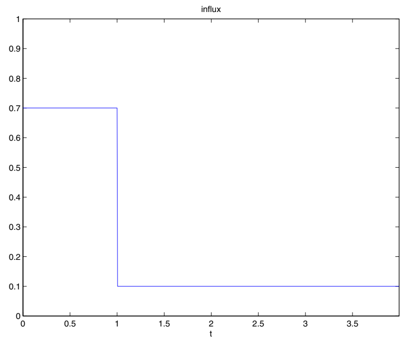


FIG. 4.4. *Experiment* 3. *Transient influx density.*

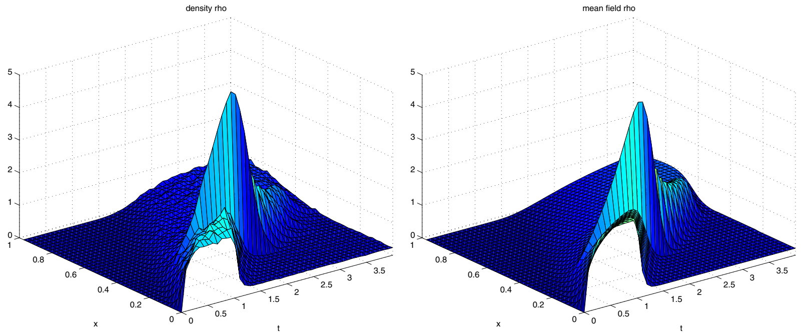


FIG. 4.5. *Experiment* 3. *Left panel: Density ρ from the deterministic conservation law* (1.1) *with random capacities and transient influx,* 40 *cells, and bottleneck in cells* 11–20. *Averaged over* 500 *realizations. Right panel: Expectation $\langle\rho\rangle$ of the density ρ according to the mean field model* (1.4) *with transient influx,* 40 *cells, and bottleneck in cells* 11–20.

the second experiment but use a transient influx density, shown in Figure 4.4. Note that the initial influx density $\lambda = 0.7$ is below the effective capacity $\frac{c_m \tau_m^{up}}{\tau_m^{up}+\tau_m^{down}}$ for most processors but exceeds the effective capacity for the bottleneck processors for $m = 11, \dots, 20$. Beyond $t = 1$ the transient influx $\lambda(t)$ is then well below the effective capacity for all processors. Thus, we will see a wave propagating through the first 10 processors $0 < x < 0.25$, the buildup of queues in the next 10 bottleneck processors $0.25 < x < 0.5$, and relaxation towards steady state after $t = 1$. Figure 4.5 shows the expectation $\langle\rho(x, t)\rangle$ computed by averaging 500 realizations of (1.1) and by solving (1.4). We observe again that the size of the peaks (the maximal queue length in front of the processors) as well as their location in the (x, t) plane (the transient response) are given accurately by the mean field model (1.4).

5. Appendix.

We start by proving the evolution equation (3.1) for the joint probability density $F(X, Z, t)$ of the particle positions and processor states.

Proof of Theorem 3.1. Once the random variables r_m in (2.6) are chosen, the rest of the evolution is completely deterministic. Summing up over all possible choices of the vector $R = (r_1, \dots, r_M)$ and weighting them with their probabilities gives

$$F(X, Z, t + \Delta t) = \int F(X', Z', t) \prod_n \delta(x'_n + \Delta t v_n(X', Z') - x_n)$$
$$\cdot \prod_m [\delta((1 - r_m)z'_m + r_m(c_m - z'_m) - z_m)]$$
$$\cdot \prod_m [\delta(r_m - 1)\Delta t \omega_m(z'_m) + \delta(r_m)(1 - \Delta t \omega_m(z'_m))] dX' dZ' dR.$$

We formulate the above relation weakly in X by integrating against a test function $\psi(X)$:

$$(5.1) \int \psi(X)F(X, Z, t + \Delta t)\ dX = \int dX'Z'R\ F(X', Z', t)\psi(X' + \Delta t V(X', Z'))$$
$$\cdot \prod_m [\delta((1 - r_m)z'_m + r_m(c_m - z'_m) - z_m)]$$
$$\cdot \prod_m [\delta(r_m - 1)\Delta t \omega_m(z'_m) + \delta(r_m)(1 - \Delta t \omega_m(z'_m))],$$

where the vector V denotes $(v_1, \dots, v_N)$. We Taylor-expand the terms on the right-hand side of (5.1) in Δt up to first order and obtain, after some calculus,

$$\int \psi(X)F(X, Z, t + \Delta t)\ dX$$
$$= \int dX\ F(X, Z, t)\psi(X) + \Delta t \int dX' F(X', Z, t)V(X', Z) \cdot \nabla_X \psi(X')$$
$$+ \Delta t \int dXZ'F(X, Z', t)\psi(X)Q(Z, Z'),$$

with the integral kernel Q given by (3.2). Letting $\Delta t \to 0$, we see this gives the weak form of

$$\partial_t F + \sum_n \partial_{x_n}[v_n(X, Z, t)F] = \int dZ' Q(Z, Z')F(X, Z'). \qquad \square$$

We now proceed to prove the form of the mean field velocity $u(x_1, Z, f^c)$ in (3.10); i.e., we prove Theorem 3.2. To prove Theorem 3.2 we will need the following auxiliary lemma, giving the expectation of the minimum of m independent random numbers, which are equidistributed in the interval $[0, 1]$.

LEMMA. *Let $\omega_1, \dots, \omega_m$ be m independent random numbers, uniformly distributed in the interval $[0, 1]$; then the expectation of the random function $\min\{\omega_1, \dots, \omega_m\}$ is given by*

$$(5.2)$$
$$E_m = \int \min\{\omega_k :\ k = 1, \dots, m\} \prod_{k=1}^m \chi_{(0,1)}(\omega_k)\ d\omega_1 \dots \omega_m = \frac{1}{m+1}, \quad m = 1, 2, \dots,$$

where $\chi_{(0,1)}$ denotes the usual indicator function on the interval $[0, 1]$.

Proof. The proof is based on induction in m. We denote by $R_m(s)$ the antiderivative of the probability density of the function $\min\{\omega_k :\ k = 1, \dots, m\}$, where $\omega_1, \dots, \omega_m$ are m random variables, uniformly distributed in $[0, 1]$. So we have

$$R'_m(s)ds = d\mathcal{P}[\min\{\omega_k :\ k = 1, \dots, m\} = s], \quad R_m(0) = 0,\ R_m(1) = 1,$$

and derive a formula for R_m recursively. The derivative $R'(s)$ is given by

$$\begin{aligned}
R'_m(s) &= \int_{[0,1]^m} \delta(s - \min\{\omega_1, \dots, \omega_m\}) d\omega_1 \dots \omega_m \\
&= \int_{[0,1]^m} \delta(s - \min\{\min\{\omega_1, \dots, \omega_{m-1}\}, \omega_m\}) d\omega_1 \dots \omega_m \\
&= \int_{[0,1]^2} \delta(s - \min\{r, \omega_m\}) R'_{m-1}(r)\ dr\omega_m \\
&= \int_{[0,1]^2} [H(r - \omega_m)\delta(s - \omega_m) + H(\omega_m - r)\delta(s - r)] R'_{m-1}(r)\ dr\omega_m \\
&= \int_0^1 H(r - s) R'_{m-1}(r)\ dr + R'_{m-1}(s) \int_0^1 H(\omega_m - s)\ d\omega_m.
\end{aligned}$$

Computing these integrals gives, because of $R_{m-1}(1) = 1$, the recursion

$$R'_m(s) = 1 - R_{m-1}(s) + (1 - s)R'_{m-1}(s) = \frac{d}{ds}[s + (1 - s)R_{m-1}(s)],$$

and because of $R_m(0) = 0,\ \forall m$ we obtain the recursive formula

$$R_m(s) = s + (1 - s)R_{m-1}(s), \quad R_1(s) = s. \tag{5.3}$$

Solving the recursion (5.3) via induction gives $R_m(s) = 1 - (1 - s)^m,\ m = 1, 2, \dots$. The expectation E_m is now given by

$$E_m = \int_0^1 sR'_m(s)\ ds = 1 - \int_0^1 R_m(s)\ ds = \int_0^1 (1 - s)^m\ ds = \frac{1}{m + 1}. \qquad \square$$

With the aid of the above lemma we are able to prove the mean field result of Theorem 3.2.

Proof of Theorem 3.2. In order to prove (3.9), we have to compute the limit of the quantity

$$u(x_1, z, V, f) = \int \min\left\{V, \frac{z}{\Delta y}(x_k - x_1),\ x_k > x_1,\ k = 2 \dots N\right\} \prod_{n=2}^{N} f(x_n)\ dx_2 \dots x_n$$

as $N \to \infty,\ \Delta y \to 0$ with $N\Delta y = \Lambda$ remaining constant, for a given probability measure f and constants z and V. u can be interpreted as the expectation of the quantity $\min\{V, \frac{z}{\Delta y}(x_k - x_1),\ x_k > x_1,\ k = 2 \dots N\}$, where $x_2, \dots, x_N$ are random variables independently and identically distributed according to the measure $f(x)$. We note that the variable x_k contributes only to the minimum if $x_1 < x_k < x_1 + \frac{\Delta y V}{z}$ holds. For any $k = 2, \dots, N$ let $\Delta y \bar{p}$ denote the probability that $x_k \in (x_1, x_1 + \frac{\Delta y V}{z})$ holds. Clearly, $\bar{p}$ is given by

$$\bar{p} = \frac{1}{\Delta y} \int_{x_1}^{x_1 + \frac{\Delta y V(x)}{z}} f(s)\ ds = \frac{V f(x_1)}{z} + O(\Delta y),$$

and the probability that none of the $x_k,\ k = 2, \dots, N$, is in the interval, i.e., the probability that $u = V$ holds, is given by $(1 - \Delta y \bar{p})^{N-1}$. We now compute the probability

p_m that of the $N-1$ variables $x_2, \ldots, x_N$ precisely $m \geq 1$ lie in the interval. p_m is given by

$$p_m = \binom{N-1}{m} (\Delta y \bar{p})^m [1 - \Delta y \bar{p}]^{N-1-m} \ , \tag{5.4}$$

where the binomial coefficient denotes the number of possible ways to choose m variables, and the other terms denote the probabilities that, for such a choice, the chosen m lie in the interval and the others do not. In the case that precisely m variables lie in the interval, their probability distribution can be replaced by the conditional probability, given that we already know that they are in the interval. This conditional probability is given by

$$q(s)ds = d\mathcal{P}\left[x_k = s \mid x_k \in \left(x_1, x_1 + \frac{\Delta y V}{z}\right)\right]$$

or

$$q(s) = \chi_{[x_1, x_1 + \frac{\Delta y V}{z}]}(s) \frac{f(s)}{\Delta y \bar{p}} \ . \tag{5.5}$$

Thus we obtain

$$u(x_1, z, V, f) = p_0 V + \sum_{m=1}^{N-1} p_m \frac{z}{\Delta y} \int \min\{s_k - x_1 : \ k = 1, \ldots, m\} \prod_{k=1}^{m} q(s_k) \ ds_1 \ldots s_m.$$

Substituting $s_k = x_1 + \frac{\Delta y V}{z} \omega_k$ in the integral gives

$$\begin{aligned} u(x_1, z, V, f) = p_0 V + \sum_{m=1}^{N-1} p_m V \int \min\{\omega_k : \ k = 1, \ldots, m\} \\ \cdot \prod_{k=1}^{m} \left[\frac{\Delta y V}{z} q\left(x_1 + \frac{\Delta y V}{z} \omega_k\right)\right] \ d\omega_1 \ldots \omega_m. \end{aligned}$$

Computing the probability density according to (5.5) gives

$$\frac{\Delta y V}{z} q\left(x_1 + \frac{\Delta y V}{z} \omega_k\right) = \frac{V}{z} \chi_{[0,1]}(\omega_k) \frac{f(x_1 + \frac{\Delta y V}{z} \omega_k)}{\bar{p}} = \chi_{[0,1]}(\omega_k) + O(\Delta y).$$

Thus, the ω_k are up to order $O(\Delta y)$ uniformly distributed in $[0, 1]$, and we have

$$u(x_1, z, V, f) = p_0 V + \sum_{m=1}^{N-1} p_m V [E_m + O(\Delta y)], \tag{5.6}$$

with the integral E_m given by

$$E_m = \int \min\{\omega_k : \ k = 1, \ldots, m\} \prod_{k=1}^{m} \chi_{(0,1)}(\omega_k) \ d\omega_1 \ldots \omega_m.$$

E_m is the expectation of the minimum of m uniformly distributed random variables and, according to the auxiliary lemma (5.2), $E_m = \frac{1}{m+1}$ holds.

Using this result in (5.6) gives

$$u(x_1, z, V, f) = V \sum_{m=0}^{N-1} p_m \left[\frac{1}{m+1} + O(\Delta y) \right].$$

Because of (5.4) we have that $\sum_{m=0}^{N-1} p_m = 1$ holds. Therefore, the $O(\Delta y)$ term can be neglected, although it appears inside the summation, and we have

$$(5.7) \qquad u(x_1, z, V, f) = V \sum_{m=0}^{N-1} \frac{1}{m+1} \binom{N-1}{m} (\Delta y \bar{p})^m (1 - \Delta y \bar{p})^{N-1-m} + O(\Delta y).$$

A simple application of the binomial theorem yields that

$$\sum_{m=0}^{N-1} \frac{1}{m+1} \binom{N-1}{m} a^m b^{N-1-m} = \frac{(a+b)^N - b^N}{Na} \quad \forall a, b$$

holds. With the obvious choice of a and b, we obtain from (5.7)

$$u(x_1, z, V, f) = V \frac{1 - (1 - \Delta y \bar{p})^N}{N \Delta y \bar{p}} + O(\Delta y) = V \frac{1 - e^{-\Lambda \bar{p}}}{\Lambda \bar{p}} + O(\Delta y).$$

(Remember $\Lambda = N\Delta y = \text{const}$ holds!) Together with $\bar{p} = \frac{Vf(x_1)}{z} + O(\Delta y)$, this gives (3.9). □

Finally, we prove the structure of the kernel of the collision operator $\mathbf{Q_G}$ in Theorem 3.3.

Proof of Theorem 3.3. From (3.17) we have that the collision kernel Q of the operator $\mathbf{Q_G}$ is of the form

$$Q(Z, Z') = \sum_m q_m(z_m, z'_m) \prod_{k \neq m} \delta(z_k - z'_k),$$
$$q_m(z_m, z'_m) = \omega_m(z'_m)[\delta(c_m - z_m - z'_m) - \delta(z'_m - z_m)].$$

At the same time, we have from (3.18) that the steady state $G(Z, \infty)$ of the processor state distribution is supported only on the hypercube $\prod_{m=0}^{M-1} \{0, c_m\}$. So $G(Z, \infty)$ is of the form

$$G(Z, \infty) = \prod_m g_m(z_m), \quad g_m(z_m) = g_m^0 \delta(z_m) + g_m^1 \delta(z_m - c_m).$$

Inserting this into the definition (3.21)(b) of the collision operator $\mathbf{Q_G}$ gives

$$\mathbf{Q_G}[f] = \sum_m \int q_m(z_m, z'_m)[g_m^0 \delta(z'_m) + g_m^1 \delta(z'_m - c_m)] f(Z') \prod_{k \neq m} \delta(z_k - z'_k) g_k(z'_k)\, dZ'.$$

Integrating out all variables except z'_m in each term of the sum above yields

$$\begin{aligned}\mathbf{Q_G}[f] &= \sum_m \int q_m(z_m, z'_m)[g_m^0 \delta(z'_m) + g_m^1 \delta(z'_m - c_m)] f(z_1 \dots z'_m \dots z_M)\, dz'_m \prod_{k \neq m} g_k(z_k) \\ &= \sum_m [q_m(z_m, 0) g_m^0 f(z_1 0 z_M) + q_m(z_m, c_m) g_m^1 f(z_1 \dots c_m \dots z_M)] \prod_{k \neq m} g_k(z_k).\end{aligned}$$

Using the form (3.17)(b) of the individual kernels q_m gives

$$\mathbf{Q_G}[f] = \sum_m \{\omega_m(0)[\delta(c_m - z_m) - \delta(z_m)]g_m^0 f(z_1 \dots 0 \dots z_M) \\ +\omega_m(c_m)[\delta(z_m) - \delta(c_m - z_m)]g_m^1 f(z_1 \dots c_m \dots z_M)\} \times \prod_{k \neq m} g_k(z_k) \ .$$

Using the form (3.19) of the coefficients $g_m^0(\infty)$ and $g_m^1(\infty)$ of the steady distribution, and collecting terms, gives

$$\mathbf{Q_G}[f] = \sum_m \frac{\omega_m(0)\omega_m(c_m)}{\omega_m(0) + \omega_m(c_m)}[\delta(c_m - z_m) - \delta(z_m)] \\ \cdot[f(z_1 \dots 0 \dots z_M) - f(z_1 \dots c_m \dots z_M)] \prod_{k \neq m} g_k(z_k) \ .$$

Therefore $\mathbf{Q_G}[f]$ can vanish identically $\forall Z$ only if

$$f(z_1 \dots c_m \dots z_M) - f(z_1 \dots 0 \dots z_M) = 0 \ \forall m, \ \forall Z \in \prod_m \{0, c_m\}$$

holds. $\square$

REFERENCES

[1] D. Armbruster, D. Marthaler, and C. Ringhofer, *Kinetic and fluid model hierarchies for supply chains*, Multiscale Model. Simul., 2 (2003), pp. 43–61.

[2] D. Armbruster and C. Ringhofer, *Thermalized kinetic and fluid models for reentrant supply chains*, Multiscale Model. Simul., 3 (2005), pp. 782–800.

[3] D. Armbruster, P. Degond, and C. Ringhofer, *A model for the dynamics of large queuing networks and supply chains*, SIAM J. Appl. Math., 66 (2006), pp. 896–920.

[4] C. Cercignani, R. Illner, and M. Pulvirenti, *The Mathematical Theory of Dilute Gases,* Springer-Verlag, New york, 1994.

[5] C. Daganzo, *A Theory of Supply Chains*, Lecture Notes in Econom. and Math. Systems 526, Springer-Verlag, Berlin, 2003.

[6] A. Fügenschuh, M. Herty, A. Klar, and A. Martin, *Combinatorial and continuous models for the optimization of traffic flows on networks*, SIAM J. Optim., 16 (2006), pp. 1155–1176.

[7] A. Fügenschuh, S. Göttlich, M. Herty, A. Klar, and A. Martin, *A Mixed-Integer Programming Approach for the Optimization of Continuous Models in Supply Chain Management*, preprint, Universität Kaiserslautern, Kaiserslautern, Germany, 2006.

[8] S. Göttlich, M. Herty, and A. Klar,, *Network models for supply chains*, Commun. Math. Sci., 3 (2005), pp. 545–559.

[9] S. Göttlich, M. Herty, C. Kirchner, and A. Klar, *Optimal control for continuous supply network models*, Netw. Heterog. Media, 1 (2006), pp. 675–688 (electronic).

[10] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles,* McGraw-Hill, Maidenhead, UK, 1981.

[11] U. Karmarkar, *Capacity loading and release planning with WIP and lead time,* J. Manuf. Oper. Management, 2 (1989), pp. 105–123.

[12] R. LeVeque, *Finite Volume Methods for Hyperbolic Problems,* Cambridge University Press, Cambridge, UK, 2002.

[13] D. Marthaler, D. Armbruster, C. Ringhofer, K. Kempf, and T. C. Jo, *A continuum model for a re-entrant factory,* Oper. Res., 54 (2006), pp. 933–951.

[14] H. Missbauer, *Aggregate order release planning for time varying demand,* Int. J. Production Res., 40 (2002), pp. 699–718.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 80–97

CREEP, RECOVERY, AND WAVES IN A NONLINEAR FIBER-REINFORCED VISCOELASTIC SOLID[*]

M. DESTRADE[†] AND G. SACCOMANDI[‡]

Abstract. We present a constitutive model capturing some of the experimentally observed features of soft biological tissues: nonlinear viscoelasticity, nonlinear elastic anisotropy, and nonlinear viscous anisotropy. For this model we derive the equation governing rectilinear shear motion in the plane of the fiber reinforcement; it is a nonlinear partial differential equation for the shear strain. Specializing the equation to the quasi-static processes of creep and recovery, we find that usual (exponential-like) time growth and decay exist in general, but that for certain ranges of values for the material parameters and for the angle between the shearing direction and the fiber direction, some anomalous behaviors emerge. These include persistence of a nonzero strain in the recovery experiment, strain growth in recovery, strain decay in creep, disappearance of the solution after a finite time, and similar odd comportments. For the full dynamical equation of motion, we find kink (traveling wave) solutions which cannot reach their assigned asymptotic limit.

Key words. fiber reinforcement, nonlinear creep and recovery, traveling waves

AMS subject classifications. 74D10, 74A10, 74H05, 74G30

DOI. 10.1137/060664483

1. Introduction.

Many biological, composite, and synthetic materials must be modeled as fiber-reinforced nonlinearly elastic solids. Hence, the anisotropy due to the presence of collagen fibers in many biological materials has been studied extensively within the constitutive context of fiber-reinforced materials by several authors (see, for example, Humphrey (2002) and the references therein.) In nonlinear elasticity, the macroscopic response of an anisotropic material is given in terms of a strain-energy function, which itself depends on a set of independent deformation invariants. This formulation captures a great variety of phenomena related to the behavior of fiber-reinforced materials, e.g., the examination of fiber instabilities, using loss of ellipticity (see Merodio and Ogden (2002), (2003), and the references therein).

Generally speaking, a reinforcement is added to a given material with the aim of avoiding a possible failure under operating conditions. Therefore it is important to develop a detailed study showing how to introduce reinforcements into a material in order to control the possible development of a boundary layer structure. Our goal here is to provide a first step in this direction. We make several simplifications and ad hoc assumptions. First, we limit ourselves to the consideration of *only one fiber direction* and second, we consider a *one-dimensional motion* in the bulk of an infinite body. Here the motion is linearly polarized in a direction normal to the plane containing the direction of propagation and the direction of the fibers. We acknowledge that more complex anisotropies, geometries, and couplings arise in biomechanical applications. For instance, the mechanics of the aorta involves two families of parallel fibers, triaxial motions, and blood flow/arterial wall coupling. However, we argue that some

[*] Received by the editors July 7, 2006; accepted for publication (in revised form) June 12, 2007; published electronically October 10, 2007. This work was supported by a Fellowship from the City of Paris.

http://www.siam.org/journals/siap/68-1/66448.html

[†] Institut Jean Le Rond d'Alembert (UMR 7190), CNRS/Université Pierre et Marie Curie, 4 place Jussieu, case 162, 75252 Paris Cedex 05, France (destrade@lmm.jussieu.fr).

[‡] Dipartimento di Ingegneria Industriale, Università degli Studi di Perugia, 06125 Perugia, Italy (giuseppe.saccomandi@unile.it).

major characteristics of biological soft tissues are encompassed in the choices of transverse isotropy, of infinite extent, and of a motion governed by an ordinary differential equation. Indeed the anisotropy due to the presence of one family of parallel fibers complicates the governing equations to an extent which is only marginally less than that due to the presence of two families of parallel fibers. Also, soft biological tissues are nearly incompressible, and a (compressive) longitudinal wave is difficult to observe; it thus make sense to focus on transverse shear motions, which are useful in imaging technologies. Our third assumption is that the elastic strain energy is the sum of an isotropic part and an anisotropic part (called a *reinforcing model*), in order to model an isotropic base material augmented by a uniaxial reinforcement in the *fiber direction.* Albeit strong, this constitutive assumption is now common and used by many authors (e.g., Triantafyllidis and Abeyaratne (1983), Qiu and Pence (1997), Merodio and Ogden (2002)). Finally, we assume that the solid is viscoelastic, and here we assume not only Newtonian viscosity (proportional to the stretching tensor) but also fiber-oriented (anisotropic) viscosity. That latter assumption is strong but can be removed from our calculations by taking a constant to be zero. We believe that it might be useful in modeling the well-documented physiological effect of stretching training in sport medicine, which is that it affects the viscosity of tendon structures but not their elasticity (Taylor et al. (1990); Kubo, Kanehisa, and Fukunaga (2002)).

We divide the article into the following sections. Section 2 presents the constitutive model and the derivation of the equation governing the rectilinear shear motion. As expected, this equation is nonlinear in the shear strain: it is a second-order partial differential equation, with cubic nonlinearity. To initiate its resolution, we first look at the quasi-static experiment of recovery in section 3. Then we have a first-order ordinary differential equation, and we find that it can lead to unusual behaviors when certain conditions (strong anisotropy, large angle between the shearing direction and the fibers) are met. The same is true of the case of creep, treated in section 4. Basically, it turns out that the nonlinearity introduces ranges of material parameters and angles for which an expected behavior—say, strain growth in creep—can be turned on its head, and lead to strain decay with time in creep, say. In the course of the investigation we develop synthetic tools of analysis which highlight the boundaries of these ranges. They also guide us for the resolution of the full dynamical equation of motion, which we tackle in section 5 for traveling wave solutions. Again the solution may behave in an unexpected way, provided that the anisotropy is strong enough and the fibers are in compression. Finally, section 6 recaps the results and puts them into a wider context.

2. Basic equations.

2.1. The viscoelastic anisotropic model.

We describe the motion of a body by a relation $\mathbf{x} = \mathbf{x}(\mathbf{X}, t)$, where $\mathbf{x}$ denotes the current coordinates of a point occupied by the particle of coordinates $\mathbf{X}$ in the reference configuration at the time t.

We introduce $\mathbf{F} = \partial\mathbf{x}/\partial\mathbf{X}$, the deformation gradient, and $\mathbf{C} = \mathbf{F}^{\mathrm{T}}\mathbf{F}$, the right Cauchy–Green strain tensor. We focus on *incompressible materials* for which all admissible deformations must be isochoric, or equivalently, for which the relation $\det \mathbf{F} = 1$ must hold at all times.

The body is reinforced with one family of parallel fibers. Our first assumption is that the unit vector $\mathbf{a}_0$, giving the fiber direction in the reference configuration, is independent of $\mathbf{X}$. The stretch along the fiber direction is $\sqrt{\mathbf{a}_0 \cdot \mathbf{C}\mathbf{a}_0} = \sqrt{\mathbf{a} \cdot \mathbf{a}}$, where $\mathbf{a} = \mathbf{F}\mathbf{a}_0$.

We may now introduce the elastic part of our constitutive model. We consider the so-called *standard reinforcing model*, which is a quite simple generalization to anisotropy of the neo-Hookean model (Triantafyllidis and Abeyaratne (1983); Qiu and Pence (1997)). For the standard reinforcing model, the strain-energy density is given by

$$W = \frac{\mu}{2}\left[(I_1-3)+\gamma_0(I_4-1)^2\right], \quad \text{where } I_1 = \text{tr } \mathbf{C},\ I_4 = \mathbf{a}_0\cdot\mathbf{C}\mathbf{a}_0 = \mathbf{a}\cdot\mathbf{a}. \tag{2.1}$$

Here $\mu > 0$ is the infinitesimal shear modulus of the isotropic neo-Hookean matrix, $\gamma_0 > 0$ is the *elastic anisotropy parameter*, and the invariant I_4 measures the squared stretch in the fiber direction. Mechanical tests show that the neo-Hookean strain energy function $\mu(I_1-3)/2$ fits uniaxial data rather well for arteries (Gundiah, Ratcliffe, and Pruitt (2007)), while the anisotropic term $\gamma_0(I_4-1)^2$ is adequate to describe a reinforced material which penalizes deformation in the fiber direction (Merodio and Ogden (2003)).

The spatial velocity gradient $\mathbf{L}(\mathbf{X},t)$ associated with a motion is defined as $\mathbf{L} =$ grad $\mathbf{v}$, where $\mathbf{v} = \partial\mathbf{x}/\partial t$ is the velocity, and the stretching tensor $\mathbf{D}$ is defined as $\mathbf{D} = \frac{1}{2}(\mathbf{L}+\mathbf{L}^{\mathrm{T}})$. For incompressible materials, tr $\mathbf{D} = 0$ at all times. Newtonian viscous fluids possess a constitutive term in the form $2\nu\mathbf{D}$, where ν is a constant. For our special solid, we modulate the Newtonian viscosity with an anisotropic term, by replacing ν with $\nu[1+\gamma_1(I_4-1)]$, where $\gamma_1 > 0$ is the *viscous anisotropy parameter.* We show in the course of the paper that this simple choice of anisotropic viscosity captures the essential characteristics of attenuation in soft biological fibrous tissues. According to Baldwin et al. (2006), ultrasonic measurements of freshly excised myocardium show that "the attenuation coefficient was found to increase as a function of frequency in an approximately linear manner and to increase monotonically as a function of angle of insonification from a minimum perpendicular to a maximum parallel relative to the direction of the myofibers."

We are now ready to give the complete Cauchy stress tensor of our viscoelastic, transversally isotropic material as

$$\mathbf{T} = -p\mathbf{I} + \mu[\mathbf{B} + \gamma_0(I_4-1)\mathbf{a}\otimes\mathbf{a}] + 2\nu[1+\gamma_1(I_4-1)]\mathbf{D}, \tag{2.2}$$

where the p is the yet indeterminate Lagrange multiplier introduced by the incompressibility constraint, and $\mathbf{B} = \mathbf{F}\mathbf{F}^{\mathrm{T}}$ is the left Cauchy–Green tensor.

2.2. Shear motion. We take a fixed orthonormal triad of vectors ($\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$), and call X, Y, Z the reference coordinates; hence $\mathbf{X} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}$. The triad is such that the unit vector in the fiber direction lies in the XY plane; hence,

$$\mathbf{a}_0 = \cos\theta\mathbf{i} + \sin\theta\mathbf{j} \tag{2.3}$$

(say), where $\theta \in [0,\pi]$ is the angle between the X-axis and the fibers.

We then consider the *rectilinear shearing motion*,

$$x = X + u(Y,t), \quad y = Y, \quad z = Z, \tag{2.4}$$

where the antiplane displacement u is real and finite. Then the components of the gradient of deformation $\mathbf{F}$ and of its inverse are given by

$$\mathbf{F} = \begin{bmatrix} 1 & U & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{F}^{-1} = \begin{bmatrix} 1 & -U & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.5}$$

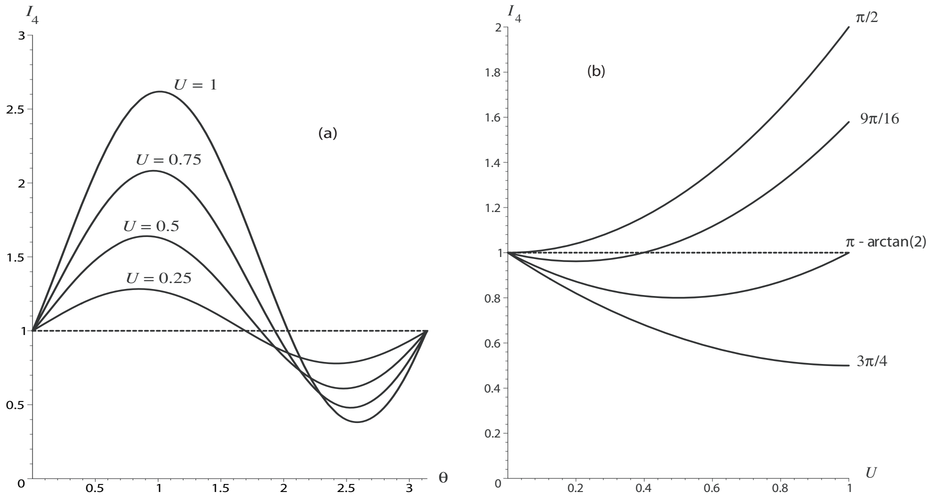


FIG. 2.1. *Variations of the squared stretch in the fiber direction*: (a) *with the angle and* (b) *with the shear. When $I_4 > 1$, the fibers are in extension; when $I_4 < 1$, they are in compression.*

where $U = \partial u/\partial Y$ is the *amount of shear*. The left and right Cauchy–Green tensors are thus

$$\mathbf{B} = \begin{bmatrix} U^2+1 & U & 0 \\ U & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 & U & 0 \\ U & U^2+1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.6}$$

respectively, from which the expressions of the invariants I_1 and I_4 follow,

$$I_1 = 3 + U^2, \quad I_4 = 1 + U \sin 2\theta + U^2 \sin^2\theta. \tag{2.7}$$

Figure 2.1(a) shows the variations of I_4 with θ for several values of U between 0 and 1. When $I_4 > 1$ the fibers are in extension, and when $I_4 < 1$ they are in compression; the figure shows that this latter behavior occurs in a smaller and smaller angular range, but is more and more pronounced, as the amount of shear is increased. Conversely, Figure 2.1(b) shows the variations of I_4 with U for several values of θ; when $0 < \theta < \pi/2$, the fibers are always in extension, and when $\pi - \tan^{-1}(2) = 2.034 < \theta < \pi$, they are always in compression for $0 \leq U \leq 1$. We refer to the paper by Qiu and Pence (1997) for similar figures and closely related discussions.

In the deformed configuration, we find that $\mathbf{a} = (\cos\theta + U\sin\theta)\mathbf{i} + \sin\theta\mathbf{j}$. The remaining tensors required to compute the Cauchy stress tensor (2.2) are

$$\mathbf{a} \otimes \mathbf{a} = \begin{bmatrix} (\cos\theta + U\sin\theta)^2 & (\cos\theta + U\sin\theta)\sin\theta & 0 \\ (\cos\theta + U\sin\theta)\sin\theta & \sin^2\theta & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{2.8}$$

and

$$\mathbf{D} = \frac{1}{2}\begin{bmatrix} 0 & U_t & 0 \\ U_t & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{2.9}$$

so that the nonzero components of $\mathbf{T}$ are $T_{33} = -p + \mu$ and

$$\begin{aligned} &T_{11} = -p + \mu(1 + U^2) + \mu\gamma_0(I_4 - 1)(\cos\theta + U\sin\theta)^2, \\ &T_{22} = -p + \mu + \mu\gamma_0(I_4 - 1)\sin^2\theta, \\ &T_{12} = \mu U + \mu\gamma_0(I_4 - 1)(\cos\theta + U\sin\theta)\sin\theta + \nu[1 + \gamma_1(I_4 - 1)]U_t. \end{aligned} \tag{2.10}$$

Now the equations of motion div $\mathbf{T} = \rho\mathbf{x}_{tt}$ reduce to the two scalar equations $-p_x + T_{12,y} = \rho u_{tt}$ and $-p_y + T_{22,y} = \rho u_{tt}$. Differentiating the former with respect to y and the latter with respect to x, and eliminating p_{xy}, we arrive at a single governing equation for the rectilinear shear motion:

$$\begin{aligned} \rho U_{tt} = \mu U_{yy} &+ \mu\gamma_0 \sin^2\theta \left[U(2\cos\theta + U\sin\theta)(\cos\theta + U\sin\theta)\right]_{yy} \\ &+ \nu U_{tyy} + \nu\gamma_1 \sin\theta \left[UU_t(2\cos\theta + U\sin\theta)\right]_{yy}. \end{aligned} \tag{2.11}$$

Using the scalings $\tilde{t} = \mu t/\nu$ and $\tilde{y} = y/L$ (where L is a characteristic length to be specified later on a case-by-case basis), we write this equation in dimensionless form as

$$\begin{aligned} \varepsilon U_{\tilde{t}\tilde{t}} = U_{\tilde{y}\tilde{y}} &+ \gamma_0 \sin^2\theta \left[U(2\cos\theta + U\sin\theta)(\cos\theta + U\sin\theta)\right]_{\tilde{y}\tilde{y}} \\ &+ U_{\tilde{t}\tilde{y}\tilde{y}} + \gamma_1 \sin\theta \left[UU_t(2\cos\theta + U\sin\theta)\right]_{\tilde{y}\tilde{y}}, \end{aligned} \tag{2.12}$$

where $\varepsilon = \rho\mu L^2/\nu^2$. This is the main equation of our study. For convenience we drop the tildes in the remainder of the paper. We also introduce the functions

$$\begin{aligned} &f(\gamma_0, U, \theta) = 1 + \gamma_0 \sin^2\theta(2\cos\theta + U\sin\theta)(\cos\theta + U\sin\theta), \\ &g(\gamma_1, U, \theta) = 1 + \gamma_1 U\sin\theta(2\cos\theta + U\sin\theta), \end{aligned} \tag{2.13}$$

so that (2.13) is now

$$\varepsilon U_{tt} = \left[Uf(\gamma_0, U, \theta) + U_t g(\gamma_1, U, \theta)\right]_{yy}. \tag{2.14}$$

3. Nonlinear anisotropic recovery. Our first investigation is placed in the quasi-static approximation, where we study the influence of elastic anisotropy and viscous anisotropy on the classic experiment of viscous *recovery*. We imagine that the material is sheared and that at $t = 0$ the shear stress is removed: $T_{12}(0) = 0$. Here the characteristic length L is the displacement at $t = 0$ from which the material will relax to the unstressed state.

In the quasi-static case, we neglect the inertia term of (2.13) and may thus integrate it twice to give the following first-order ordinary differential equation:

$$Uf(\gamma_0, U, \theta) + U_t g(\gamma_1, U, \theta) = 0. \tag{3.1}$$

Here we take the constants of integration to be zero, according to the context of recovery, as explained above. We then solve the equation as

$$\int \frac{g(\gamma_1, U, \theta)}{Uf(\gamma_0, U, \theta)} \mathrm{d}U = -t + \text{const.}, \tag{3.2}$$

where the constant is computed so that $U(0) = 1$.

When $\theta = 0$, the fibers are not active with respect to the deformation, and we recover the classical result of isotropic viscoelastic recovery: $U(t) = \mathrm{e}^{-t}$.

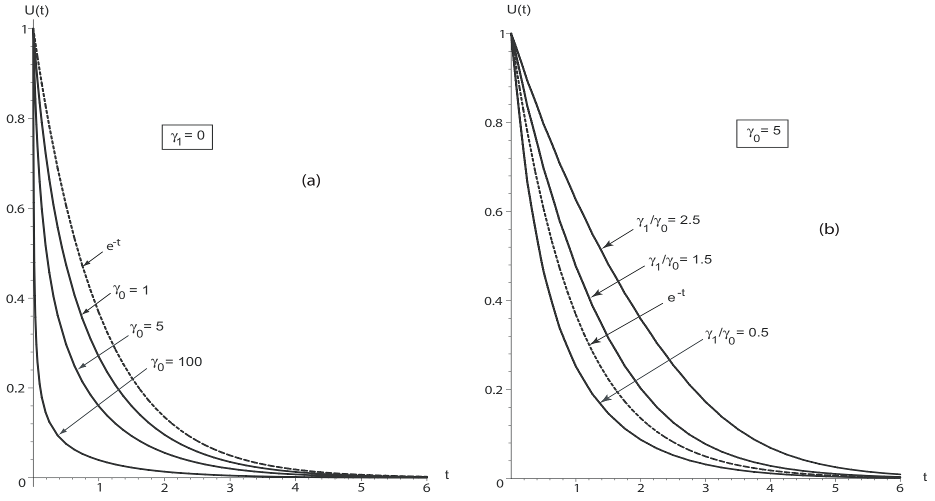


FIG. 3.1. *Time recovery function when* $\theta = \pi/2$: (a) $\gamma_1 = 0$ *and* $\gamma_0 = 1$, 5, 100; (b) $\gamma_0 = 5$ *and* $\gamma_1 = 0.5$, 1.5, 2.5. *The recovery function for an isotropic solid is also plotted (dotted curve).*

When $\theta = \pi/2$, the anisotropic effects are at their strongest. In that case the integral above has a compact expression, and we find

$$U\left[\frac{1+\gamma_0 U^2}{1+\gamma_0}\right]^{\frac{1}{2}\left(\frac{\gamma_1}{\gamma_0}-1\right)} = \mathrm{e}^{-t}. \tag{3.3}$$

We now take $\gamma_1 = 0$ (no anisotropic viscosity) and $\gamma_0 = 1, 5, 100$ (recall that the fibers are inextensible in the limit $\gamma_0 \to \infty$). Figure 3.1(a) shows that as the anisotropic effect becomes more pronounced, the recovery is quicker; in other words, the influence of elasticity becomes stronger as γ_0 increases. Then we fix γ_0 at 5, for instance, and look at the role played by the anisotropic viscosity, by taking in turn $\gamma_1/\gamma_0 = 0.5$, 1.5, 2.5. We find in Figure 3.1(b) that, as expected, the viscous recovery is slower as γ_1 increases.

When $\theta \neq 0$, $\theta \neq \pi/2$, other behaviors arise, which call for a detailed analysis. In particular, the exponential, or near-exponential, decay toward zero as $t \to \infty$ is not necessarily ensured, especially when the anisotropic effects are strong and the fibers are oriented at a large angle $\theta > \pi/2$. Clearly, $U_t = 0$ when $f = 0$, according to (3.2). Also, $U_t < 0$ when f and g are of the same sign, and $U_t > 0$ when f and g are of opposite signs. These two functions are quadratic in U. If they have no real roots in U, then they are both of the positive sign and $U_t < 0$. (This is clearly the case in the region $0 < \theta < \pi/2$.) If they have real roots, then they may change sign, and U might be an *increasing* function of t. This happens for f and for g when $\pi/2 < \theta < \pi$ and

$$\gamma_0 \geq \frac{4}{\cos^2\theta \sin^2\theta}, \quad \gamma_1 \geq \frac{1}{\cos^2\theta}, \tag{3.4}$$

respectively. In Figure 3.2, the region C corresponds to the first inequality, where the delimiting curve has a vertical asymptote at $\theta = \pi/2$, a vertical asymptote at $\theta = \pi$, and a minimum at $\theta = \pi/4$, $\gamma_0 = 16$; we recall that Qiu and Pence (1997) showed that

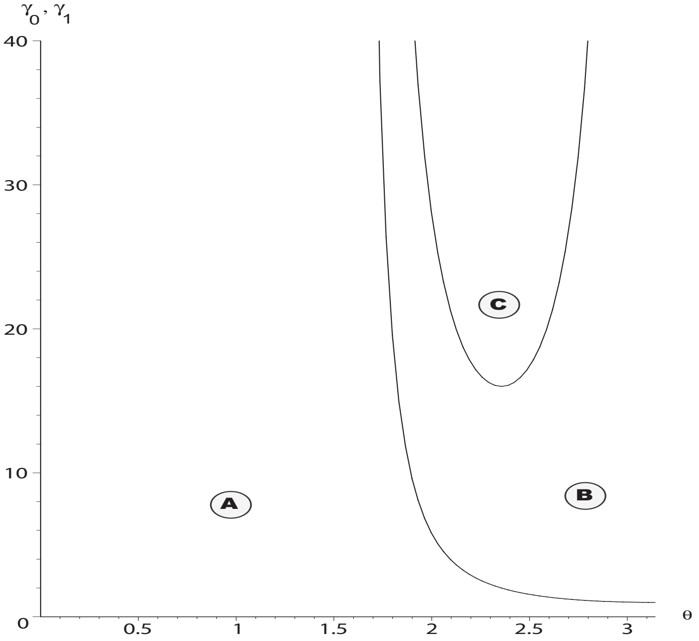


FIG. 3.2. *Recovery: regions where the sign of U_t may change.*

when $\gamma_0 > 16$, "simple shear at certain fiber orientations involves negative shear stress in the shearing direction for certain positive shears." The region B corresponds to the second inequality, where the delimiting curve has a vertical asymptote at $\theta = \pi/2$ and an horizontal asymptote at $\gamma_1 = 1$. In the region A, neither inequality is satisfied.

3.1. Weak anisotropy. First, we take both γ_0 and γ_1 in region A. This is the simplest case because f and g are then both positive, and so U_t is always negative (damped recovery). We took several representative examples in this region (say, $\theta = \pi/4$, $\gamma_0 = 20$, $\gamma_1 = 1$) and checked, through integration and implicit plotting, that the graphs are indeed of the same nature as those in Figure 3.1.

3.2. Strong elastic anisotropy. Second, we take γ_1 in region A, by fixing it at $\gamma_1 = 1$, say. In that region, $g > 0$ always, and thus the sign of U_t is the opposite of the sign of f. Then we take $\gamma_0 = 20$, which is above the minimum of region C. In Figure 3.3, we plot the locus for the values of U as functions of θ such that $f(20, U, \theta) = 0$. Outside the resulting oval shape, $f > 0$, and inside, $f < 0$. We also plotted the line $U = 1$, which intersects the oval at $\theta_{\text{min}} = 2.136$ and $\theta_{\text{max}} = 2.221$. Recall that $U(0) = 1$.

When $\theta > \theta_{\text{max}}$, $U(t)$ starts at 1 and decreases because $f > 0$ so that $U_t < 0$; as U decreases toward 0, U_t tends to zero according to $(3.2)_1$, but takes an infinite time to do so, according to $(3.2)_2$; hence $U = 0$ is a horizontal asymptote and the recovery is "classical"; see plot (i) in Figure 3.3, traced at $\theta = 2.4$ (notice, however, that the recovery is not exponential because the second derivative of U clearly changes sign as t increases, in contrast with e^{-t}, traced in dotted lines).

When $\theta_{\text{min}} < \theta < \theta_{\text{max}}$, $U(t)$ starts at 1 and then grows until it hits the upper side of the oval, taking an infinite time to do so; then this upper bound gives a horizontal asymptotic value, *above the initial value* (see plot (ii) in Figure 3.3), traced at $\theta = 2.2$.

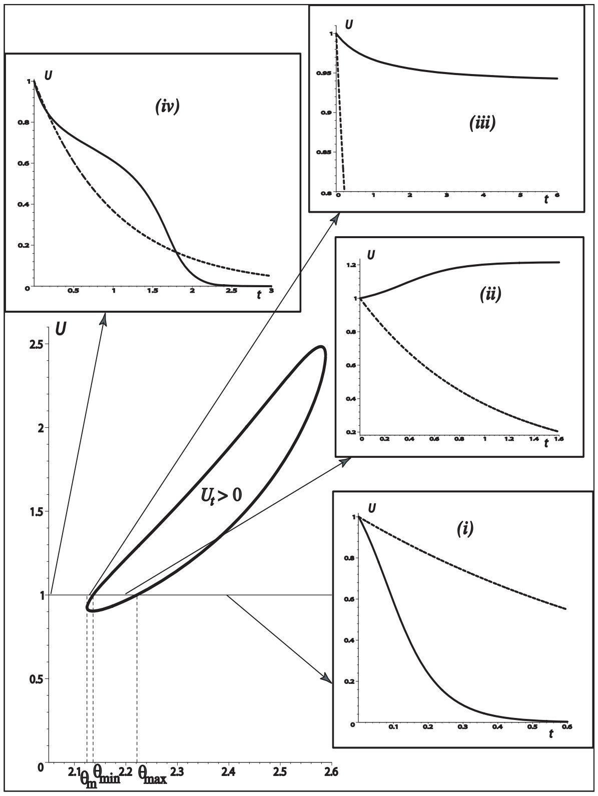


Fig. 3.3. *Types of time recovery functions for $\gamma_0 = 20$, $\gamma_1 = 1$ (strong elastic anisotropy). The amount of shear U starts at 1 for $t = 0$. Outside the oval shape, $U_t < 0$ and U decreases as in* (i) *and* (iv): *decay toward zero*; *and in* (iii): *decay toward a value > 0. Inside the oval shape, $U_t > 0$ and U increases as in* (ii): *growth toward a value > 1. The recovery function e^{-t} for an isotropic solid is also plotted (dotted curves).*

When $\theta_{\mathrm{m}} < \theta < \theta_{\min}$, where $\theta_{\mathrm{m}} = 2.124$ is the angle at which the oval plot has a vertical tangent, $U(t)$ starts at 1 and then decreases until it hits the upper side of the oval, below 1 but above 0; then this lower bound gives a horizontal asymptotic value, *above zero* (see plot (iii) in Figure 3.3), traced at $\theta = 2.125$.

Finally, when $\theta < \theta_{\mathrm{m}}$, $U(t)$ starts at 1 and then decreases until zero; then this lower bound gives zero as a horizontal asymptotic value (see plot (iv) in Figure 3.3), traced at $\theta = 2.05$. Notice that the second derivative changes signs three times as t increases.

3.3. Strong viscous anisotropy. Third, we take γ_0 outside the C region, by fixing it at $\gamma_0 = 1$, say. In that region, $f > 0$ always, and thus the sign of U_t is the opposite of the sign of g. Then we allow γ_1 to be in region B, and thus allow g (and U_t) to change sign with increasing θ, by taking $\gamma_1 = 3.0$, say. In Figure 3.4, we plotted the values of U as functions of θ such that $g(3, U, \theta) = 0$ and obtained the thick-line shape. Outside the shape, $g > 0$, and inside, $g < 0$. We also plotted the horizontal line $U = 1$, which intersects the shape at $\theta_{\min} = 2.356$ and $\theta_{\max} = 2.820$, and the vertical line $\theta = \theta_{\mathrm{m}} = 2.186$, which is tangent to the shape.

Now when $\theta < \theta_{\mathrm{m}}$ or $\theta > \theta_{\max}$, $U(t)$ starts at 1 and decreases until zero; as $U \to 0$, the denominator in the integral tends to zero, indicating that it takes an

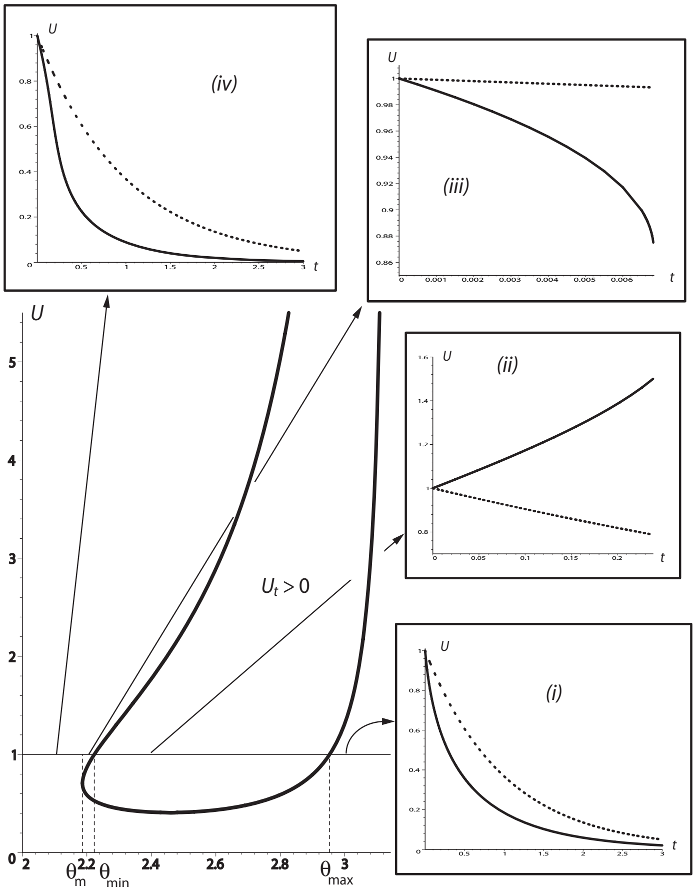


FIG. 3.4. *Types of time recovery functions for $\gamma_0 = 1$, $\gamma_1 = 2$ (strong viscous anisotropy). The amount of shear U starts at 1 for $t = 0$. Outside the thick-line shape, $U_t < 0$ and U decreases toward zero as in* (i) *and* (iv)*. Inside the thick-line shape, $U_t > 0$ and U increases rapidly as in* (ii)*, until it ceases to exist. There is also an angular region $\theta_m < \theta < \theta_{min}$ where U decreases rapidly until it ceases to exist*; *see* (iii)*. The recovery function e^{-t} for an isotropic solid is also plotted (dotted curves).*

infinite time to do so; hence, zero is a horizontal asymptote in these cases. To draw Figure 3.4(i) we took $\theta = 3.0$, and for Figure 3.4(iv) we took $\theta = 2.1$; both graphs show a somewhat classical decay with time.

However, when $\theta_{\min} < \theta < \theta_{\max}$, $U(t)$ starts at 1 and then *grows* because $U_t > 0$ inside the thick line shape. Eventually U hits the upper face of the shape, where $g = 0$; then by (3.1), either $Uf = 0$ or $U_t \to \infty$. Clearly, the first possibility is excluded because $U \neq 0$ when it is larger than 1, and $f \neq 0$ when γ_0 is outside the C region. It follows that U grows and hits the upper face of the shape with a vertical asymptote after a finite time (and then stops because it cannot increase further since $U_t < 0$ outside the shape, it cannot remain constant since $U_t \neq 0$ on the shape, and it cannot decrease since $U_t > 0$ inside the shape). Figure 3.4(ii) shows such behavior for $U(t)$, traced at $\theta = 2.4$.

Finally, when $\theta_{\rm m} < \theta < \theta_{\min}$, $U(t)$ decays from 1 until it hits the shape from above after a finite time; see Figure 3.4(iii), traced at $\theta = 2.2$. Notice how quickly the final value is reached, compared to the isotropic exponential recovery.

3.4. Strong elastic and viscous anisotropies. In the case where both γ_0 and γ_1 are in the region C, any combination and overlaps of the thick curves presented in Figures 3.3 and 3.4 may arise. The tools presented in the two previous subsections are easily transposed to those possibilities. A special situation arises when the locus of $f = 0$ intersects the locus of $g = 0$; then, the numerator and the denominator in (3.2) may have a common factor so that the integrand simplifies and a regular behavior may appear. This situation is, however, too special to warrant further investigation, and we do not pursue this line of enquiry.

4. Nonlinear anisotropic creep. Our second investigation is again placed in the quasi-static approximation, where we now study the influence of elastic anisotropy and viscous anisotropy on the classic experiment of viscous *creep*. As the resulting analysis is similar to that conducted for recovery, we simply outline the main results.

We imagine that the material is sheared and that the shear stress is maintained: $T_{12}(\infty) \neq 0$. Here the characteristic length L is an asymptotic value of the displacement. We neglect the inertial term of (2.13) and integrate it twice to give the ordinary differential equation

$$U f(\gamma_0, U, \theta) + U_t g(\gamma_1, U, \theta) = \text{const.}, \tag{4.1}$$

where we took the constant of the first integration to be zero and the constant of the second integration to correspond to the applied (constant) shear stress, as is usual in the creep problem. More specifically, this constant is taken so that $U(\infty) = 1$, and so is equal to $f(\gamma_0, 1, \theta)$; it follows that the equation above can be written as

$$h(\gamma_0, U, \theta)(U - 1) + g(\gamma_1, U, \theta)U_t = 0, \tag{4.2}$$

where h is defined by

$$\begin{aligned} h(\gamma_0, U, \theta) &= [U f(\gamma_0, U, \theta) - f(\gamma_0, 1, \theta)]/(U - 1) \\ &= 1 + \gamma_0 \sin^2\theta[1 + \cos^2\theta + (U + 1)\sin\theta(U \sin\theta + 3\cos\theta)]. \end{aligned} \tag{4.3}$$

We then solve the equation as

$$\int \frac{g(\gamma_1, U, \theta)}{(U - 1)h(\gamma_0, U, \theta)} \mathrm{d}U = -t + \text{const.}, \tag{4.4}$$

where the constant is computed so that $U(0) = 0$. Hence the equations governing creep are almost identical to those governing recovery, with the difference that f is now replaced by h.

Here we are mostly concerned with the question of how, if at all, a state of shear can be reached such that, once removed, the unusual recovery behaviors of the previous section emerge. Thus we concentrate on strong anisotropic effects, with emphasis on strong elastic anisotropy (where the new function h is involved). We traced the regions where g and h, and thus U_t, may change signs and found that the resulting graph is similar to that of Figure 3.2, with the main difference that the minimum of region C is now located at $\theta = 3\pi/4$ and $\gamma_0 = 4$. Thus unusual behavior in creep may occur at much lower levels of elastic anisotropy than in recovery (where the minimum is at $\gamma_0 = 16$). We recall that Qiu and Pence (1997) showed that when $\gamma_0 > 4$, "simple shear at certain fiber orientations involves a nonmonotonic relation between the shear stress in the shearing direction and the amount of shear."

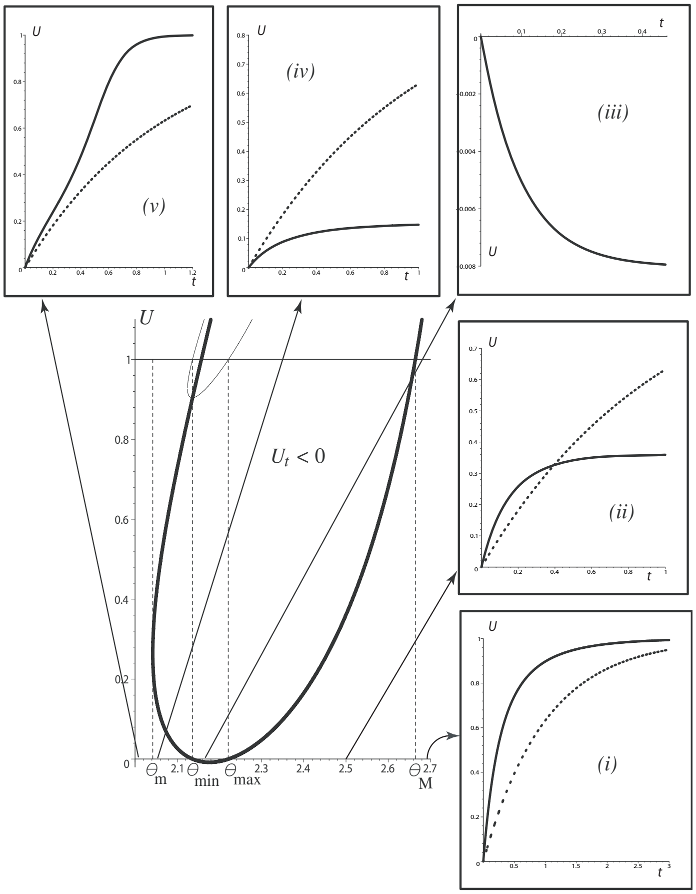


FIG. 4.1. *Types of time creep functions for $\gamma_0 = 20$, $\gamma_1 = 1$ (strong elastic anisotropy). The amount of shear U starts at* 0 *for $t = 0$. Outside the thick-line shape, $U_t > 0$ and U increases as in* (i) *and* (v)*: growth toward* 1*, and as in* (ii) *and* (iv)*: growth toward a value below* 1*. Inside the thick-line shape, U decreases as in* (iii)*: decay toward a negative value. The creep function $1 - e^{-t}$ for an isotropic solid is also plotted (dotted curves).*

4.1. Strong elastic anisotropy. We begin with the case where h plays a major role, that is, when γ_0 is greater than 4. For the purpose of direct comparison with the recovery problem, we take $\gamma_0 = 20$ and $\gamma_1 = 1$, as in section 3.2. Figure 4.1 displays the curve where $h(20, U, \theta) = 0$. Outside the thick-line curve, $U_t > 0$, and inside, $U_t < 0$. The curve intersects the line $U = 0$ twice, at $\theta_{\min} = 2.136$ and at $\theta_{\max} = 2.221$. These are the values at which $f = 0$ intersects $U = 1$ in section 3.2 (see the thin-line shape), because by (4.3), $h(20, 0, \theta_{\min}) = f(20, 1, \theta_{\min}) = 0$ and similarly $h(20, 0, \theta_{\max}) = f(20, 1, \theta_{\max}) = 0$. We also display the vertical lines $\theta = \theta_M = 2.664$, where $h = 0$ intersects $U = 1$, and $\theta = \theta_m = 2.042$, where $h = 0$ has a vertical tangent. Recall that for creep, $U(0) = 0$.

When $\theta > \theta_{\mathrm{M}}$, $U(t)$ starts at 0 and grows toward 1; then U_t tends to zero according to (4.2) but takes an infinite time to do so; hence $U = 1$ is a horizontal asymptote and the creep is "classical." See plot (i) in Figure 4.1, traced at $\theta = 2.7$ (the exponential creep function of isotropic visco-elasticity $(1 - \mathrm{e}^{-t})$ is shown by the dotted line).

When $\theta_{\max} < \theta < \theta_{\mathrm{M}}$ or when $\theta_{\mathrm{m}} < \theta < \theta_{\min}$, $U(t)$ starts at 0 and then grows until it hits the oval shape, taking an infinite time to do so; then this upper bound gives a horizontal asymptotic value, *below* 1; see plot (ii) in Figure 4.1, traced at $\theta = 2.5$, and plot (iv), traced at $\theta = 2.05$.

When $\theta_{\text{min}} < \theta < \theta_{\text{max}}$, $U(t)$ starts at 0 inside the oval shape, and thus it *decreases* until it hits the lower side of the shape, taking an infinite time to do so; then this lower bound gives a horizontal asymptotic value, *below* 0; see plot (iii) in Figure 4.1, traced at $\theta = 2.17$.

Finally, when $\theta < \theta_{\text{m}}$, $U(t)$ can again grow toward 1; see plot (v) in Figure 4.1, traced at $\theta = 2.0$. Notice, however, that the concavity of the curve changes as t increases.

4.2. Strong viscous anisotropy. Here we remark that the function governing the strength of the viscous anisotropy, namely g, is the same for creep as it is for recovery. Thus, the region where U_t might change sign because of strong viscous anisotropy is the region B of Figure 3.1. Also, the locus of points where $g = 0$ is typically displayed by the thick-line shape of Figure 3.4, and because $g(\gamma_1, 0, \theta) = 1 > 0$ always, this curve never crosses the abscissa $U = 0$. It follows that there is only one situation where viscous anisotropy leads to anomalous creep, when $\theta_{\text{min}} < \theta < \theta_{\text{max}}$; then $U(t)$ starts at zero and grows toward the thick-line shape, which it reaches after a finite time with a vertical asymptote.

4.3. Prestretch and nonlinear anisotropic creep. Here we show how anomalous creep can be avoided (amplified) by stretching (compressing) the solid prior to the shear. Hence, instead of (2.4), we consider the motion

$$x = \lambda^{-\frac{1}{2}} X + \lambda u(Y,t), \quad y = \lambda Y, \quad z = \lambda^{-\frac{1}{2}} Z. \tag{4.5}$$

The following decomposition of the associated deformation gradient shows that the solid is stretched by a ratio λ in the Y direction:

$$\mathbf{F} = \mathbf{F_2}\mathbf{F_1}, \quad \text{where} \quad \mathbf{F_2} = \begin{bmatrix} 1 & U & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{F_1} = \begin{bmatrix} \lambda^{-\frac{1}{2}} & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda^{-\frac{1}{2}} \end{bmatrix}. \tag{4.6}$$

(Note that $\mathbf{F_2}\mathbf{F_1} \neq \mathbf{F_1}\mathbf{F_2}$.) The kinematic quantities of section 2.2 are modified accordingly. In particular,

$$I_4 = \lambda^{-1}\cos^2\theta + \lambda^2\sin^2\theta + U\lambda^{\frac{1}{2}}\sin 2\theta + U^2\lambda^2\sin^2\theta. \tag{4.7}$$

The end result is that the differential equation governing creep is changed from (4.2) to

$$h^\lambda(\gamma_0, U, \theta)(U-1) + g^\lambda(\gamma_1, U, \theta)U_t = 0, \tag{4.8}$$

where h^λ and g^λ are defined by

$$\begin{aligned} h^\lambda(\gamma_0, U, \theta) = \lambda^2 \{ 1 + \gamma_0 \sin^2\theta [& 2\lambda^2\sin^2\theta + 3\lambda^{-1}\cos^2\theta - 1 \\ & + (U+1)\sin\theta(U\sin\theta + 3\cos\theta)] \}, \end{aligned}$$

$$g^\lambda(\gamma_1, U, \theta) = 1 + \gamma_1(\lambda^{-1}\cos^2\theta + \lambda^2\sin^2\theta - 1 + U\lambda^{\frac{1}{2}}\sin 2\theta + U^2\lambda^2\sin^2\theta). \tag{4.9}$$

Figure 4.2 shows the loci of $h^\lambda = 0$ in the case of a strong elastic anisotropy ($\gamma_0 = 30$,

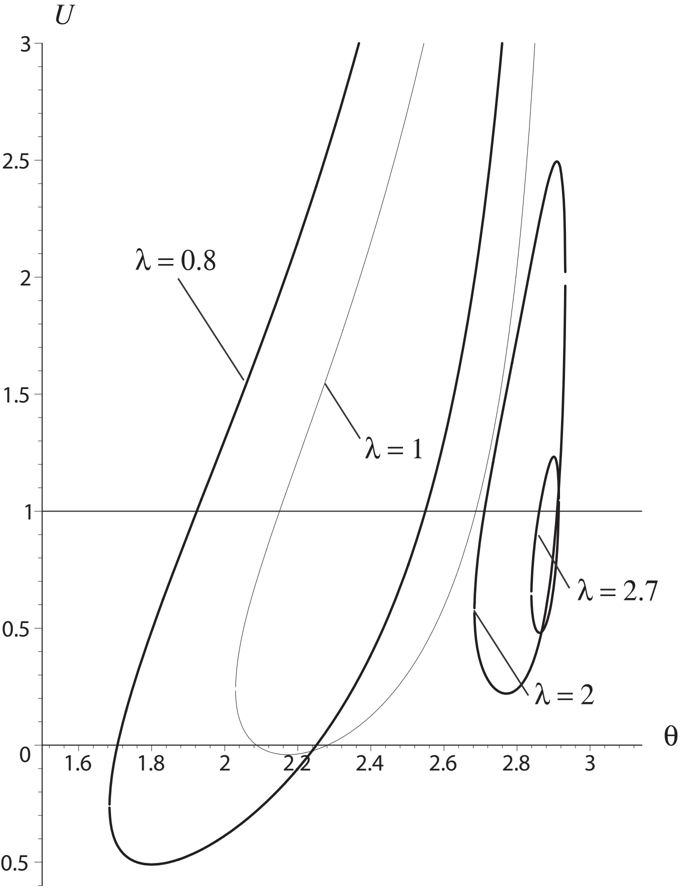


FIG. 4.2. *Effect of prestretch on time creep functions for $\gamma_0 = 30$, $\gamma_1 = 1$ (strong elastic anisotropy). The amount of shear U starts at 0 for $t = 0$. Inside the thick-line shapes, U decreases; this situation may arise when the solid is not prestretched ($\lambda = 1$) or when it is compressed ($\lambda = 0.8$). Outside the thick-line shapes, $U_t > 0$ and U increases; this situation may arise when the solid is in extension ($\lambda = 2$, $\lambda = 2.7$).*

$\gamma_1 = 1$), for several values of λ. The figure clearly shows that the prestretch λ can be used to control the shape of these curves: if the solid is put in compression first, and sheared for creep next, then the region of potential anomalous creep is increased; if it is put under tension, then the area of the region rapidly decreases and eventually disappears altogether.

5. Nonlinear traveling waves. So far we have looked at how the presence of elastic and viscous fibers affects some quasi-static processes. Typically, creep and recovery connect one state of constant shear (initial) to another (final). Now we examine another class of solutions connecting two constant states of shear, this time *dynamically*, by looking for traveling wave (kink) solutions.

The mathematical theory of one-dimensional transverse traveling waves in isotropic viscoelastic materials with a Kelvin–Voigt type of constitutive equation is well grounded; see, for example, Nishihara (1995) for a clear and complete mathematical approach, or Jordan and Puri (2005) for a specific and explicit example. A traveling wave is a solution to the equations of motion in the form

$$U(Y,t) = U(\xi), \qquad \xi = Y - ct, \tag{5.1}$$

where c is the constant speed; also, U is such that

$$\lim_{\xi \to -\infty} U(\xi) = U_L, \qquad \lim_{\xi \to \infty} U(\xi) = U_R, \tag{5.2}$$

where U_L and U_R are distinct constants. In what follows, we focus on the case where $U_L = 0$, $U_R = 1$. This case is general up to a rigid translation. Here we take the displacement corresponding to U_R as the characteristic length L.

Substituting (5.1) into (3.2), we obtain

$$\varepsilon c^2 U'' = (Uf - cU'g)'', \tag{5.3}$$

and then by integration,

$$cU'g = \left(f - \varepsilon c^2\right) U + \text{const.} \tag{5.4}$$

By the requirement $U_L = 0$, the constant must be zero. By the requirement $U_R = 1$, we have

$$f(\gamma_0, 1, \theta) = \varepsilon c^2. \tag{5.5}$$

This equation prompts three remarks.

First, we must ensure that $f(\gamma_0, 1, \theta) > 0$. Recall that, according to (2.13),

$$f(\gamma_0, 1, \theta) = 1 + \gamma_0 \sin^2\theta(2\cos\theta + \sin\theta)(\cos\theta + \sin\theta), \tag{5.6}$$

and so

$$\partial f(\gamma_0, 1, \theta)/\partial\theta = \gamma_0 \sin\theta(4\cos^3\theta + 9\sin\theta\cos^2\theta - 3\sin^3\theta). \tag{5.7}$$

In Figure 5.1(a) we plot the variations of $[f(\gamma_0, 1, \theta) - 1]/\gamma_0$ with θ, as well as those of its derivative with respect to θ (scaled to 1/8). Clearly, the function (5.6), viewed as a function of θ, has an absolute minimum and an absolute maximum. The minimum is at $\hat{\theta}$, say, such that $\tan\hat{\theta}$ is that root of the cubic $4 + 9x - 3x^3 = 0$ corresponding to $\pi/2 < \hat{\theta} < \pi$; numerically, $\hat{\theta} = 2.1777$. Then, solving $f(\gamma_0, 1, \hat{\theta}) = 0$ for γ_0, we find that $f(\gamma_0, 1, \theta) > 0$ when $0 < \gamma_0 < \hat{\gamma_0} = 18.490$; and that when $\gamma_0 > \hat{\gamma_0}$, there appears a range for θ where $f(\gamma_0, 1, \theta) > 0$ is not insured. Placing ourselves outside that possibility, we deduce from (5.5) that, for a given γ_0 and a given θ, the wave travels with speed

$$c = \pm\sqrt{f(\gamma_0, 1, \theta)/\varepsilon}. \tag{5.8}$$

This is of course expressed in the dimensionless variables of length$/L$ and time$\times\mu/\nu$. Turning back, if required, to physical variables, we would find that the wave travels with the dimensional speed $\sqrt{\mu f(\gamma_0, 1, \theta)/\nu}$.

The second remark is that, according to (5.5) and (5.6), the wave (when it exists) travels with maximum speed at the angle $\tilde{\theta}$, say, such that $\tan\tilde{\theta}$ is that root of the cubic $4 + 9x - 3x^3 = 0$ corresponding to $0 < \tilde{\theta} < \pi/2$; numerically, $\tilde{\theta} = 1.0910$. Hence the directions of extremal speeds of propagation are always the same, whatever the values of the constitutive parameters μ, γ_0, and γ_1. This observation indicates the way for

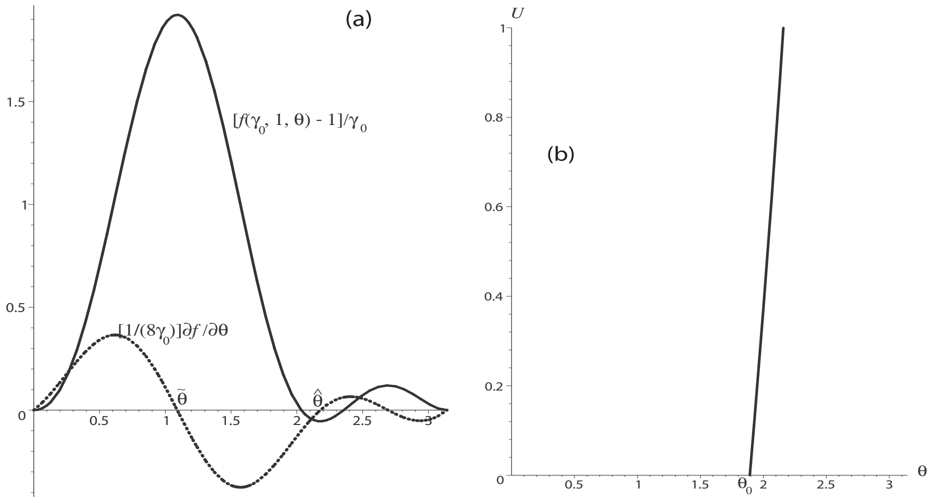


FIG. 5.1. (a) *Variations with* θ *of* $[f(\gamma_0, 1, \theta) - 1]/\gamma_0$ *and of its derivative, showing an absolute minimum at* $\hat{\theta} = 2.1777$. (b) *Variations of* $-3/\tan\theta - 1$ *with* θ, *crossing the abscissa line at* $\theta_0 = 1.8926$.

an acoustic determination of the fiber orientation: if an experimental measurement of the shear wave speed can be made in every direction of a fiber-reinforced viscoelastic nonlinear material, then the fibers are at an angle $\hat{\theta}$ from the direction of the slowest wave and at an angle $\tilde{\theta}$ from the direction of the fastest wave. We recall that for waves in an *isotropic* deformed neo-Hookean material, Ericksen (1953) found that the fastest waves propagate along the direction of greatest initial stretch.

The third remark is that when (5.5) holds,

$$f(\gamma_0, U, \theta) - \varepsilon c^2 = \gamma_0 U(U-1)\sin^3\theta\,[U\sin\theta + 3\cos\theta + \sin\theta]\,. \tag{5.9}$$

Then the separation of variables, followed by integration of the first-order differential equation (5.4), leads to

$$\int \frac{g(\gamma_1, U, \theta)}{U(U-1)\sin^3\theta[U\sin\theta + 3\cos\theta + \sin\theta]}\mathrm{d}U = \frac{\gamma_0}{c}\xi + \text{const.}, \tag{5.10}$$

where the constant of integration is arbitrary; without loss of generality, we take it to be such that $U(0) = 1/2$.

Clearly, critical issues arise when either the numerator or the denominator change signs (because then U' changes sign, and it might not be possible to find a solution satisfying the requirements (5.2)). We may take care of the numerator's sign by considering only elastic anisotropy ($\gamma_0 \neq 0$) and discarding viscous anisotropy ($\gamma_1 = 0$); then $g = 1$. For the denominator, however, we note that $U\sin\theta + 3\cos\theta + \sin\theta$ can change sign for certain ranges of U and θ. Figure 5.1(b) shows the curve $U = -3/\tan\theta - 1$; on its left side, the denominator is positive; on its right side, it is negative. Accordingly, the wave connects 0 to 1 (see Figure 5.2(a)) or is unable to do so (see Figure 5.2(b)). In that latter case, the wave front grows toward an asymptotic value which is less than 1; a second solution exists (dotted curve) with 1 as an asymptotic value, but in the $\xi \to -\infty$ direction.

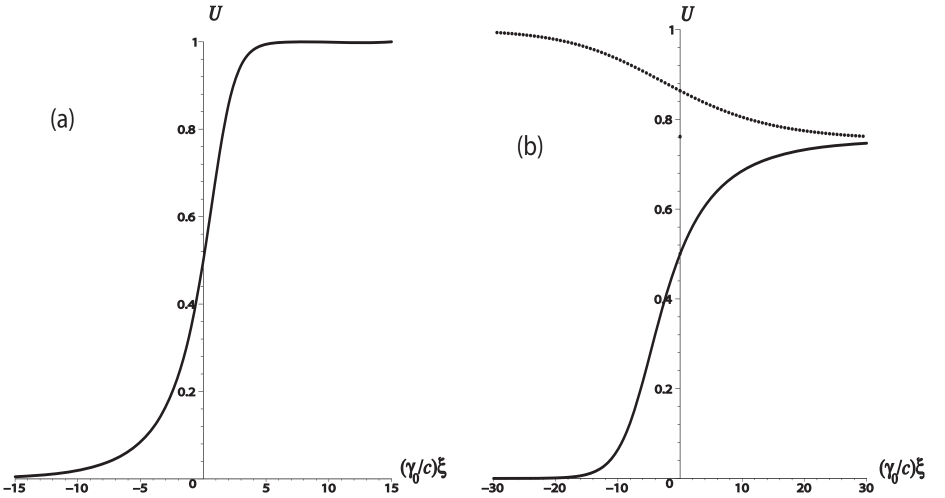


FIG. 5.2. *Traveling wave solution for anisotropic elasticity* ($\gamma_1 = 0$): (a) *at* $\theta = 1.8$, (b) *at* $\theta = 2.1$.

As a final remark, we note that when γ_1 is large enough to allow for the possibility that $g = 0$ (strong viscous anisotropy), then a "singular barrier" arises; see Pettet, McElwain, and Norbury (2000).

6. Discussion.

In the course of this investigation on nonlinear anisotropic creep, recovery, and waves for fiber-reinforced nonlinear elastic materials, we unearthed some complex mechanical responses. For some range of the constitutive parameters and for some angle ranges of the fiber arrangement, we saw that unusual and possibly aberrant behaviors can emerge.

From a mathematical point of view, we gave a detailed explanation of the reasons for these behaviors, by linking them to the singularities of the determining equations for the amount of shear.

From the mechanical point of view, we pointed out that nonstandard behaviors always occur when the angle between the fiber family and the direction of shear is such that the fibers are compressed; see Figure 2.1. It has been widely demonstrated that several types of instabilities may develop in the case of fiber contraction; see the detailed studies by Triantafyllidis and Abeyaratne (1983), Qiu and Pence (1997), Merodio and Ogden (2002), (2003), (2005a), (2005b), or Fu and Freidin (2004). For example, Merodio, Saccomandi, and Sgura (2007) recently investigated a nonhomogeneous rectilinear shear static deformation for the standard reinforcing model (2.1) and found nonregular solutions (that is, deformations characterized by a discontinuous amount of shear) in fiber-contracted materials.

From a numerical point of view, we recall that a simple model, together with a simple class of solutions, allows a step-by-step control of the simulations. It would indeed be hard to detect nonstandard behaviors by relying solely on a complex numerical finite element method (omitting to conduct a simple analytical methodology such as the one presented in this paper). For example, Holzapfel and Gasser (2001) present a detailed computational study of some viscoelastic fiber-reinforced nonlinear materials, but use values for the material parameters and for the angles which place their simulations outside the problematic ranges. Other studies are placed in the framework of linear models (even for polymeric materials; see Liu, Kasyanov, and Schoephoerster (2007)), which fail to capture nonstandard behaviors.

From an experimental point of view, our results suggest some simple yet revealing protocols. In particular, it would be most valuable to investigate the existence and the persistence of asymptotic residual shear strains, sustained after the shear stress is removed, at levels not only below the value at initial time but also above (as in section 3). So far we have identified only reports of experimental results concerned with elastomeric materials reinforced with *inextensible* fibers (and therefore with a ratio between the shear modulus of the bulk matrix and that of the fibers of several orders of magnitude), or concerned with moderate angles between the direction of shear and the fiber direction.

From a biomechanical point of view, the results have meaningful implications for biological soft tissues. First, the model captures adequately the elastic and the viscous anisotropies of biological materials (Baldwin et al. (2006), Taylor et al. (1990)). Second, although anomalous creep behaviors might preclude anomalous recovery behaviors, it is still useful to study the latter, because they might nonetheless arise in vivo following a stress-driven fiber orientation remodeling (Hariton et al. (2006)). Third, the effect of the prestretch on nonstandard behaviors is significant theoretically (section 4.3) as well as practically. (In vivo experiments show that large static prestretches of tendons reduce the risk of unexpected behaviors; see Kubo, Kanehisa, and Fukunaga (2002).) Finally, the results of the traveling wave study (section 5) may eventually lead to an acoustic (elastographic) determination of the fiber angle in soft tissues, through an efficient, simple, and noninvasive investigation.

Obviously, our results must be improved, and several directions are possible. Hence, two families of fibers have to be considered to give a better comparison with in vivo results for soft tissues. Also, the more realistic models of fiber reinforcements (such as the one proposed by Horgan and Saccomandi (2005) and by Gasser, Ogden, and Holzapfel (2006)) must be incorporated into the present study, to identify with a greater precision the range of parameters for which strange behaviors may occur.

Acknowledgment. G. S. thanks Gérard Maugin, Head of the Institut Jean Le Rond d'Alembert, for his hospitality.

REFERENCES

S. L. Baldwin, K. R. Marutyan, M. Yang, K. D. Wallace, M. R. Holland, and J. G. Miller (2006), *Measurements of the anisotropy of ultrasonic attenuation in freshly excised myocardium*, J. Acoust. Soc. Amer., 119, pp. 3130–3139.

J. L. Ericksen (1953), *On the propagation of waves in isotropic incompressible perfectly elastic materials*, J. Ration. Mech. Anal., 2, pp. 329–337.

Y. B. Fu and A. B. Freidin (2004), *Characterization and stability of two-phase piecewise-homogeneous deformations*, Proc. Royal Soc. Lond. A, 460, pp. 3065–3084.

T. C. Gasser, R. W. Ogden, and G. A. Holzapfel (2006), *Hyperelastic modelling of arterial layers with distributed collagen fibre orientations*, J. Roy. Soc. Interfaces, 3, pp. 15–25.

N. Gundiah, M. B. Ratcliffe, and L. A. Pruitt (2007), *Determination of strain energy function for arterial elastin: Experiments using histology and mechanical tests*, J. Biomech., 40, pp. 586–594.

I. Hariton, G. Debotton, T. C. Gasser, and G. A. Holzapfel (2006), *Stress-driven collagen fiber remodeling in arterial walls*, TU Graz Online preprint reports, 68, pp. 1–26.

G. A. Holzapfel and T. C. Gasser (2001), *A viscoelastic model for fiber-reinforced composites at finite strains: Continuum basis, computational aspects and applications*, Comput. Methods Appl. Mech. Engrg., 190, pp. 4379–4430.

C. O. Horgan and G. Saccomandi (2005), *A new constitutive theory for fiber-reinforced incompressible nonlinearly elastic solids*, J. Mech. Phys. Solids, 53, pp. 1985–2025.

J. D. Humphrey (2002), *Cardiovascular Solid Mechanics, Cells, Tissues and Organs*, Springer, New York.

P. M. Jordan and A. Puri (2005), *A note on traveling wave solutions for a class of nonlinear viscoelastic media*, Phys. Lett. A, 335, pp. 150–156.

K. Kubo, H. Kanehisa, and T. Fukunaga (2002), *Effect of stretching training on the viscoelastic properties of human tendon structures in vivo*, J. Appl. Physiol., 92, pp. 595–601.

Y. Liu, K. Kasyanov, and R. T. Schoephoerster (2007), *Effect of fiber orientation on the stress distribution within a leaflet of a polymer composite hearth valve in the closed position*, J. Biomech., 40, pp. 1099–1106.

J. Merodio and R. W. Ogden (2002), *Material instabilities in fiber-reinforced nonlinearly elastic solids under plane deformation*, Arch. Mech., 54, pp. 525–552.

J. Merodio and R. W. Ogden (2003), *Instabilities and loss of ellipticity in fiber-reinforced compressible non-linearly elastic solids under plane deformation*, Int. J. Solids Struct., 40, pp. 4707–4727.

J. Merodio and R. W. Ogden (2005a), *Mechanical response of fiber-reinforced incompressible nonlinear elastic solids*, Int. J. Nonlinear Mech., 40, pp. 213–227.

J. Merodio and R. W. Ogden (2005b), *On tensile instabilities and ellipticity loss in fiber-reinforced incompressible nonlinearly elastic solids*, Mech. Res. Comm., 32, pp. 290–299.

J. Merodio, G. Saccomandi, and I. Sgura (2007), *The rectilinear shear of fiber-reinforced incompressible non-linearly elastic solids*, Int. J. Non-Linear Mech., 42, pp. 342–354.

K. Nishihara (1995), *Stability of traveling waves with degenerate shock for systems of one-dimensional viscoelastic model*, J. Differential Equations, 120, pp. 304–318.

G. J. Pettet, D. L. S. McElwain, and J. Norbury (2000), *Lotka-Volterra equations with chemotaxis: Walls, barriers and travelling waves*, IMA J. Math. Control Inform., 17, pp. 395–413.

G. Y. Qiu and T. J. Pence (1997), *Remarks on the behavior of simple directionally reinforced incompressible nonlinearly elastic spheres*, J. Elasticity, 49, pp. 1–30.

D. C. Taylor, J. D. Dalton, A. V. Seaber, and W. E. Garrett (1990), *Viscoelastic properties of muscle-tendon units. The biomechanical effects of stretching*, Amer. J. Sports Med., 18, pp. 300–309.

N. Triantafyllidis and R. Abeyaratne (1983), *Instabilities of a finitely deformed fiber-reinforced elastic material*, ASME J. Appl. Mech., 50, pp. 149–156.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 98–108

ACTIVATION THROUGH A NARROW OPENING[*]

A. SINGER[†] AND Z. SCHUSS[‡]

Abstract. The escape of a Brownian motion through a narrow absorbing window in an otherwise reflecting boundary of a domain is a rare event. In the presence of a deep potential well, there are two long time scales, the mean escape time from the well and the mean time to reach the absorbing window. We derive a generalized Kramers formula for the mean escape time through the narrow window.

Key words. stochastic differential equations, exit problem, narrow escape, mixed boundary value problem

AMS subject classifications. 60H10, 60J60, 35J25

DOI. 10.1137/060663477

1. Introduction.

Kramers' theory [12], [6] concerns the thermal activation of a Brownian particle over a high potential barrier. It assumes that the barrier height is much larger than the thermal energy. Its application to the theory of chemical kinetics [20] gives the activation rate of the stochastic dynamics of a reactant molecule over a potential barrier ΔE as the Arrhenius law

$$k = Ae^{-\Delta E/k_BT}, \tag{1.1}$$

where A is a function of temperature, friction, and the potential landscape. A similar, but different situation arises, if the chemical reaction can be described as the diffusion of a Brownian particle through a small opening in the boundary of a domain, whose remaining boundaries are practically reflecting. Such a situation can occur, if the reflecting boundaries are due to a high potential barrier with a small opening, whose energy is not necessarily much higher than the thermal energy. This can happen, for example, if the reflecting boundaries are due to a dielectric barrier, as in biological membranes, and the small opening is a protein channel embedded in an otherwise impenetrable membrane [7]. The small absorbing window setup is also a model for the forward rate of chemical reactions, in which there are small binding sites for the diffusing reacting molecule in the boundary of the domain [9]. The same setup also describes the process of trafficking receptors on biological membranes [8]. The escape of a free Brownian motion (without drift) through a small window was discussed in [17], [18], [19]. Here we consider the narrow escape problem for a Brownian motion in a field of force. The closely related problem of computing the principal eigenvalue of the Laplace operator for mixed boundary conditions on large and small pieces of the boundary was considered in [22], [23], [24], [11] (see section 6 for discussion).

We derive an Arrhenius-like formula (1.1) for the activation rate through narrow openings. Specifically, we consider the diffusion of a Brownian particle in a potential

[*] Received by the editors June 21, 2006; accepted for publication (in revised form) July 6, 2007; published electronically October 10, 2007.
http://www.siam.org/journals/siap/68-1/66347.html

[†] Department of Mathematics, Program in Applied Mathematics, Yale University, 10 Hillhouse Ave., P.O. Box 208283, New Haven, CT 06520-8283 (amit.singer@yale.edu).

[‡] Department of Applied Mathematics, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel (schuss@post.tau.ac.il). The research of this author was partially supported by a grant from the US-Israel Binational Foundation.

field in a bounded domain Ω, where activation occurs if the particle goes through a small opening $\partial\Omega_a$ in the boundary $\partial\Omega$ of the domain. We assume that the remaining boundary $\partial\Omega_r$ reflects the Brownian trajectories. We find the dependence of the rate constant on the potential, specific geometry of the opening and on the volume or surface area of the domain. As in Kramers' theory, we obtain different rate constants for low and high barriers. The activation rates for the different geometries are summarized in (4.6)–(4.13).

2. Formulation. As in classical theories [12], [6], [20], our point of departure is the Langevin dynamics in $\mathbb{R}^n$ $(n = 2, 3)$,

$$m\ddot{\boldsymbol{x}} + \eta\dot{\boldsymbol{x}} + \nabla\Phi(\boldsymbol{x}) = \sqrt{2\eta k_B T}\dot{\boldsymbol{w}}, \tag{2.1}$$

where m is the mass, η is the friction coefficient, $\Phi(x)$ is the potential, T is temperature, k_B is Boltzmann's constant, and $\dot{\boldsymbol{w}}$ is a vector of n independent δ-correlated Gaussian white noises. In the Smoluchowski (Kramers) limit of large friction, the Langevin dynamics (2.1) reduces to the Smoluchowski equation [16], [4], [6]

$$\dot{\boldsymbol{x}} + \frac{1}{\gamma}\nabla\phi(\boldsymbol{x}) = \sqrt{\frac{2k_B T}{m\gamma}}\,\dot{\boldsymbol{w}}, \tag{2.2}$$

where $\gamma = \eta/m$ is the dynamics viscosity and $\phi = \Phi/m$ is the potential per unit mass.

The motion of the Brownian particle is confined to a bounded domain Ω, whose boundary $\partial\Omega$ is reflecting, but for a small absorbing window $\partial\Omega_a$ $(\partial\Omega = \partial\Omega_a \cup \Omega_r)$. The assumption that the window is small means that

$$\delta = \left(\frac{|\partial\Omega_a|}{|\partial\Omega|}\right)^{1/(n-1)} \ll 1 \tag{2.3}$$

(δ is a small parameter).

The probability density function (pdf) $p_\delta(\boldsymbol{x}, t)$ of finding the Brownian particle at location $\boldsymbol{x}$ at time t satisfies the Fokker–Planck equation

$$\gamma\frac{\partial p_\delta}{\partial t} = \varepsilon\Delta p_\delta + \nabla\cdot(p_\delta\nabla\phi) \equiv \mathcal{L}_\delta p_\delta, \tag{2.4}$$

with the initial condition

$$p_\delta(\boldsymbol{x}, 0) = p_0(\boldsymbol{x}), \tag{2.5}$$

and the mixed Dirichlet–Neumann boundary conditions for $t > 0$

$$p_\delta = 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_a, \tag{2.6}$$

$$\varepsilon\frac{\partial p_\delta}{\partial n} + p_\delta\frac{\partial\phi}{\partial n} = 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_r, \tag{2.7}$$

where $\varepsilon = k_B T/m$, $\boldsymbol{n}$ is the unit outer normal at the boundary, and $p_0(\boldsymbol{x})$ is the initial pdf (e.g., $p_0(\boldsymbol{x}) = \frac{1}{|\Omega|}$ for a uniform distribution). The function

$$u_\delta(\boldsymbol{x}) = \int_0^\infty p_\delta(\boldsymbol{x}, t)\,dt, \tag{2.8}$$

which is the mean time the particle spends at $\boldsymbol{x}$ before it escapes through the narrow window, is the solution of the boundary value problem

$$\mathcal{L}_\delta u_\delta = -\gamma p_0 \quad \text{for} \quad \boldsymbol{x} \in \Omega, \tag{2.9}$$

$$u_\delta = 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_a, \tag{2.10}$$

$$\varepsilon \frac{\partial u_\delta}{\partial n} + u_\delta \frac{\partial \phi}{\partial n} = 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_r. \tag{2.11}$$

The function $g_\delta = u_\delta e^{\phi/\varepsilon}$ is the solution of the adjoint problem

$$\mathcal{L}_\delta^* g_\delta = -\gamma p_0 e^{\phi/\varepsilon} \quad \text{for} \quad \boldsymbol{x} \in \Omega, \tag{2.12}$$

$$\begin{aligned} \frac{\partial g_\delta(\boldsymbol{x})}{\partial n} &= 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_r, \\ g_\delta(\boldsymbol{x}) &= 0 \quad \text{for} \quad \boldsymbol{x} \in \partial\Omega_a. \end{aligned} \tag{2.13}$$

Equation (2.12) can be written in the divergence form

$$\nabla \left(e^{-\phi/\varepsilon} \nabla g_\delta \right) = -\frac{\gamma p_0}{\varepsilon}. \tag{2.14}$$

The adjoint operators $\mathcal{L}_\delta$ and $\mathcal{L}_\delta^*$, defined by (2.4), (2.9), (2.10), (2.11), and (2.12), (2.13), respectively, have biorthogonal systems of normalized eigenfunctions, $\{\psi_i(\boldsymbol{x},\delta)\}$ and $\{\varphi_i(\boldsymbol{x},\delta)\}$ $(i = 0, 1, \ldots)$, and we can expand

$$p_\delta(\boldsymbol{x}, t) = \sum_{i=0}^{\infty} a_i(\delta) \psi_i(\boldsymbol{x}, \delta) e^{-\lambda_i(\delta) t/\gamma}, \tag{2.15}$$

where $\lambda_i(\delta)$ are the eigenvalues of $\mathcal{L}_\delta$. The $a_i(\delta)$ are the Fourier coefficients of the initial function $p_0(\boldsymbol{x})$. In the limit $\delta \to 0$ the Dirichlet part of the boundary conditions, (2.6), is dropped, so that $\lambda_0(\delta) \to 0$ (the first eigenvalue of the problem (2.4), (2.7) with $\partial\Omega_r = \partial\Omega$), with the normalized eigenfunction

$$\psi_0(\boldsymbol{x}, 0) = \frac{\exp\{-\phi(\boldsymbol{x})/\varepsilon\}}{\displaystyle\int_\Omega \exp\{-\phi(\boldsymbol{x})/\varepsilon\}\, d\boldsymbol{x}}, \tag{2.16}$$

and $a_0(\delta) \to 1$. It follows from (2.8) and (2.15) that for all $\boldsymbol{x} \in \Omega$

$$u_\delta(\boldsymbol{x}) = \gamma \sum_{i=0}^{\infty} \frac{a_i(\delta)\psi_i(\boldsymbol{x},\delta)}{\lambda_i(\delta)} \to \infty \quad \text{as} \quad \delta \to 0. \tag{2.17}$$

In particular, the first passage time $\tau_\delta = \inf\{t > 0 \mid \boldsymbol{x}(t) \in \partial\Omega_a\}$ diverges. That is, $\lim_{\delta\to 0} \tau_\delta = \infty$ on almost every trajectory $\boldsymbol{x}(t)$. Obviously, the mean first passage time,

$$\langle \tau_\delta \rangle = \int_\Omega u_\delta(\boldsymbol{x})\, d\boldsymbol{x} = \gamma \sum_{i=0}^{\infty} \frac{a_i(\delta)}{\lambda_i(\delta)}, \tag{2.18}$$

also diverges as $\delta \to 0$. It is the purpose of this paper to find the orders of magnitude of $u_\delta(\boldsymbol{x})$ and $\langle \tau_\delta \rangle$ for small δ.

3. The Neumann function. The Neumann function for Ω is the solution of the boundary value problem

$$\begin{aligned} \Delta_{\boldsymbol{y}} N(\boldsymbol{x},\boldsymbol{y}) &= -\delta(\boldsymbol{x}-\boldsymbol{y}) \quad \text{for} \quad \boldsymbol{x},\boldsymbol{y}\in\Omega, \\ \frac{\partial N(\boldsymbol{x},\boldsymbol{y})}{\partial n_{\boldsymbol{y}}} &= -\frac{1}{|\partial\Omega|} \quad \text{for} \quad \boldsymbol{x}\in\Omega,\ \boldsymbol{y}\in\partial\Omega, \end{aligned} \tag{3.1}$$

with $N(\boldsymbol{x},\boldsymbol{y})$ fixed at a given point, to ensure uniqueness. Using Green's identity and the boundary conditions (2.10)–(2.11) and (3.1) gives

$$\begin{aligned} &\int_\Omega N(\boldsymbol{x},\boldsymbol{y})\Delta_{\boldsymbol{y}} u_\delta(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int_\Omega u_\delta(\boldsymbol{y})\Delta_{\boldsymbol{y}} N(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y} + \int_{\partial\Omega}\left(N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}} - u_\delta(\boldsymbol{y})\frac{\partial N(\boldsymbol{x},\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\right)dS_{\boldsymbol{y}} \\ &= -u_\delta(\boldsymbol{x}) + \int_{\partial\Omega_a} N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\,dS_{\boldsymbol{y}} - \frac{1}{\varepsilon}\int_{\partial\Omega_r} N(\boldsymbol{x},\boldsymbol{y})u_\delta(\boldsymbol{y})\frac{\partial\phi(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\,dS_{\boldsymbol{y}} \\ &\quad + \frac{1}{|\partial\Omega|}\int_{\partial\Omega_r} u_\delta(\boldsymbol{y})\,dS_{\boldsymbol{y}}. \end{aligned} \tag{3.2}$$

On the other hand, (2.9) gives

$$\begin{aligned} &\int_\Omega N(\boldsymbol{x},\boldsymbol{y})\Delta_{\boldsymbol{y}} u_\delta(\boldsymbol{y})\,d\boldsymbol{y} \\ &\quad = \int_\Omega N(\boldsymbol{x},\boldsymbol{y})\left[-\frac{\gamma p_0}{\varepsilon} - \frac{1}{\varepsilon}\nabla\cdot(u_\delta\nabla\phi)\right]d\boldsymbol{y} \\ &\quad = -\frac{\gamma}{\varepsilon}\int_\Omega N(\boldsymbol{x},\boldsymbol{y})p_0(\boldsymbol{y})\,d\boldsymbol{y} - \frac{1}{\varepsilon}\int_\Omega \nabla_{\boldsymbol{y}}\cdot[N(\boldsymbol{x},\boldsymbol{y})u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})]\,d\boldsymbol{y} \\ &\qquad + \frac{1}{\varepsilon}\int_\Omega u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}N(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y} \\ &\quad = -\frac{\gamma}{\varepsilon}\int_\Omega N(\boldsymbol{x},\boldsymbol{y})p_0(\boldsymbol{y})\,d\boldsymbol{y} - \frac{1}{\varepsilon}\int_{\partial\Omega_r} N(\boldsymbol{x},\boldsymbol{y})u_\delta(\boldsymbol{y})\frac{\partial\phi(\boldsymbol{y})}{\partial n}\,dS_{\boldsymbol{y}} \\ &\qquad + \frac{1}{\varepsilon}\int_\Omega u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}N(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}. \end{aligned} \tag{3.3}$$

Combining (3.2) and (3.3) yields

$$\begin{aligned} &-u_\delta(\boldsymbol{x}) + \frac{1}{|\partial\Omega|}\int_{\partial\Omega_r} u_\delta(\boldsymbol{y})\,dS_{\boldsymbol{y}} + \int_{\partial\Omega_a} N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\,dS_{\boldsymbol{y}} \\ &= -\frac{\gamma}{\varepsilon}\int_\Omega N(\boldsymbol{x},\boldsymbol{y})p_0(\boldsymbol{y})\,d\boldsymbol{y} + \frac{1}{\varepsilon}\int_\Omega u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}N(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}. \end{aligned} \tag{3.4}$$

In view of (2.17), the integral $\int_\Omega N(\boldsymbol{x},\boldsymbol{y})p_0(\boldsymbol{y})\,d\boldsymbol{y}$ can be neglected to leading order, because it is uniformly bounded for smooth initial distributions[1] p_0 as $\delta\to 0$, while all other terms in (3.4) are unbounded. For $\boldsymbol{x}\in\Omega$, at a distance $O(1)$ away from the window, the Neumann function is uniformly bounded.

[1] For nonsmooth p_0 the integral is not uniformly bounded. For example, for $p_0 = \delta(\boldsymbol{x}-\boldsymbol{x}_0)$ we have $\int_\Omega N(\boldsymbol{x},\boldsymbol{y})p_0(\boldsymbol{y})\,d\boldsymbol{y} = N(\boldsymbol{x},\boldsymbol{x}_0)$, which becomes singular as $\boldsymbol{x}\to\boldsymbol{x}_0$. However, this is an integrable singularity, and as such it does not affect the leading order asymptotics in δ.

Note that integrating (2.14) and using the boundary conditions (2.13), we obtain the compatibility condition

$$\int_{\partial\Omega_a} \frac{\partial u_\delta}{\partial n}\, dS = -\frac{\gamma}{\varepsilon}. \tag{3.5}$$

Because of the fact that the normal derivative $\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}$ is negative on $\partial\Omega_a$, (3.5) implies that $\int_{\partial\Omega_a} N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\, dS_{\boldsymbol{y}}$ is uniformly bounded. It follows that for $\boldsymbol{x} \in \Omega$, at a distance $O(1)$ (with respect to δ) away from the window, the integral equation (3.4) is to leading order

$$u_\delta(\boldsymbol{x}) \sim \frac{1}{|\partial\Omega|}\int_{\partial\Omega} u_\delta(\boldsymbol{y})\, dS_{\boldsymbol{y}} - \frac{1}{\varepsilon}\int_\Omega u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot\nabla N(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y}, \tag{3.6}$$

which is the integral representation of the boundary value problem $\mathcal{L}_\delta u_\delta = 0$ with the no flux boundary condition (2.11) on the entire boundary (i.e., with $\partial\Omega_r = \partial\Omega$), whose solution is the Boltzmann distribution

$$u_\delta(\boldsymbol{x}) \sim C_\delta e^{-\phi(\boldsymbol{x})/\varepsilon}. \tag{3.7}$$

Equation (3.7) represents the averaged time the particle spent at a point $\boldsymbol{x}$ at a distance $O(1)$ away from the absorbing window prior to absorption.

Due to the absorbing boundary condition (2.10), (3.4) reduces to

$$\begin{aligned}\int_{\partial\Omega_a} & N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\, dS_{\boldsymbol{y}} \\ &= \left\{\frac{-1}{|\partial\Omega|}\int_{\partial\Omega_r} u_\delta(\boldsymbol{y})\, dS_{\boldsymbol{y}} + \frac{1}{\varepsilon}\int_\Omega u_\delta(\boldsymbol{y})\nabla_{\boldsymbol{y}}\phi(\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}N(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y}\right\}(1+o(1))\end{aligned} \tag{3.8}$$

for all $\boldsymbol{x} \in \partial\Omega_a$. Substituting (3.7) into (3.8) yields an integral equation for the flux $\frac{\partial u_\delta}{\partial n}$ into the absorbing window,

$$\int_{\partial\Omega_a} N(\boldsymbol{x},\boldsymbol{y})\frac{\partial u_\delta(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\, dS_{\boldsymbol{y}} = -C_\delta e^{-\phi(\boldsymbol{x})/\varepsilon}(1+o(1)) \quad \text{for} \quad \delta \ll 1. \tag{3.9}$$

If $\phi(\boldsymbol{x})$ does not change much in the window, we can use the constant approximation $\phi(\boldsymbol{x}) \approx \phi(\text{window}) = \phi_0$.

In three dimensions

$$N(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{4\pi|\boldsymbol{x}-\boldsymbol{y}|} + v_S(\boldsymbol{x},\boldsymbol{y}), \tag{3.10}$$

where v_S is a regular harmonic function [10], and so the leading order contribution to (3.9) is due to the singular part of the Neumann function. Thus the leading order approximation $\frac{\partial u_0}{\partial n}$ to the absorption flux is the solution of

$$\frac{1}{2\pi}\int_{\partial\Omega_a} \frac{\partial u_0(\boldsymbol{y})}{\partial n_{\boldsymbol{y}}}\frac{dS_{\boldsymbol{y}}}{|\boldsymbol{x}-\boldsymbol{y}|} = -C_\delta e^{-\phi_0/\varepsilon}. \tag{3.11}$$

Note that the singularity of the Neumann function at the boundary is twice as large as it is inside the domain, due to the contribution of the regular part (the "image charge"). For that reason the factor $\frac{1}{4\pi}$ in (3.10) was replaced by $\frac{1}{2\pi}$.

4. Narrow escape. von Helmholtz [21] (see also Rayleigh [1] and others, e.g., [13]) solved the integral equation (3.11) analytically for the case of an elliptical absorbing window $\partial\Omega_a$,

$$\frac{\partial u_0(y_1,y_2)}{\partial n} = -\frac{C_\delta e^{-\phi_0/\varepsilon}}{\sqrt{1-\frac{y_1^2}{a^2}-\frac{y_2^2}{b^2}}}, \tag{4.1}$$

where a and b are the ellipse semiaxes, and $\boldsymbol{y} = (y_1, y_2)$ are local Cartesian coordinates in the ellipse. The value of the constant C_δ is calculated using the compatibility condition (3.5) to be

$$C_\delta = \frac{\gamma K(e)}{2\pi\varepsilon a} e^{\phi_0/\varepsilon}, \tag{4.2}$$

where e is the eccentricity of the ellipse and $K(\cdot)$ is the complete elliptic integral of the first kind. In a three-dimensional domain, the averaged time spent at point $\boldsymbol{x}$ before escape through an elliptical absorbing window is given by (see (3.7))

$$u_\delta(\boldsymbol{x}) \approx \frac{\gamma K(e)}{2\pi\varepsilon a} \exp\left\{\frac{\phi_0-\phi(\boldsymbol{x})}{\varepsilon}\right\}. \tag{4.3}$$

Equations (2.18) and (4.3) now give the mean escape time as

$$\langle\tau_\delta\rangle = \frac{\gamma K(e) e^{\phi_0/\varepsilon}}{2\pi\varepsilon a} \int_\Omega \exp\left\{-\frac{\phi(\boldsymbol{x})}{\varepsilon}\right\} d\boldsymbol{x}. \tag{4.4}$$

If the barrier is sufficiently high, we evaluate the integral in (4.4) by the Laplace method, assuming that ϕ has a single global minimum ϕ_m at $\boldsymbol{x}_m$,

$$\int_\Omega \exp\left\{-\frac{\phi(\boldsymbol{x})}{\varepsilon}\right\} d\boldsymbol{x} \approx \frac{(2\pi\varepsilon)^{n/2}}{\prod_{i=1}^{n}\omega_i} \exp\left\{-\frac{\phi_m}{\varepsilon}\right\}, \tag{4.5}$$

where ω_i are the frequencies at the minimum $\boldsymbol{x}_m$. For reactions that consist in passing through a small elliptical window (assuming no returns are possible), the reaction rate is the modified Kramers formula

$$\kappa_\delta = \frac{1}{\langle\tau_\delta\rangle} \sim \frac{a\omega_1\omega_2\omega_3}{\sqrt{2\pi\varepsilon}\,\gamma K(e)}\, e^{-\Delta E/\varepsilon}, \tag{4.6}$$

where $\Delta E = \phi_0 - \phi_m$. In the special case of a circular window, we obtain

$$\kappa_\delta \sim \frac{4a\omega_1\omega_2\omega_3}{(2\pi)^{3/2}\gamma\sqrt{\varepsilon}}\, e^{-\Delta E/\varepsilon}, \tag{4.7}$$

where a is the radius of the window. Note that ΔE is not the barrier height. We conclude that the activation rate is of Arrhenius form and has two contributions. The first is due to the potential, while the second is due to geometry of the absorbing window alone. Unlike the free diffusion case [17], [18], [19], geometrical properties of the domain, such as its volume, are not included in the leading order asymptotics of the reaction rate.

Second, in the limit of large ε, the power series approximation

$$e^{-(\phi(\boldsymbol{x})-\phi_0)/\varepsilon}=1-\frac{\phi(\boldsymbol{x})-\phi_0}{\varepsilon}+\frac{(\phi(\boldsymbol{x})-\phi_0)^2}{2\varepsilon^2}\cdots$$

in (4.4) gives

$$k\sim\frac{2\pi\varepsilon a}{\gamma K(e)|\Omega|}\left(1-\frac{\langle\phi\rangle-\phi_0}{\varepsilon}+O\left(\varepsilon^{-2}\right)\right)^{-1}, \tag{4.8}$$

where $\langle\phi\rangle=\frac{1}{|\Omega|}\int_\Omega\phi(\boldsymbol{x})\,d\boldsymbol{x}$ is the spatial average of the potential. The rate can also be rewritten into an Arrhenius form as

$$k\sim\frac{2\pi\varepsilon a}{\gamma K(e)|\Omega|}\,e^{-\langle\Delta E\rangle/\varepsilon}, \tag{4.9}$$

where $\langle\Delta E\rangle=\phi_0-\langle\phi\rangle$. In the case of large ε the reaction rate depends not merely on the geometry of the window but also on the geometry of the domain itself through its volume. Large ε means that the motion is diffusion limited; therefore, fine details of the potential are less important and the spatial averaged potential has only an $O(\varepsilon^{-1})$ effect.

Finally, we give rate functions for small and large ε for several geometries. For the case of diffusion in a ball of radius R, the results of [17] show that

$$\begin{aligned} k&\sim\frac{4\varepsilon a}{\gamma|\Omega|}\left[1+\frac{a}{R}\ln\frac{R}{a}+O\left(\frac{a}{R}\right)\right]^{-1}e^{-\langle\Delta E\rangle/\varepsilon}\quad\text{for}\quad\varepsilon\gg\Delta E,\\ k&\sim\frac{4\varepsilon a\omega_1\omega_2\omega_3}{\gamma(2\pi)^{3/2}}\left[1+\frac{a}{R}\ln\frac{R}{a}+O\left(\frac{a}{R}\right)\right]^{-1}e^{-\Delta E/\varepsilon}\quad\text{for}\quad\varepsilon\ll\Delta E.\end{aligned} \tag{4.10}$$

We conjecture that the second order term is $O(\delta\ln\delta)$ also for a general three-dimensional domain, though we were unable to prove it so far.

In two dimensions the singularity of the Neumann function is logarithmic, and so the leading order approximation to the activation rate is

$$\begin{aligned} k&\sim\frac{\pi\varepsilon}{\gamma|\Omega|}\frac{e^{-\langle\Delta E\rangle/\varepsilon}}{\left[\ln\frac{1}{\delta}+O(1)\right]}\quad\text{for}\quad\varepsilon\gg\Delta E,\\ k&\sim\frac{\varepsilon\sqrt{\omega_1\omega_2}}{2\gamma}\frac{e^{-\Delta E/\varepsilon}}{\left[\ln\frac{1}{\delta}+O(1)\right]}\quad\text{for}\quad\varepsilon\ll\Delta E.\end{aligned} \tag{4.11}$$

The remainder $O(1)$ is important, because in real life applications even if δ is small, $\ln\frac{1}{\delta}$ is not necessarily large. In [18], [19] we have calculated the $O(1)$ term for diffusion in a circular disk, in a circular annulus, and on a sphere. These results extend in a straightforward way to domains that can be mapped conformally onto these shapes (e.g., all simply connected planar domains).

If the boundary of the absorbing window contains a singular point of $\partial\Omega$, such as a corner or a cusp, the order of magnitude of the activation rate may change. Thus, if the window is at a corner of angle α, then the rate is [19]

$$
\begin{aligned}
k &\sim \frac{\alpha\varepsilon}{\gamma|\Omega|}\frac{e^{-\langle\Delta E\rangle/\varepsilon}}{\left[\ln\dfrac{1}{\delta}+O(1)\right]} \quad \text{for} \quad \varepsilon \gg \Delta E,\\
k &\sim \frac{\alpha\varepsilon\sqrt{\omega_1\omega_2}}{2\pi\gamma}\frac{e^{-\Delta E/\varepsilon}}{\left[\ln\dfrac{1}{\delta}+O(1)\right]} \quad \text{for} \quad \varepsilon \ll \Delta E.
\end{aligned}
\tag{4.12}
$$

If the absorbing window is near a cusp, then $\langle\tau_\delta\rangle$ grows algebraically rather than logarithmically. For example, in the domain bounded between two tangent circles, the activation rate is

$$
\begin{aligned}
k &\sim \frac{(d^{-1}-1)\varepsilon}{\gamma|\Omega|}\left[\delta+O(\delta^2)\right]e^{-\langle\Delta E\rangle/\varepsilon} \quad \text{for} \quad \varepsilon \gg \Delta E,\\
k &\sim \frac{(d^{-1}-1)\varepsilon\sqrt{\omega_1\omega_2}}{2\pi\gamma}\left[\delta+O(\delta^2)\right]e^{-\Delta E/\varepsilon} \quad \text{for} \quad \varepsilon \ll \Delta E,
\end{aligned}
\tag{4.13}
$$

where $d<1$ is the ratio of the radii.

5. Deep well—a Markov chain model. The modified Kramers formulas (4.6) or (4.11) can be explained by coarse-graining the diffusive motion into a simplified 3-state Markov model, when the domain contains a deep well $\Omega_W \subset \Omega$. The three states of the Markov process are (i) state W—the trajectory is trapped in the deep well; (ii) state D—the trajectory diffuses in the domain $\Omega_D = \Omega - \Omega_W$, outside the well; (iii) state A—the trajectory is absorbed into the small hole. Once the trajectory is absorbed into the small hole, its motion is terminated, and so A is a terminal state of the Markov chain. For simplicity, we assume $\Omega \subset \mathbb{R}^2$.

Not all transition times between the different states are finite with probability 1, and so not all mean transition times are finite. The particle leaves the well to the outer in finite mean time, that is,

$$
\Pr\{\tau_{W\to D}<\infty\}=1, \quad \mathbb{E}\tau_{W\to D}<\infty. \tag{5.1}
$$

For small ε, the mean time spent in the well, $\mathbb{E}\tau_{W\to D}$, is exponentially large and is given by [14]

$$
\mathbb{E}\tau_{W\to D} \sim \frac{2\pi\sqrt{\dfrac{\partial^2\phi(x_S)}{\partial s^2}}}{\sqrt{-\dfrac{\partial^2\phi(x_S)}{\partial\nu^2}}\sqrt{H(x_W)}}\exp\left\{\frac{\phi(x_S)-\phi(x_W)}{\varepsilon}\right\}, \tag{5.2}
$$

where ν and s are the distance to and arclength on $\partial\Omega_W$, respectively, x_W is the deepest point of the well, x_S is the point on $\partial\Omega_W$, where ϕ achieves its minimum, and H is the Hessian of ϕ.

The time $\tau_{D\to W}$, however, is not finite with probability 1, because there is a finite probability $\Pr\{\tau_{D\to A}<\tau_{D\to W}\}$ of termination at A without returning to W, and there is no return from A to W. Consequently, $\mathbb{E}\tau_{D\to W}=\infty$. However, $\mathbb{E}\tau_{D\to A}$ and $\mathbb{E}[\tau_{D\to W} \mid \tau_{D\to W}<\tau_{D\to A}]$ are finite. For small ε,δ, the conditional mean time $\mathbb{E}[\tau_{D\to W} \mid \tau_{D\to W}<\tau_{D\to A}]$ is asymptotically the same as $\mathbb{E}\tau_{D\to W}$ for a problem without the small absorbing window, because the conditioning changes the drift only near A, to repel the trajectory from the window, and so the effect on the conditional mean time is small, regardless of whether this mean time is long or short. The transition probabilities from the outer domain to the absorbing window and to the well are

$$\Pr\{\tau_{D\to A}<\tau_{D\to W}\}\sim\frac{\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]}{\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]+\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]},$$
$$\Pr\{\tau_{D\to W}<\tau_{D\to A}\}\sim\frac{\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]}{\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]+\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]},\tag{5.3}$$

respectively. The conditional mean transition time $\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]$ from Ω_D to Ω_W is similar to (5.2),

$$\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]\sim\frac{2\pi\sqrt{\dfrac{\partial^2\phi(x_S)}{\partial s^2}}}{\sqrt{-\dfrac{\partial^2\phi(x_S)}{\partial\nu^2}}\sqrt{H(x_D)}}\exp\left\{\frac{\phi(x_S)-\phi(x_D)}{\varepsilon}\right\},\tag{5.4}$$

where x_D is the deepest point of the potential in the outer domain, $\phi(x_W)<\phi(x_D)<\phi(x_S)$. The mean transition time $\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]$ from Ω_D to the absorbing window is given by (4.11)

$$\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]\sim\frac{2\gamma\ln\delta^{-1}}{\varepsilon\sqrt{H(x_D)}}\exp\left\{\frac{\phi_0-\phi(x_D)}{\varepsilon}\right\}.\tag{5.5}$$

If we assume that the effect of the small window on the mean escape time, $\ln\delta^{-1}$ (or $1/\delta$ in three dimensions), is larger than that of the energy barrier, $\exp\{[\phi_0-\phi(x_S)]/\varepsilon\}$, then, according to our assumption that the potential is relatively flat outside the deep well, $\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]\ll\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]$, and so (5.3) implies

$$\Pr\{\tau_{D\to A}<\tau_{D\to W}\}\sim\frac{\mathbb{E}[\tau_{D\to W}\mid\tau_{D\to W}<\tau_{D\to A}]}{\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]}.\tag{5.6}$$

The mean absorption times $\mathbb{E}\tau_{i\to A}$ are finite for $i=D,W$. They satisfy the renewal equations

$$\mathbb{E}\tau_{D\to A}=\Pr\{\tau_{D\to A}<\tau_{D\to W}\}\mathbb{E}\left[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}\right]$$
$$\qquad+\Pr\{\tau_{D\to W}<\tau_{D\to A}\}\mathbb{E}\tau_{W\to A},\tag{5.7}$$
$$\mathbb{E}\tau_{W\to A}=\mathbb{E}\tau_{W\to D}+\mathbb{E}\tau_{D\to A}\tag{5.8}$$

(see [15]). Adding (5.7) and (5.8), and dividing by $\Pr\{\tau_{D\to A}<\tau_{D\to W}\}=1-\Pr\{\tau_{D\to W}<\tau_{D\to A}\}$, we obtain

$$\mathbb{E}\tau_{W\to A}=\mathbb{E}\left[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}\right]+\frac{\mathbb{E}\tau_{W\to D}}{\Pr\{\tau_{D\to A}<\tau_{D\to W}\}}.\tag{5.9}$$

Both $\mathbb{E}[\tau_{D\to A}\mid\tau_{D\to A}<\tau_{D\to W}]$ and $1/\Pr\{\tau_{D\to A}<\tau_{D\to W}\}$ have the same order of magnitude as functions of δ; however, $\mathbb{E}\tau_{W\to D}$ is exponentially large. Therefore,

$$\mathbb{E}\tau_{W\to A}\sim\frac{\mathbb{E}\tau_{W\to D}}{\Pr\{\tau_{D\to A}<\tau_{D\to W}\}}.\tag{5.10}$$

Now, by (5.8), we have

$$\mathbb{E}\tau_{D\to A} \sim \mathbb{E}\tau_{W\to D}\left(\frac{1}{\Pr\{\tau_{D\to A} < \tau_{D\to W}\}} - 1\right) \sim \frac{\mathbb{E}\tau_{W\to D}}{\Pr\{\tau_{D\to A} < \tau_{D\to W}\}}, \tag{5.11}$$

because $\Pr\{\tau_{D\to A} < \tau_{D\to W}\} \to 0$ as $\delta \to 0$. The meaning of (5.10) and (5.11) is that for each realization of the Markov chain, e.g., $DWDWDWDWDWDWDWDA$, the number of visits in state D is larger by 1, or equal to the number of visits at state W. The mean time that the particle spends at state W is exponentially larger than the mean time spent at state D. Therefore, the mean time to absorption is approximately the average number of visits at state D times the average time of a single visit in the deep well. The average number of visits in state D prior to absorption is $1/\Pr\{\tau_{D\to A} < \tau_{D\to W}\}$, as in a geometric distribution, and (5.10) follows. We conclude that

$$\mathbb{E}\tau_{D\to A} \sim \mathbb{E}\tau_{W\to A}; \tag{5.12}$$

i.e., the initial state (or location) of the particle has no (leading order) significance for the mean absorption time $\langle\tau_\delta\rangle$, which by (5.6) and (5.10) is

$$\langle\tau_\delta\rangle \sim \mathbb{E}\tau_{W\to A} \sim \frac{\mathbb{E}\tau_{W\to D}}{\Pr\{\tau_{D\to W} < \tau_{D\to A}\}}. \tag{5.13}$$

Substituting (5.2), (5.4)–(5.6) into (5.13) yields

$$\langle\tau_\delta\rangle = \frac{2\gamma \ln\dfrac{1}{\delta}}{\varepsilon\sqrt{H(x_W)}} \exp\left\{\frac{\phi_0 - \phi(x_W)}{\varepsilon}\right\}, \tag{5.14}$$

in agreement with (4.11).

6. Summary and discussion.

The narrow escape problem of a Brownian particle through a small absorbing window in an otherwise reflecting boundary was discussed in [8], [17], [18], and [19]. Here we solve the narrow escape problem for a Brownian particle in a force field. In cases where there is a deep potential well inside the domain, there are two time scales in the problem, the mean time to escape the well and the mean time to reach the small window. We give explicit asymptotic expressions for the mean escape time when the time scales are comparable and in the case where one is much longer than the other.

Matched asymptotics of two- and three-dimensional problems [22], [23], [24], [11] yield the leading term in the expansion of the principal eigenvalue in three dimensions and a full expansion in two dimensions. For the special case of the mixed Neumann problem with a small Dirichlet window in the boundary, the leading term obtained in [17], [18], [19] can be obtained by the application of the matched asymptotics expansion to this problem. In this paper we generalize the method of [17], [18], [19] to obtain the leading term for the corresponding boundary value problem for the Fokker–Planck operator, though matched asymptotics can be applied to this problem as well. The advantage of our method, as demonstrated in [17], is that it reveals the order of magnitude of the second term in three dimensions, while the matched asymptotics method does not indicate this in a simple way. In the particular case of a ball with a small Dirichlet cap, the application of the special functions method of Collins [2], [3] gave in [17] the unexpected estimate on the remainder term $O(\delta^2 \log\delta)$

to the expected leading term $O(\delta)$. Another advantage of the present method is the Helmholtz integral equation (3.11) for the flux and capacity of the small window. This equation is easier to solve numerically than the mixed Neumann–Dirichlet problem for a half space, as required in the boundary layer equation of the matched asymptotics expansion.

REFERENCES

[1] J. W. S. Baron Rayleigh, *The Theory of Sound*, Vol. 2, 2nd ed., Dover, New York, 1945.

[2] W. D. Collins, *On some dual series equations and their application to electrostatic problems for spheroidal caps*, Proc. Cambridge Philos. Soc., 57 (1961), pp. 367–384.

[3] W. D. Collins, *Note on an electrified circular disk situated inside an earthed coaxial infinite hollow cylinder*, Proc. Cambridge Philos. Soc., 57 (1961), pp. 623–627.

[4] C. W. Gardiner, *Handbook of Stochastic Methods*, 2nd ed., Springer, New York, 1985.

[5] I. V. Grigoriev, Y. A. Makhnovskii, A. M. Berezhkovskii, and V. Y. Zitserman, *Kinetics of escape through a small hole*, J. Chem. Phys., 116 (2002), pp. 9574–9577.

[6] P. Hänngi, P. Talkner, and M. Borkovec, *Reaction-rate theory: Fifty years after Kramers*, Rev. Modern Phys., 62 (1990), pp. 251–341.

[7] B. Hille, *Ionic Channels of Excitable Membranes*, 2nd ed., Sinauer, Sunderland, MA, 1992.

[8] D. Holcman and Z. Schuss, *Escape through a small opening: Receptor trafficking in a synaptic membrane*, J. Statist. Phys., 117 (2004), pp. 975–1014.

[9] D. Holcman and Z. Schuss, *Stochastic chemical reactions in microdomains*, J. Chem. Phys., 122 (2005), 114710.

[10] J. D. Jackson, *Classical Electrodynamics*, 2nd ed., Wiley, New York, 1975.

[11] T. Kolokolnikov, M. Titcombe, and M. J. Ward, *Optimizing the fundamental Neumann eigenvalue for the Laplacian in a domain with small traps*, European J. Appl. Math., 16 (2005), pp. 161–200.

[12] H. Kramers, *Brownian motion in a field of force*, Physica, 7 (1940), pp. 284–304.

[13] A. I. Lur'e, *Three-Dimensional Problems of the Theory of Elasticity*, Interscience, New York, 1964.

[14] B. J. Matkowsky and Z. Schuss, *Eigenvalues of the Fokker–Planck operator and the approach to equilibrium in potential fields*, SIAM J. Appl. Math., 40 (1981), pp. 242–254.

[15] B. J. Matkowsky, Z. Schuss, and C. Tier, *Uniform expansion of the transition rate in Kramers' problem*, J. Statist. Phys., 35 (1984), pp. 443–456.

[16] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*, Wiley Ser. Probab. Stat., Wiley, New York, 1980.

[17] A. Singer, Z. Schuss, D. Holcman, and R. S. Eisenberg, *Narrow escape, Part* I, J. Stat. Phys., 122 (2006), pp. 437–463.

[18] A. Singer, Z. Schuss, and D. Holcman, *Narrow escape, Part* II: *The circular disk*, J. Stat. Phys., 122 (2006), pp. 465–489.

[19] A. Singer, Z. Schuss, and D. Holcman, *Narrow escape, Part* III: *Non-smooth domains and Riemann surfaces*, J. Stat. Phys., 122 (2006), pp. 491–509.

[20] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed., North–Holland, Amsterdam, 1992.

[21] H. L. F. von Helmholtz, *Theorie der Luftschwingungen in Röhren mit offenen Enden*, Crelle, 57 (1860), pp. 1–72.

[22] M. J. Ward and J. B. Keller, *Strong localized perturbations of eigenvalue problems*, SIAM J. Appl. Math., 53 (1993), pp. 770–798.

[23] M. J. Ward, W. D. Henshaw, and J. B. Keller, *Summing logarithmic expansions for singularly perturbed eigenvalue problems*, SIAM J. Appl. Math., 53 (1993), pp. 799–828.

[24] M. J. Ward and E. Van De Velde, *The onset of thermal runaway in partially insulated or cooled reactors*, IMA J. Appl. Math., 48 (1992), pp. 53–85.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 109–132

A SPATIO-TEMPORAL DESIGN PROBLEM FOR A DAMPED WAVE EQUATION[*]

FAUSTINO MAESTRE[†], ARNAUD MÜNCH[‡], and PABLO PEDREGAL[†]

Abstract. We analyze in this work a spatio-temporal optimal design problem governed by a linear damped one-dimensional wave equation. The problem consists of simultaneously seeking the spatio-temporal layout of two isotropic materials and the static position of the damping set in order to minimize a functional depending quadratically on the gradient of the state. The lack of classical solutions for this kind of nonlinear problem is well known. We examine a well-posed relaxation by using the representation of a two-dimensional divergence-free vector as a rotated gradient. We transform the original optimal design problem into a nonconvex vector variational problem. By means of gradient Young measures we compute an explicit form of the *"constrained quasi convexification"* of the cost density. Moreover, this quasi convexification is recovered by first order laminates which give the optimal distribution of materials and damping set at every point. Finally, we analyze the relaxed problem, and some numerical experiments are performed. The novelty here lies in the optimization with respect to two independent subdomains, and our contribution consists of understanding their mutual interaction.

Key words. optimal design, wave equation, relaxation, Young measure

AMS subject classifications. 35L05, 49J05, 65K10

DOI. 10.1137/07067965X

1. Introduction—Problem statement.

Let us consider the following damped wave equation posed in $(0,T)\times\Omega$:

$$
(1.1)\quad \begin{cases} u_{tt}-\nabla_x([\alpha\mathcal{X}_{\omega_1}+\beta(1-\mathcal{X}_{\omega_1})]u_x)+d(x)\mathcal{X}_{\omega_2}u_t=0 & \text{in}\quad (0,T)\times\Omega,\\ u=0 & \text{on}\quad (0,T)\times\partial\Omega,\\ u(0,x)=u_0(x),\ u_t(0,x)=u_1(x) & \text{in}\quad \Omega, \end{cases}
$$

for any bounded interval Ω of $\mathbb{R}$ and any positive time T. $\mathcal{X}_{\omega_1}$ and $\mathcal{X}_{\omega_2}$ designate respectively the characteristic function of two subsets $\omega_1\subset\Omega\times(0,T)$ and $\omega_2\subset\Omega$, both of positive Lebesgue measure $|\omega_1|$ and $|\omega_2|$. We assume that $0<\alpha<\beta$ and that the damping potential $d\in L^\infty(\Omega;\mathbb{R}^+)$ is such that $d(x)\geq d>0$ for all $x\in\omega_2$. Finally, we assume that the initial data (u_0,u_1) are in $H_0^1(\Omega)\times L^2(\Omega)$ and are independent of ω_1,ω_2, and d. System (1.1) is then well posed, and there exists a unique weak solution such that $u\in C\left([0,T];H_0^1(\Omega)\right)\cap C^1\left([0,T];L^2(\Omega)\right)$ (see [16]).

As is well known, system (1.1) models the stabilization of an elastic string made of two materials α and β located on ω_1 and $((0,T)\times\Omega))\backslash\omega_1$, respectively, by an internal dissipative mechanism located on ω_2. The unknown $u(t,x)$ represents the transversal displacement of the string at the point x and at time t, while u_0 and u_1 designate the initial position and velocity, respectively.

[*] Received by the editors January 9, 2007; accepted for publication (in revised form) May 18, 2007; published electronically October 17, 2007.
http://www.siam.org/journals/siap/68-1/67965.html

[†] ETSI Industriales, Universidad de Castilla-La Mancha, 13071 Ciudad Real, Spain (Faustino.Maestre@uclm.es, Pablo.Pedregal@uclm.es). The research of the first author was supported by project MTM2004-07114 from Ministerio de Educación y Ciencia (Spain), by project PAI05-029 from JCCM (Castilla-La Mancha), and by Ph.D. grant 03/034 of JCCM. The research of the third author was supported by project MTM2004-07114 from Ministerio de Educación y Ciencia (Spain) and by project PAI05-029 from JCCM (Castilla-La Mancha).

[‡] Corresponding author. UMR CNRS 6623, Université de Franche-Comte, 16 route de Gray, 25030 Besançon cedex, France (arnaud.munch@univ-fcomte.fr).

Following similar works [9, 15], we address the very important question of determining the best space-time layout of materials α and β in $\Omega\times(0,T)$ and the best space distribution of damping material in order to minimize some cost depending on the square of the gradient of the underlying state u. Precisely, introducing the functions $a_\alpha, a_\beta \in L^\infty((0,T)\times\Omega;\mathbb{R}^\star_+)$ and

$$a(t,x,\mathcal{X}_{\omega_1}) = \mathcal{X}_{\omega_1} a_\alpha(t,x) + (1-\mathcal{X}_{\omega_1})a_\beta(t,x), \tag{1.2}$$

we consider the following nonlinear optimal shape design problem:

$$\text{(P)} \quad \inf_{\mathcal{X}_{\omega_1},\mathcal{X}_{\omega_2}} I(\mathcal{X}_{\omega_1},\mathcal{X}_{\omega_2}) = \int_0^T\int_\Omega (u_t^2 + a(t,x,\mathcal{X}_{\omega_1})|u_x|^2)dxdt \tag{1.3}$$

subject to

$$\begin{cases} u \text{ fulfills (1.1)}, \\ \mathcal{X}_{\omega_1} \in L^\infty(\Omega\times(0,T);\{0,1\}), \quad \mathcal{X}_{\omega_2} \in L^\infty(\Omega;\{0,1\}), \\ \displaystyle\int_\Omega \mathcal{X}_{\omega_1}(t,x)dx \le L_\alpha|\Omega| \quad \forall t\in(0,T), \;\; L_\alpha\in(0,1), \\ \displaystyle\int_\Omega \mathcal{X}_{\omega_2}(x)dx \le L_d|\Omega|, \quad L_d\in(0,1). \end{cases} \tag{1.4}$$

The constraint $(1.4)_3$ requires that for all $t\in(0,T)$ the volume fraction of the α-material be lower than L_α given in $(0,1)$. The constraint $(1.4)_4$ requires that the volume fraction of the damping material be lower than L_d given in $(0,1)$.

Optimal design problems in conductivity and elasticity have been extensively studied in the last decade from various perspectives (e.g., the homogenization approach [1, 24], shape derivative [6, 7], topological derivative [27], variational formulation [5, 26], simulation-oriented approaches [4, 12], etc). Under the hyperbolic laws, much less is known. A pioneer work in this direction is [18], where the author analyzes the hyperbolic G-closure for a similar optimal control problem (see also [17] for a general report on dynamic materials). On the other hand, an interesting analysis for optimal control problems under the wave equation in greater dimensions is described in [8], where the control is a time dependent coefficient. Let us also mention [3], where the authors examine time-harmonic solutions of the wave equation, prove a relaxation result for the corresponding design problem, and obtain existence of classical solutions for some particular cases. Finally, shape analysis for noncylindrical evolution problems is considered in [7] (and the references therein).

More recently, a one-dimensional (1-D) hyperbolic optimal control design problem with designs depending both on x and t has been addressed in [20]. This corresponds to the problem (P) with $\omega_2=\emptyset$ and a minimization with respect to ω_1 only. A full relaxation of the associated problem is given and numerically justified if the gap $\beta-\alpha>0$ is large enough. On the other hand, the pure damping case (corresponding to $\omega_1=\emptyset$ and a minimization with respect to ω_2 only) has been studied similarly in [13, 21, 22]. Once again, it appears that the well-posed character of the problem relies on the amplitude of the function d. In this work, we aim at mixing these two cases and minimize I with respect to ω_1 and ω_2 simultaneously. In this respect, we derive and analyze a well-posed relaxation of (P). The approach is based on an equivalent variational reformulation of the original problem as a nonconvex vector variational problem: following [2, 26], we transform our scalar problem with differential constraints into a vector variational problem with integral constraints (where the state equation is implicit in the new cost function). It is well known that the nonexistence of optimal solutions for vector variational problems is related to the lack of quasi convexity of the cost functional I (see [11]). Therefore, by using gradient Young measures as generalized solutions of variational problems, we compute an explicit relaxation of the original problem in the form of a relaxed (quasi-convexified) variational problem.

To the knowledge of the authors, this work is the first considering a bidesign problem. Our contribution consists, first, of adapting relaxation techniques in this case, and then, of studying the interaction between the two optimal designs ω_1 and ω_2.

The rest of the paper is organized as follows. In section 2, we describe in detail the equivalent variational reformulation (denoted by (VP)) as well as a general relaxation result when integrands are not continuous and may take on infinite values abruptly. Section 3 presents the computation of the constrained quasi convexification of the underlying integrand of (VP). The first part is concerned with the computation of a lower bound—the constrained polyconvexification—by using in a fundamental way the weak continuity of the determinant. The second part is concerned with the search for laminates furnishing the precise value of the lower bound in an attempt to show equality of the three convex hulls (poly-, quasi-, and rank one convex hulls). This provides the well-posed relaxation (RP) stated in Theorem 3.4. In addition, the optimal Young measure permits us to describe precisely the optimal microstructure (see Theorem 3.5). Section 4 is devoted to the analysis of the relaxed formulation. In section 5, we present some numerical experiments which justify the introduction of the relaxed formulation (RP) and present a simple penalization technique to obtain some elements of a minimizing sequence for (P) from the relaxed optimal solution of (RP).

2. Variational reformulation and relaxation.

In order to apply suitable results of calculus of variations [11, 25], we first reformulate the problem (P) into a classical vector variational one. To this end, following [2, 19, 26], we use a characterization of divergence-free vector fields. Precisely, since the subset ω_2 is time independent, the state equation of system (1.1) can be written as

$$div(u_t + d(x)\mathcal{X}_{\omega_2}u,\ -[\alpha\mathcal{X}_{\omega_1} + \beta(1-\mathcal{X}_{\omega_1})]u_x) = 0, \tag{2.1}$$

where the operator div is defined as $div = (\partial_t, \nabla_x)$. Then, under the hypothesis of simple-connectedness of Ω and from the characterization of the 2-D divergence-free vector fields (see, for instance, [14], Chapter I), there exists a potential $v \in H^1(\Omega \times (0,T))$ such that the above formula is equivalent to the pointwise constraint

$$\begin{pmatrix} u_t \\ -(\alpha\mathcal{X}_{\omega_1} + \beta(1-\mathcal{X}_{\omega_1}))u_x \end{pmatrix} - R\nabla v = -d(x)\mathcal{X}_{\omega_2}\bar{u}, \tag{2.2}$$

where

$$\bar{u} = \begin{pmatrix} u \\ 0 \end{pmatrix}, \quad \nabla v = \begin{pmatrix} v_t \\ v_x \end{pmatrix}, \quad R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{2.3}$$

R is the counterclockwise $\pi/2$-rotation in the (x,t)-plane. We then introduce the vector field $U = (u,v) \in (H^1(\Omega \times (0,T)))^2$ and the manifolds $\Lambda_{\gamma,\lambda}$ as follows:

$$\Lambda_{\gamma,\lambda} = \{A \in M^{2\times 2} : M_{-\gamma}A^{(1)} - RA^{(2)} = \lambda e_1\}, \quad \gamma = \alpha, \beta, \text{ and } \lambda \in \mathbb{R},$$

where $A^{(i)}$, $i = 1,2$, stands for the ith row of the matrix and

$$M_{-\gamma} = \begin{pmatrix} 1 & 0 \\ 0 & -\gamma \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{2.4}$$

It is clear that we can identify the design variable $(\mathcal{X}_{\omega_1}, \mathcal{X}_{\omega_2})$ with the vector field $U = (u,v)$; conversely, a pair $U = (u,v)$ which verifies (2.2) determines characteristic functions $(\mathcal{X}_{\omega_1}, \mathcal{X}_{\omega_2})$, so that we can consider the new design variables $U = (u,v)$, where $U : \mathbb{R}^2 \to \mathbb{R}^2$ and $\nabla U(t,x) \in \mathbb{R}^{2\times 2}$. Then, for any 2×2 matrix $A = (a_{ij})_{(1\le i,j\le 2)}$, we consider the following three functions:

$$W(t,x,U,A) = \begin{cases} a_{11}^2 + a_\alpha(t,x)a_{12}^2 & \text{if } A \in \Lambda_{\alpha,0} \cup \Lambda_{\alpha,-d(x)U^{(1)}}, \\ a_{11}^2 + a_\beta(t,x)a_{12}^2 & \text{if } A \in (\Lambda_{\beta,0} \cup \Lambda_{\beta,-d(x)U^{(1)}}) \\ & \qquad \setminus(\Lambda_{\alpha,0} \cup \Lambda_{\alpha,-d(x)U^{(1)}}), \\ +\infty & \text{else,} \end{cases}$$

$$V_\alpha(t,x,U,A) = \begin{cases} 1 & \text{if } A \in \Lambda_{\alpha,0} \cup \Lambda_{\alpha,-d(x)U^{(1)}}, \\ 0 & \text{if } A \in (\Lambda_{\beta,0} \cup \Lambda_{\beta,-d(x)U^{(1)}}) \setminus (\Lambda_{\alpha,0} \cup \Lambda_{\alpha,-d(x)U^{(1)}}), \\ +\infty & \text{else,} \end{cases}$$

$$V_d(t,x,U,A) = \begin{cases} 1 & \text{if } A \in (\Lambda_{\beta,-d(x)U^{(1)}} \cup \Lambda_{\alpha,-d(x)U^{(1)}}), \\ 0 & \text{if } A \in (\Lambda_{\beta,0} \cup \Lambda_{\alpha,0}) \setminus (\Lambda_{\beta,-d(x)U^{(1)}} \cup \Lambda_{\alpha,-d(x)U^{(1)}}), \\ +\infty & \text{else.} \end{cases}$$

Then, noting that

$$\{x \in \Omega, \mathcal{X}_{\omega_1}(x,t) = 1\} = \{x \in \Omega, V_\alpha(t,x,U,\nabla U) = 1\} \quad \forall t \in (0,T) \tag{2.5}$$

and

$$\{x \in \Omega, \mathcal{X}_{\omega_2}(x) = 1\} = \{x \in \Omega, V_d(t,x,U,\nabla U) = 1 \quad \forall t \in (0,T)\}, \tag{2.6}$$

the optimization problem (P) is equivalent to the following vector variational problem:

$$\text{(VP)} \quad m = \inf_U \int_0^T \int_\Omega W(t,x,U(t,x),\nabla U(t,x))dxdt \tag{2.7}$$

subject to

(2.8)
$$\begin{cases} U = (U^{(1)}, U^{(2)}) \in H^1((0,T)\times\Omega)^2, \\ U^{(1)}(0,x) = u_0(x), \ U_t^{(1)}(0,x) = u_1(x) \quad \text{in} \quad \Omega, \\ U^{(1)} = 0 \quad \text{in} \quad (0,T)\times\partial\Omega, \\ \displaystyle\int_\Omega V_\alpha(t,x,U(t,x),\nabla U(t,x))dx \le L_\alpha|\Omega| \quad \forall t \in [0,T], \\ \displaystyle\int_\Omega V_d(t,x,U(t,x),\nabla U(t,x)) \times V_d(0,x,U(0,x),\nabla U(0,x))dx \le L_d|\Omega| \quad \forall t \in [0,T]. \end{cases}$$

Therefore, this procedure transforms the scalar dynamical problem (P), with differentiable, integrable, and pointwise constraints, into a nonconvex vector variational problem (VP) with only pointwise and integral constraints.

We are now going to analyze the nonconvex vector problem (VP) by seeking its relaxation. We use Young measures (see [25]) as a main tool in the computation of the suitable density for the relaxed problem. Let us recall the following definition.

Definition 2.1. *The constrained quasi convexification of the functional W is defined as*

$$CQW(t,x,U,A,s,r) = \inf_{\nu} \left\{ \int_{M^{2\times 2}} W(t,x,U,A) d\nu(A) : \nu \in \mathcal{A} \right\}, \tag{2.9}$$

where

$$\begin{aligned} \mathcal{A} = \Big\{ &\nu : \nu \textit{ is a homogeneous } H^1\textit{-Young measure}, \\ &F = \int_{M^{2\times 2}} A d\nu(A), \int_{M^{2\times 2}} V_\alpha(t,x,U,A) d\nu(A) = s, \\ &\int_{M^{2\times 2}} V_d(t,x,U,A) d\nu(A) = r \quad \forall t \in [0,T] \Big\}. \end{aligned} \tag{2.10}$$

We then introduce the following minimization problem:

$$\text{(RP)} \quad \overline{m} = \inf_{(U,s,r)} \int_0^T \int_\Omega CQW(t,x,U(t,x),\nabla U(t,x),s(t,x),r(x)) dx dt \tag{2.11}$$

subject to

$$\begin{cases} U = (U^{(1)}, U^{(2)}) \in H^1((0,T)\times\Omega)^2, \\ U^{(1)}(0,x) = u_0(x), \ U_t^{(1)}(0,x) = u_1(x) \quad \text{in} \quad \Omega, \\ U^{(1)} = 0, \quad \text{in} \quad (0,T)\times\partial\Omega, \\ 0 \le s(t,x) \le 1, \quad \displaystyle\int_\Omega s(t,x) dx \le L_\alpha |\Omega| \quad \forall t \in [0,T], \\ 0 \le r(x) \le 1, \quad \displaystyle\int_\Omega r(x) dx \le L_d |\Omega|. \end{cases} \tag{2.12}$$

The functions s and r denote the pointwise volume fraction associated with the α-material and the damping set, respectively.

Then, the following relaxation result (initially obtained in the elliptic case in [2, 26]) can be proved: (RP) is a full relaxation of (VP) in the sense of the following theorem.

Theorem 2.2. *Assume that the initial data of system* (1.1) *have the regularity*

$$(u_0, u_1) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega). \tag{2.13}$$

Then, problem (RP) *is well posed and the following equality holds:*

$$m = \overline{m} \quad (\textit{i.e.}, \inf(\text{VP}) = \min(\text{RP})). \tag{2.14}$$

Moreover, the minimum (U,s,r) codifies (in the sense of Young measures) the optimal microstructures of the original optimal design problem.

Remark 2.3. In order to represent the limit of the cost function I associated with a minimizing sequence, say $\{\mathcal{X}_{\omega_1,j}, \mathcal{X}_{\omega_2,j}\}_j$, through its associated Young measure, we need equi-integrability for the sequence $|u_{t,j}|^2 + a(t,x,\mathcal{X}_{\omega_1,j})|\nabla u_j|^2$ (see [25]). Equation (2.13) is a sufficient condition to get this equi-integrability. We refer to [22, 23] for the details.

Therefore, Theorem 2.2 reduces the determination of a relaxed formulation to the computation of the constrained quasi convexification CQW associated with W.

3. Constrained quasi convexification.

In this section, we solve the optimization problem (2.9), leading for all (U,F,s,r) to the value of $CQW(t,x,U,F,s,r)$. The main difficulty is that we do not know explicitly the set of the admissible measures $\mathcal{A}$ defined in (2.10). We then follow the same strategy as in [26]. Consider two classes of a family of probability measures $\mathcal{A}_\star, \mathcal{A}^\star$ such that

$$\mathcal{A}_\star \subset \mathcal{A} \subset \mathcal{A}^\star.$$

We first calculate the minimum over the greater class of probability measures $\mathcal{A}^\star$, and then we check that the optimal value is attained by at least one measure over the narrower class $\mathcal{A}_\star$. This fact tells us that the optimal value so achieved is the same in $\mathcal{A}$, and hence we will have in fact computed the exact value $CQW(t,x,U,F,s,r)$.

Following [26], we choose $\mathcal{A}^\star$ as the set of polyconvex measures, which are not necessarily gradient Young measures, and therefore obtain a lower bound (the constrained polyconvexification). The main property of these measures is that they commute with the determinant. This constraint can be imposed in a more-or-less manageable way. We also choose $\mathcal{A}_*$ as the class of *laminates* which is a subclass of the gradient Young measures. By working with this class, we would get an upper bound (the constrained rank one convexification).

In what follows, in order to simplify the expression, we note $\Lambda_{\gamma,1}$ for $\Lambda_{\gamma,-d(x)U^{(1)}}$.

3.1. Lower bound: Polyconvexification.

We compute the constrained polyconvexification defined as follows.

Definition 3.1. *The constrained polyconvexification CPW of the functional W is given by the following minimization problem:*

$$CPW(U,F,s,r) = \min_\nu \left\{ \int_{M^{2\times 2}} W(U,A)d\nu(A) : \nu \in \mathcal{A}^\star \right\}, \tag{3.1}$$

where

$$\begin{aligned} \mathcal{A}^\star(F,s,r) = \Big\{ \nu :& \nu \textit{ is a homogeneous Young measure,} \\ & \nu \textit{ commutes with the determinant,} \\ & F = \int_{M^{2\times 2}} A d\nu(A), \\ & s = \int_{M^{2\times 2}} V_\alpha(U,A)d\nu(A), \quad r = \int_{M^{2\times 2}} V_d(U,A)d\nu(A) \Big\}. \end{aligned} \tag{3.2}$$

In this respect, we exploit that ν belongs to the class $\mathcal{A}^\star$. First, from the volume constraints $(3.2)_4$, the measure ν has the following decomposition:

$$\nu = s(r\nu_{\alpha,1} + (1-r)\nu_{\alpha,0}) + (1-s)(r\nu_{\beta,1} + (1-r)\nu_{\beta,0}) \tag{3.3}$$

with $\operatorname{supp}(\nu_{\gamma,\lambda}) \subset \Lambda_{\gamma,\lambda}, \gamma = \alpha, \beta,\ \lambda = 0, 1$. Therefore, if we introduce

$$F^{\gamma,\lambda} = \int_{\Lambda_{\gamma,\lambda}} A d\nu_{\gamma,\lambda}, \quad \gamma = \alpha, \beta,\ \lambda = 0, 1,$$

then the first moment constraint $(3.2)_3$ leads to the following expression:

$$F = s(rF^{\alpha,1} + (1-r)F^{\alpha,0}) + (1-s)(rF^{\beta,1} + (1-r)F^{\beta,0}). \tag{3.4}$$

Now, from the property $F^{\gamma,\lambda} \in \Lambda_{\gamma,\lambda}$, we have, for $\gamma = \alpha, \beta$,

$$\begin{cases} F^{\gamma,0}_{11} + F^{\gamma,0}_{22} = 0, \\ -F^{\gamma,0}_{21} - \gamma F^{\gamma,0}_{12} = 0, \end{cases} \quad \text{and} \quad \begin{cases} F^{\gamma,1}_{11} + F^{\gamma,1}_{22} = \lambda, \\ -F^{\gamma,1}_{21} - \gamma F^{\gamma,1}_{12} = 0. \end{cases} \tag{3.5}$$

Substituting (3.5) into the system (3.4), we obtain a noncompatible system on $F^{\gamma,\lambda}$ unless the condition

$$F_{11} + F_{22} = r\lambda \tag{3.6}$$

holds. Assuming henceforth this compatibility condition, (3.4)–(3.6) lead to

$$\begin{cases} F^{\alpha,1}_{11} = c_1, \quad F^{\alpha,0}_{11} = c_2, \quad F^{\beta,0}_{11} = c_3, \quad F^{\alpha,1}_{12} = c_4, \quad F^{\beta,1}_{12} = c_5, \\ F^{\beta,1}_{11} = \dfrac{F_{11} - rsc_1 - s(1-r)c_2 - (1-s)(1-r)c_3}{(1-s)r}, \\ F^{\alpha,0}_{12} = \dfrac{F_{21} + \beta F_{12} - (\beta-\alpha)rsc_4}{(1-r)s(\beta-\alpha)} \equiv f_4(c_4), \\ F^{\beta,0}_{12} = \dfrac{-F_{21} - \alpha F_{12} - (\beta-\alpha)r(1-s)c_5}{(1-r)(1-s)(\beta-\alpha)} \equiv f_5(c_5), \end{cases} \tag{3.7}$$

where $c_i \in \mathbb{R}$, $i = 1, \ldots, 5$, are parameters.

On the other hand, if we take a matrix $A = (a_{ij})_{(1 \le i,j \le 2)} \in \Lambda_{\gamma,\lambda}$ with $\gamma = \alpha, \beta$ and $\lambda = 0, 1$, then the equality

$$\det A = -A^{(1)} M_{-\gamma} A^{(1)} - \lambda A^{(1)} e_1$$

and the constraint on the commutation yield

$$\begin{aligned} \det F &= \int_{M^{2\times 2}} \det A d\nu(A) \\ &= -S_1 + \lambda r(sF^{\alpha,1}_{11} + (1-s)F^{\beta,1}_{11}) + \alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) \\ &\qquad + \beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0}), \end{aligned} \tag{3.8}$$

where

$$S_{\gamma,\lambda} = \int_{\Lambda_{\gamma,\lambda}} a^2_{12} d\nu_{\gamma,\lambda}(A), \quad \gamma = \alpha, \beta, \lambda = 1, 0, \quad S_1 = \int_{M^{2\times 2}} a^2_{11} d\nu(A). \tag{3.9}$$

Similarly, the cost function can be written as

$$\int_{M^{2\times 2}} W(U, A) d\nu(A) = S_1 + a_\alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) + a_\beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0}). \tag{3.10}$$

Finally, using Jensen's inequality, we obtain

$$S_{\gamma,\lambda} = \int_{\Lambda_{\gamma,\lambda}} a_{12}^2 d\nu_{\gamma,\lambda} \geq \left| \int_{\Lambda_{\gamma,\lambda}} a_{12} d\nu_{\gamma,\lambda} \right|^2 = |F_{12}^{\gamma,\lambda}|^2 \tag{3.11}$$

and

$$S_1 \geq \left| \int_{M^{2\times 2}} a_{11} d\nu(A) \right|^2 = |F_{11}|^2.$$

As a conclusion, from (3.8), (3.10), (3.11), the polyconvexification problem (3.1) is reduced to the following mathematical programming problem:

$$\text{(MPP)} \quad \min_{(S_1, S_{\gamma,\lambda}, c_i)} \quad S_1 + a_\alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) + a_\beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0})$$

subject to

$$\begin{cases} \det F = \lambda r(sF_{11}^{\alpha,1} + (1-s)F_{11}^{\beta,1}) - S_1 \\ \qquad\qquad + \alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) + \beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0}), \\ S_{\gamma,\lambda} \geq (F_{12}^{\gamma,\lambda})^2, \quad \gamma = \alpha, \beta, \quad \lambda = 0, 1; \qquad S_1 \geq (F_{11})^2. \end{cases}$$

The resolution of this problem leads to the following expression of CPW.

Proposition 3.2. *The polyconvexification* (3.1) *is explicitly given by*

$$CPW(U, F, s, r) = \begin{cases} |F_{11}|^2 + \dfrac{a_\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 \\ \quad + \dfrac{a_\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21}|^2 & \text{if } \psi(F, s, r) = 0, \\ +\infty & \text{else,} \end{cases} \tag{3.12}$$

where

$$\begin{aligned} \psi(F, s, r) = -\det F - |F_{11}|^2 + \lambda r F_{11} + \frac{\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 \\ + \frac{\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21}|^2. \end{aligned} \tag{3.13}$$

Proof. From (3.7), we obtain that

$$r(sF_{11}^{\alpha,1} + (1-s)F_{11}^{\beta,1}) = F_{11} - s(1-r)c_2 - (1-s)(1-r)c_3. \tag{3.14}$$

Consequently, the problem is

$$\min_{(S_1, S_{\gamma,\lambda}, c_i)} \quad S_1 + a_\alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) + a_\beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0})$$

subject to

$$\begin{cases} \det F = \lambda\big(F_{11} - s(1-r)c_2 - (1-s)(1-r)c_3\big) - S_1 \\ \qquad\qquad + \alpha s(rS_{\alpha,1} + (1-r)S_{\alpha,0}) + \beta(1-s)(rS_{\beta,1} + (1-r)S_{\beta,0}), \\ S_{\alpha,1} \geq c_4^2, \quad S_{\beta,1} \geq c_5^2, \quad S_{\alpha,0} \geq f_4^2(c_4), \quad S_{\beta,0} \geq f_5^2(c_5), \quad S_1 \geq (F_{11})^2. \end{cases} \tag{3.15}$$

Since a_α and a_β are positive, the minimum is obtained when the equalities hold in $(3.15)_2$ with a suitable choice of the constant c_2 and c_3 in $(3.15)_1$. Therefore, the minimum is

$$|F_{11}^2| + a_\alpha s(rc_4^2 + (1-r)f_4^2(c_4)) + a_\beta(1-s)(rc_5^2 + (1-r)f_5^2(c_5)). \tag{3.16}$$

The minimization of $(rc_4^2 + (1-r)f_4^2(c_4))$ with respect to c_4 leads to

$$c_4 = \frac{1}{s(\beta-\alpha)}(\beta F_{12} + F_{21}) = F_{12}^{\alpha,1} \tag{3.17}$$

and then

$$(rc_4^2 + (1-r)f_4^2(c_4)) = \left(\frac{1}{s(\beta-\alpha)}(\beta F_{12} + F_{21})\right)^2 = c_4^2 = S_{\alpha,1}. \tag{3.18}$$

Similarly, we obtain

$$c_5 = -\frac{1}{(1-s)(\beta-\alpha)}(\alpha F_{12} + F_{21}) = F_{12}^{\beta,1}. \tag{3.19}$$

Then, writing $\det F = F_{11}F_{22} - F_{12}F_{21} = -F_{11}^2 + \lambda r F_{11} - F_{12}F_{21}$ from (3.6), the relation $(3.15)_1$ becomes

$$\begin{aligned}\lambda r F_{11} - F_{12}F_{21} = \lambda\Bigg(F_{11} - s(1-r)c_2 - (1-s)(1-r)c_3\Bigg) + \frac{\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 \\ + \frac{\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21}|^2\end{aligned}$$

and implies the equality $\lambda(1-r)F_{11} = \lambda(1-r)(sc_2 + (1-s)c_3)$, and then $(sc_2 + (1-s)c_3) = F_{11}$. This leads to the expression of CPW. Moreover, note that since $c_2 = F_{11}^{\alpha,0}$ and $c_3 = F_{11}^{\beta,0}$, the relation $F_{11} = sF_{11}^{\alpha,0} + (1-s)F_{11}^{\beta,0}$ implies

$$F_{11}^{\alpha,0} = F_{11}^{\beta,0} = F_{11}, \tag{3.20}$$

and then, from $(3.15)_2$,

$$F_{11}^{\alpha,1} = F_{11}^{\beta,1} = F_{11}. \qquad \square \tag{3.21}$$

Remark 3.3. From (3.6), $-\det F - |F_{11}|^2 + \lambda r F_{11}$ is simply $F_{12}F_{21}$ and

$$\begin{aligned}\psi(F,s,r) &= F_{12}F_{21} + \frac{\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 + \frac{\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21}|^2 \\ &= \frac{1}{s(1-s)(\beta-\alpha)^2}\Big[F_{21} + F_{12}(\alpha s + \beta(1-s))\Big]\Big[\alpha\beta F_{12} + F_{21}(\alpha(1-s) + \beta s)\Big]\end{aligned} \tag{3.22}$$

does not depend explicitly on r.

The polyconvexification CPW gives a lower bound of the constrained quasi convexification. In the next section, we prove that this bound is in fact attained.

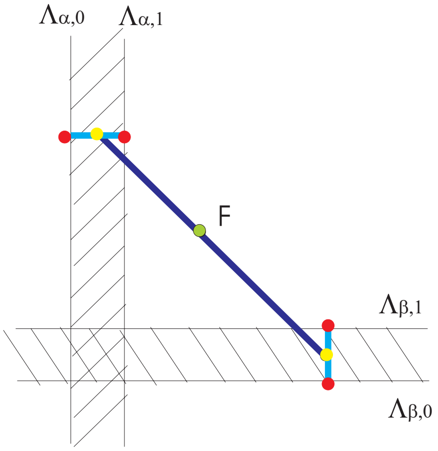


Fig. 3.1. *Geometrical decomposition of F.*

3.2. Upper bound: Searching laminates. In order to prove that the lower bound given by the polyconvexification is in fact the optimal value, we now search a measure ν in the class $\mathcal{A}_\star$ of laminates which recover it. Precisely, we exhibit a ν with the decomposition (3.3) and first moment F which satisfies a rank one condition.

First, from the optimality conditions (3.5), (3.17), (3.21) and the strict convexity of the square function, we deduce that

$$\nu^{(11)} = \delta_{F_{11}} \quad \text{and} \quad \nu^{(12)}_{\gamma,\lambda} = \delta_{F_{12}^{\gamma,\lambda}},$$

and therefore

$$\nu_{\gamma,\lambda} = \delta_{F^{\gamma,\lambda}} \qquad \text{with } \gamma = \alpha, \beta,\ \lambda = 0, 1,$$

where the matrices $F^{\gamma,\lambda}$ are

$$F^{\gamma,1} = \begin{pmatrix} F_{11} & y_\gamma \\ -\gamma y_\gamma & -F_{11} - \lambda \end{pmatrix}, \qquad F^{\gamma,0} = \begin{pmatrix} F_{11} & y_\gamma \\ -\gamma y_\gamma & -F_{11} \end{pmatrix}$$

with $\gamma = \alpha, \beta$ and

$$y_\alpha \equiv \frac{1}{s(\beta - \alpha)}(\beta F_{12} + F_{21}), \qquad y_\beta \equiv \frac{-1}{(1-s)(\beta - \alpha)}(\alpha F_{12} + F_{21}). \tag{3.23}$$

The unique possible measure ν which admits the decomposition (3.3) is then (geometrically; see Figure 3.1)

$$\nu = s(r\delta_{F^{\alpha,1}} + (1-r)\delta_{F^{\alpha,0}}) + (1-s)(r\delta_{F^{\beta,1}} + (1-r)\delta_{F^{\beta,0}}). \tag{3.24}$$

Let us now check that ν is actually a laminate; i.e., we check that there is a rank one connection between the support of deltas. On the one hand, for $\gamma = \alpha, \beta$, the relation

$$F^{\gamma,1} - F^{\gamma,0} = \begin{pmatrix} 0 & 0 \\ 0 & -\lambda \end{pmatrix} = b \otimes e_2 \ \text{ with } \ b = (0, -\lambda), e_2 = (0, 1)$$

indicates that the direction of lamination of the set of damping has to be with normal e_2. On the other hand, the relation

$$\begin{aligned}(rF^{\alpha,1}+(1-r)F^{\alpha,0})-(rF^{\beta,1}+(1-r)F^{\beta,0}) &= \begin{pmatrix} 0 & y_\alpha - y_\beta \\ \beta y_\beta - \alpha y_\alpha & 0 \end{pmatrix} \\ &= (0, y_\alpha - y_\beta)\otimes e_1 + (\alpha y_\alpha - \beta y_\beta, 0)\otimes e_2 \end{aligned} \tag{3.25}$$

implies that ν is a laminate if and only if

$$\det\begin{pmatrix} 0 & y_\alpha - y_\beta \\ \beta y_\beta - \alpha y_\alpha & 0 \end{pmatrix} = 0 \quad \Longleftrightarrow \quad (y_\alpha - y_\beta)(\beta y_\beta - \alpha y_\alpha) = 0.$$

Furthermore, from (3.22) and (3.23), we obtain that

$$\psi(F,s,r) = s(1-s)(\alpha y_\alpha - \beta y_\beta)(y_\alpha - y_\beta). \tag{3.26}$$

Consequently, the above rank one condition is equivalent to $\psi(F,s,r)=0$, which is precisely the necessary condition for the polyconvexification to be finite (see Proposition 3.2). We then conclude that ν is a first order laminate, i.e., belongs to the class $\mathcal{A}_\star$. Then, we remark that the conditions $y_\alpha - y_\beta = 0$ and $\alpha y_\alpha - \beta y_\beta = 0$ are not compatible because they imply $y_\alpha = y_\beta = 0$ and then $F_{12} = F_{21} = 0$. We conclude that the direction of lamination of the α or β material is $e_2 = (0,1)$ if $y_\alpha - y_\beta = 0$ or $e_1 = (1,0)$ if $\beta y_\beta - \alpha y_\alpha = 0$.

In conclusion, for the measure (3.24), the quasi convexification CQW defined by (2.9) coincides with CPW. Moreover, this provides an explicit expression of the full relaxation problem (2.11) stated in the following paragraph.

3.3. Well-posed full relaxation (RP). From Proposition 3.2 and by setting $\lambda = -d(x)U^{(1)}(t,x) = -d(x)u(t,x)$ and $F = \nabla U$ in (3.6), we obtain that the optimization problem

$$\text{(RP)} \quad \min_{U,s,r} \ \hat{I}\ (U) = \int_0^T \int_\Omega CQW(t,x,U(t,x),\nabla U(t,x),s(t,x),r(x))\,dxdt \tag{3.27}$$

subject to

$$\begin{cases} U=(u,v)\in (H^1([0,T]\times\Omega))^2, \quad \psi(t,x,\nabla U(t,x),s(t,x),r(x))=0, \\ u_t + v_x = d(x)r(x)u(t,x) \quad \text{in} \quad \Omega\times(0,T), \\ U^{(1)}(0,x)=u_0(x),\ U_t^{(1)}(0,x)=u_1(x) \quad \text{in} \quad \Omega, \\ U^{(1)}=0 \quad \text{in} \quad \partial\Omega\times[0,T], \\ 0\le s(t,x)\le 1, \quad \displaystyle\int_\Omega s(t,x)\,dx \le L_\alpha|\Omega| \quad \forall t\in[0,T], \\ 0\le r(x)\le 1, \quad \displaystyle\int_\Omega r(x)\,dx \le L_d|\Omega|, \end{cases}$$

where

(3.28)

$$CQW(U,F,s,r) = |F_{11}|^2 + \frac{a_\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 + \frac{a_\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21})|^2$$

and

$$\psi(F,s,r) = -\det F - |F_{11}|^2 + \lambda r F_{11} + \frac{\alpha}{s(\beta-\alpha)^2}|\beta F_{12} + F_{21}|^2 + \frac{\beta}{(1-s)(\beta-\alpha)^2}|\alpha F_{12} + F_{21}|^2 \tag{3.29}$$

for any

$$F = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}, \quad s, r \in \mathbb{R},$$

is a full well-posed relaxation of (VP) in the following sense.

THEOREM 3.4. *The variational problem* (RP) *is a relaxation of the initial optimization problem* (VP) *in the sense that*

(a) *the infima of both problems coincide;*

(b) *there are optimal solutions for the relaxed problem;*

(c) *these solutions codify (in the sense of the Young measures) the optimal microstructures of the original optimal design problem (see Theorem* 3.5*).*

Moreover, we can compute explicitly optimal microstructures, as follows.

THEOREM 3.5. *Optimal Young measures leading to the relaxed formulation are always first order laminates, which can be given in a completely explicit form:*

- *for the damping case the optimal microstructures are*

  $$r(x)\delta_1 + (1 - r(x))\delta_0 \tag{3.30}$$

  with normal direction of lamination $e_2 = (0,1)$*;*
- *for the material case, the optimal microstructures are always*

  $$s(x,t)\delta_\alpha + (1 - s(x,t))\delta_\beta \tag{3.31}$$

  with normal direction of lamination $e_2 = (0,1)$ *(if* $y_\alpha - y_\beta = 0$*) or* $e_1 = (1,0)$ *(if* $\alpha y_\alpha - \beta y_\beta = 0$*), depending on each point.*

Remark 3.6.

- The direction of lamination of the set of damping equal to $e_2 = (0,1)$ is in full agreement with the time independence of the subset ω_2, support of the dissipative term.
- It is interesting to note the influence of the damping term $\mathcal{X}_{\omega_2} d(x) u_t$ on the order of the laminates associated with the optimal Young measure. Without this damping term (i.e., when $\omega_2 = \emptyset$), the analysis of the relaxation of (P) (see [20]) reveals that the constrained quasi convexification is recovered by either first- or second order laminates, obtained when $\psi(\nabla U, s) \le 0$ and $\psi(\nabla U, s) > 0$, respectively. Here, even for arbitrarily small positive value of $||d||_{L^\infty(\Omega)}$ or $|\omega_2|$, the optimal laminates are always of first order, obtained on the set $\psi(\nabla U, s) = 0$. This clearly highlights the smoother effect of this term.

4. Interpretation of the relaxed problem (RP) in terms of u.

The quasiconvexified density depends on the gradient of U, verifies pointwise constraints, and may take the value $+\infty$ abruptly. For these reasons, the numerical approximation of the problem (RP) is not standard and is a priori tricky. In this section, taking advantage of the compatibility conditions, we analyze more deeply the relaxed formulation (RP) and eliminate the auxiliary variable $v = U^{(2)}$ introduced in section 2.

From the relation (3.26), the set $\{F; \psi(F,s)=0\}$ is decomposed into two disjoint sets, $\{F; y_\alpha - y_\beta = 0\}$ and $\{F; \alpha y_\alpha - \beta y_\beta = 0\}$. Then, noticing that

$$\begin{cases} y_\alpha - y_\beta = 0 \iff F_{21} + F_{12}(\alpha s + \beta(1-s)) = 0, \\ \alpha y_\alpha - \beta y_\beta = 0 \iff F_{21} + F_{12}\dfrac{1}{\alpha^{-1}s + \beta^{-1}(1-s)} = 0, \end{cases} \tag{4.1}$$

we may eliminate the variable F_{21} (i.e., v_t) and write the quasi-convexified in terms of F_{11} and F_{12} only, as follows:

$$CQW(U,F,s,r) = \begin{cases} |F_{11}|^2 + (a_\alpha s + a_\beta(1-s))|F_{12}^2| & \text{if } y_\alpha - y_\beta = 0, \\ |F_{11}|^2 + \frac{a_\alpha \beta^2 s + a_\beta \alpha^2(1-s)}{(\alpha(1-s)+\beta s)^2}|F_{12}^2| & \text{if } \alpha y_\alpha - \beta y_\beta = 0, \\ +\infty & \text{else.} \end{cases} \tag{4.2}$$

We can now invoke the following lemma (we refer to [12] for the proof).

Lemma 4.1. *For all $s \in (0,1)$ and $0 < \alpha < \beta$, we have*

$$\begin{aligned} &\frac{a_\beta}{a_\alpha} \le \frac{2\beta}{\alpha+\beta} \implies a_\alpha s + a_\beta(1-s) \le \frac{a_\alpha \beta^2 s + a_\beta \alpha^2(1-s)}{(\alpha(1-s)+\beta s)^2}, \\ &\frac{a_\beta}{a_\alpha} \ge \frac{\alpha+\beta}{2\alpha} \implies a_\alpha s + a_\beta(1-s) \ge \frac{a_\alpha \beta^2 s + a_\beta \alpha^2(1-s)}{(\alpha(1-s)+\beta s)^2}. \end{aligned} \tag{4.3}$$

We are thus led to introducing the following problem:

$$(\widetilde{\mathrm{RP}}) : \inf_{s,r} \tilde{I}(s,r) = \int_0^T \int_\Omega \Big(u_t(t,x)^2 + G(s) u_x(t,x)^2 \Big) dxdt \tag{4.4}$$

subject to

$$\begin{cases} u_{tt} - \nabla_x(H(s)u_x) + d(x)r(x)u_t = 0 & \text{in} \quad (0,T)\times\Omega, \\ u = 0 & \text{on} \quad (0,T)\times\partial\Omega, \\ u(0,x) = u_0(x),\ u_t(0,x) = u_1(x) & \text{in} \quad \Omega, \\ 0 \le s(t,x) \le 1, \quad \int_\Omega s(t,x)\,dx \le L_\alpha|\Omega| & \text{in} \quad [0,T]\times\Omega, \\ 0 \le r(x) \le 1, \quad \int_\Omega r(x)\,dx \le L_d|\Omega| & \text{in} \quad \Omega, \end{cases} \tag{4.5}$$

where

$$G(s) = a_\alpha s + a_\beta(1-s), \quad H(s) = \alpha s + \beta(1-s) \quad \text{if} \quad \frac{a_\beta}{a_\alpha} \le \frac{2\beta}{\alpha+\beta}, \tag{4.6}$$

and

$$G(s) = \frac{a_\alpha \beta^2 s + a_\beta \alpha^2(1-s)}{(\alpha(1-s)+\beta s)^2}, \quad H(s) = \frac{1}{\alpha^{-1}s + \beta^{-1}(1-s)} \quad \text{if} \quad \frac{a_\beta}{a_\alpha} \ge \frac{\alpha+\beta}{2\alpha}. \tag{4.7}$$

We assume henceforth that the positive functions a_α and a_β fulfill, for all $x \in \Omega$, either the property $a_\beta/a_\alpha \le 2\beta/(\alpha+\beta)$ or $a_\beta/a_\alpha \ge (\alpha+\beta)/2\alpha$.

Problem $(\widetilde{\mathrm{RP}})$ with (4.6) (resp., (4.7)) is obtained from (RP) assuming that CQW is given by $(4.2)_1$ (resp., $(4.2)_2$), then putting $F = \nabla U$ and $\lambda = -d(x)u(t,x)$, and finally by eliminating the auxiliary variable v. Note that in the first case, H is the *arithmetic* mean of (α, β), while in the second case, H is the *harmonic* mean.

Moreover, one cannot affirm, a priori, that problem $(\widetilde{\mathrm{RP}})$ is equivalent to (RP) because the pair $U = (u, v)$ which solves (RP) does not necessarily fulfill for all $(t,x) \in (0,T) \times \Omega$ the relation $v_t + u_x(\alpha s + \beta(1-s)) = 0$ (i.e., $y_\alpha - y_\beta = 0$; see (4.1) with $F = \nabla U$) or for all (t,x) the relation $v_t + u_x(\alpha^{-1}s + \beta^{-1}(1-s))^{-1} = 0$ (i.e., $\alpha y_\alpha - \beta y_\beta = 0$). However, we may conjecture this equivalence thanks to the following property.

Lemma 4.2. *The equality* $\inf(\widetilde{\mathrm{RP}}) = \min(\mathrm{RP})$ *holds.*

Proof. Let us consider the first case in Lemma 4.1, i.e., $a_\alpha/a_\beta \leq 2\beta/(\alpha+\beta)$, leading to the arithmetic situation (4.6). In this case, $(\widetilde{\mathrm{RP}})$ is simply derived from (VP) by replacing the set of characteristic functions $\mathcal{X}_{\omega_1} \in L^\infty((0,T) \times \Omega, \{0,1\})$ by the larger set of density functions $s \in L^\infty((0,T) \times \Omega, (0,1))$. Therefore $\inf(\widetilde{\mathrm{RP}}) \leq \inf(\mathrm{VP})$, and the conclusion follows from $\min(\mathrm{RP}) = \inf(\mathrm{VP})$ (see Theorem 2.2) and $\min(\mathrm{RP}) \leq \inf(\widetilde{\mathrm{RP}})$. In the harmonic situation, we obtain the result using the same arguments and Lemma 4.1. □

We have transformed the problem (RP) into the problem $(\widetilde{\mathrm{RP}})$, where the auxiliary variable v does not occur anymore and is much easier to solve numerically. We observe, however, that, since $(\widetilde{\mathrm{RP}})$ is not convex, one cannot ensure the existence of solutions. The next section aims at investigating the numerical resolution of $(\widetilde{\mathrm{RP}})$.

5. Numerical analysis of the relaxed problem.

We address in this section the numerical resolution of the problem $(\widetilde{\mathrm{RP}})$ in the quadratic case for which $(a_\alpha, a_\beta) = (1,1)$ and in the compliance case for which $(a_\alpha, a_\beta) = (\alpha, \beta)$. We first describe an algorithm of minimization and then present some numerical experiments. In order to simplify the presentation, we replace the volume constraint inequalities $(4.5)_4$ and $(4.5)_5$ by constraint equalities.

5.1. Algorithm of minimization.

We present the resolution of the relaxed problem $(\widetilde{\mathrm{RP}})$ using a gradient descent method. In this respect, we compute the first variation of the cost function with respect to s and r.

For any $\eta \in \mathbb{R}^+$, $\eta \ll 1$, and any $s_1 \in L^\infty((0,T) \times \Omega)$, we associate with the perturbation $s^\eta = s + \eta s_1$ of s the derivative of $\widetilde{I}$ with respect to s in the direction s_1 as follows:

$$\frac{\partial \widetilde{I}(s,r)}{\partial s} \cdot s_1 = \lim_{\eta \to 0} \frac{\widetilde{I}(s + \eta s_1, r) - \widetilde{I}(s,r)}{\eta}.$$

Theorem 5.1. *If* $(u_0, u_1) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega)$, *then the first derivative of* $\widetilde{I}$ *with respect to* s *in any direction* s_1 *exists and takes the form*

$$\frac{\partial \widetilde{I}(s,r)}{\partial s} \cdot s_1 = \int_0^T \int_\Omega \Big(G_{,s}(s) u_x^2 + H_{,s}(s) u_x p_x \Big) s_1 \, dxdt, \tag{5.1}$$

where u *is the solution of* (4.5) *and* p *is the solution in* $C^1([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ *of the adjoint problem*

$$\begin{cases} p_{tt} - \nabla_x(H(s)p_x) - d(x)r(x)p_t = u_{tt} + \nabla_x(G(s)u_x) & \textit{in} \quad (0,T) \times \Omega, \\ p = 0 & \textit{on} \quad (0,T) \times \partial\Omega, \\ p(T,x) = 0, \quad p_t(T,x) = u_t(T,x) & \textit{in} \quad \Omega. \end{cases} \tag{5.2}$$

Similarly, the first derivative of $\widetilde{I}$ with respect to r in any direction $r_1 \in L^\infty(\Omega)$ is given by

$$\frac{\partial \widetilde{I}(s,r)}{\partial r} \cdot r_1 = \int_\Omega d(x) r_1(x) \int_0^T u_t(t,x) p(t,x) dt dx. \tag{5.3}$$

Proof. We introduce the Lagrangian

$$\mathcal{L}(s,\phi,\psi) = \int_0^T \int_\Omega (\phi_t^2 + G(s)\phi_x^2)\, dxdt + \int_0^T \int_\Omega \Big[\phi_{tt} - \nabla_x (H(s)\phi_x) + d(x) r \phi_t\Big] \psi \, dxdt$$

for any $s \in L^\infty((0,T) \times \Omega)$, $\phi \in C([0,T]; H^2(\Omega) \cap H_0^1(\Omega)) \cap C^1([0,T]; H_0^1(\Omega))$, and $\psi \in C([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ and then write formally that

$$\frac{d\mathcal{L}}{ds} \cdot s_1 = \frac{\partial}{\partial s} \mathcal{L}(s,\phi,\psi) \cdot s_1 + \left\langle \frac{\partial}{\partial \phi} \mathcal{L}(s,\phi,\psi), \frac{\partial \phi}{\partial s} \cdot s_1 \right\rangle + \left\langle \frac{\partial}{\partial \psi} \mathcal{L}(s,\phi,\psi), \frac{\partial \psi}{\partial s} \cdot s_1 \right\rangle .$$

The first term is

$$\frac{\partial}{\partial s} \mathcal{L}(s,\phi,\psi) \cdot s_1 = \int_0^T \int_\Omega \Big(G_{,s}(s)\phi_x^2 + H_{,s}(s) \phi_x \psi_x \Big) s_1 \, dxdt \tag{5.4}$$

for any s, ϕ, ψ, whereas the third term is equal to zero if $\phi = u$ is the solution of (4.5). We then determine the solution p so that, for all $\phi \in C([0,T]; H^2(\Omega) \cap H_0^1(\Omega)) \cap C^1([0,T]; H_0^1(\Omega))$, we have

$$\left\langle \frac{\partial}{\partial \phi} \mathcal{L}(s,\phi,p), \frac{\partial \phi}{\partial s} \cdot s_1 \right\rangle = 0,$$

which leads to the formulation of the adjoint problem (5.2). Next, writing that $\widetilde{I}(s) = \mathcal{L}(s,u,p)$, we obtain (5.1) from (5.4). The relation (5.3) is obtained in a similar way. □

In order to take into account the volume constraint on s and r, we introduce the Lagrange multipliers $\gamma_s \in L^\infty((0,T); \mathbb{R})$, $\gamma_r \in \mathbb{R}$ and the functional

$$\widetilde{I}_\gamma(s,r) = \widetilde{I}(s,r) + \int_0^T \gamma_s(t) \int_\Omega s(t,x) dx dt + \gamma_r \int_\Omega r(x) dx.$$

Using Theorem 5.1, we then obtain easily that the first derivatives of $\widetilde{I}_\gamma$ are

$$\frac{\partial \widetilde{I}_\gamma(s,r)}{\partial s} \cdot s_1 = \int_0^T \int_\Omega (G_{,s}(s) u_x^2 + H_{,s}(s) u_x p_x) s_1 \, dxdt + \int_0^T \gamma_s(t) \int_\Omega s_1 dx dt,$$
$$\frac{\partial \widetilde{I}_\gamma(s,r)}{\partial r} \cdot r_1 = \int_\Omega d(x) r_1(x) \int_0^T u_t p \, dxdt + \gamma_r \int_\Omega r_1(x) dx,$$

which lets us define the following descent directions, respectively:

$$s_1(t,x) = -(G_{,s}(s) u_x^2 + H_{,s}(s) u_x p_x + \gamma_s(t)) \quad \forall (t,x) \in (0,T) \times \Omega, \tag{5.5}$$

and

$$r_1(t,x) = -\Big(d(x) \int_0^T u_t(t,x) p(t,x) dt + \gamma_r \Big) \quad \forall x \in \Omega. \tag{5.6}$$

Consequently, for any function $\eta_s \in L^\infty(\Omega \times (0,T), \mathbb{R}^+)$ with $||\eta_s||_{L^\infty((0,T)\times\Omega)}$ small enough, we have $\widetilde{I}_\gamma(s+\eta_s s_1, r) \leq \widetilde{I}_\gamma(s,r)$. The multiplier function γ_s is then determined so that, for any function $\eta_s \in L^\infty((0,T)\times\Omega, \mathbb{R}^+)$, $||s+\eta_s s_1||_{L^1(\Omega)} = L_\alpha|\Omega|$ for all $t \in (0,T)$, leading to

$$\gamma_s(t) = \frac{(\int_\Omega s(t,x)dx - L_\alpha|\Omega|) - \int_\Omega \eta_s(t,x)(G_{,s}(s)u_x^2 + H_{,s}(s)u_x p_x)\,dx}{\int_\Omega \eta_s(t,x)dx}. \tag{5.7}$$

Finally, the function η_s is chosen so that $s+\eta s_1 \in [0,1]$ for all $(t,x) \in (0,T)\times\Omega$. A simple and efficient choice consists of taking $\eta_s(t,x) = \varepsilon s(t,x)(1-s(t,x))$ for all $(t,x) \in (0,T)\times\Omega$ with ε small and positive.

Similarly, the choice

$$\gamma_r = \frac{(\int_\Omega r(x)dx - L_d|\Omega|) - \int_\Omega \eta_r(x)d(x)\int_0^T u_t(t,x)p(t,x)\,dtdx}{\int_\Omega \eta_r(x)dx} \tag{5.8}$$

with $\eta_r(x) = \varepsilon r(x)(1-r(x))$ for all $x \in \Omega$ permits us to ensure the condition $||r + \eta_r r_1||_{L^1(\Omega)} = L_d|\Omega|$.

The descent algorithm to solve numerically the relaxed problem $(\widetilde{\mathrm{RP}})$ may be structured as follows.

Let $\Omega \subset \mathbb{R}$, $(u_0,u_1) \in (H^2(\Omega)\cap H_0^1(\Omega)) \times H_0^1(\Omega)$, $L_\alpha, L_d \in (0,1)$, $T>0$, $0<\alpha<\beta$, $a_\beta, a_\alpha \in L^\infty((0,T)\times\Omega; \mathbb{R}_+^\star)$, and $\varepsilon < 1$, $\varepsilon_1 \ll 1$ be given;

- Initialization of the densities $s^0 \in L^\infty((0,T\times\Omega;]0,1[)$ and $r^0 \in L^\infty(\Omega;]0,1[)$;
- For $k \geq 0$, iteration until convergence (i.e., $|\widetilde{I}_\gamma(s^{k+1},r^{k+1}) - \widetilde{I}_\gamma(s^k,r^k)| \leq \varepsilon_1|\widetilde{I}_\gamma(s^0,r^0)|$) as follows:
  - Compute the solution u_{s^k,r^k} of (4.5) and then the solution p_{s^k,r^k} of (5.2), both corresponding to $(s,r) = (s^k,r^k)$.
  - Compute the descent direction s_1^k defined by (5.5), where the multiplier γ^k is defined by (5.7). Similarly, compute the descent direction r_1^k defined by (5.6), where the multiplier γ^k is defined by (5.8).
  - Update the density s^k in $(0,T)\times\Omega$ and the density r^k in Ω:

    $$s^{k+1} = s^k + \varepsilon s^k(1-s^k)s_1^k, \quad r^{k+1} = r^k + \varepsilon r^k(1-r^k)r_1^k$$

    with $\varepsilon \in \mathbb{R}^+$ small enough to ensure the decrease of the cost function, $s^{k+1} \in L^\infty((0,T)\times\Omega, [0,1])$ and $r^{k+1} \in L^\infty(\Omega, [0,1])$.

5.2. Numerical experiments. In this section, we present some numerical simulations for $\Omega = (0,1)$ in the quadratic case—$(a_\alpha, a_\beta) = (1,1)$—and in the compliance case—$(a_\alpha, a_\beta) = (\alpha,\beta)$. Recalling the assumption $0<\alpha<\beta$, these two cases fall into the arithmetic (see (4.6)) and harmonic (4.7) cases, respectively. From a numerical viewpoint, we highlight that the numerical resolution of the descent algorithm is a priori delicate in the sense that the descent direction depends on the derivative of u and p, both solutions of a wave equation with space and time coefficients only in $L^\infty((0,T)\times\Omega; \mathbb{R}_+^\star)$. To the knowledge of the authors, there does not exist any numerical analysis for this kind of equation. We use a C^0-finite element approximation for u and p with respect to x and a finite difference centered approximation with respect to t. Moreover, we add a vanishing viscosity and dissipative term of the type $(\beta-\alpha)\epsilon^2 \mathrm{div}(H(s)u_{xtt})$ with ϵ of order h—the space discretization parameter. This term has the effect of regularizing the descent term (5.5) and leading to a convergent algorithm. Finally, this provides an implicit and unconditionally stable scheme, consistent with (4.5) and (5.2), and of order two in time and space.

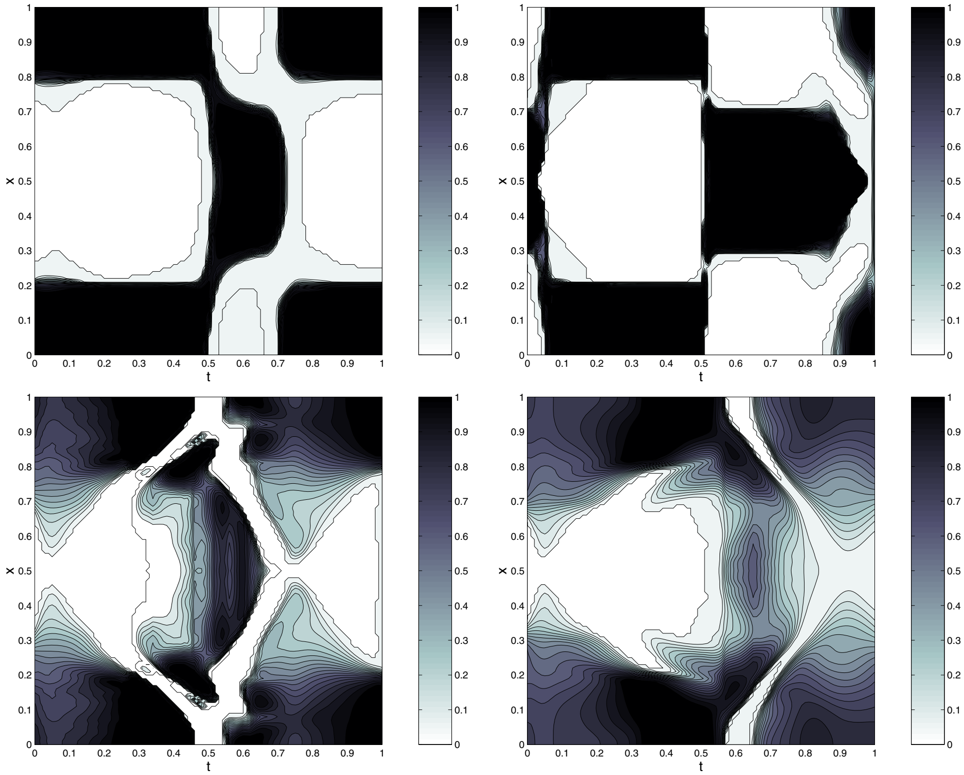


FIG. 5.1. $(a_\alpha, a_\beta) = (\alpha, \beta)$. *Optimal density* s^{lim} *on* $(0,T) \times \Omega$ *for* $(\alpha, \beta, d) = (1, 1.1, 1)$ *(top left),* $(\alpha, \beta, d) = (1, 1.1, 10)$ *(top right),* $(\alpha, \beta, d) = (1, 4, 1)$ *(bottom left), and* $(\alpha, \beta, d) = (1, 4, 10)$ *(bottom right).*

In what follows, we treat the following simple and smooth initial conditions on $\Omega = (0, 1)$:

$$u_0(x) = \sin(\pi x), \quad u_1(x) = 0, \tag{5.9}$$

and $\alpha = 1$. Results are obtained with $h = \Delta t = 10^{-2}$ (Δt designates the time discretization parameter), $\varepsilon_1 = 10^{-5}$, $L_\alpha = 2/5$, $L_d = 1/5$, $T = 1$, $s^0(t, x) = L_\alpha$ on $[0, T] \times \Omega$, $r^0(x) = L_d$ on Ω, and $\varepsilon = 10^{-2}$ (see the algorithm).

We highlight that the gradient algorithm may lead to local minima of $\widetilde{I}$ with respect to s and r. For this reason, we consider constant initial density s^0 and r^0 as indicated above, which does not privilege any location for ω_1 and ω_2.

We discuss the result obtained with respect to the value of β and of the damping function $d(x) = d\mathcal{X}_\Omega$ assumed constant in Ω: precisely, for $(\beta, d) = (1.1, 1)$, $(\beta, d) = (1.1, 10)$, $(\beta, d) = (4, 1)$, and $(\beta, d) = (4, 10)$.

5.2.1. The compliance case—$(a_\alpha, a_\beta) = (\alpha, \beta)$. The compliance choice is the most usual one, because the corresponding cost function I (see (1.3)) coincides with the energy of the vibrating membrane described by system (1.1). This case falls into the harmonic situation (4.7), $G(s) = H(s) = (\alpha^{-1}s + \beta^{-1}(1-s))^{-1}$, and we get easily that $G_{,s}(s) = (\alpha - \beta)G^2(s)/(\alpha\beta)$. We present some results obtained with the following data: Figures 5.1 and 5.2 depict the iso-values of the optimal density

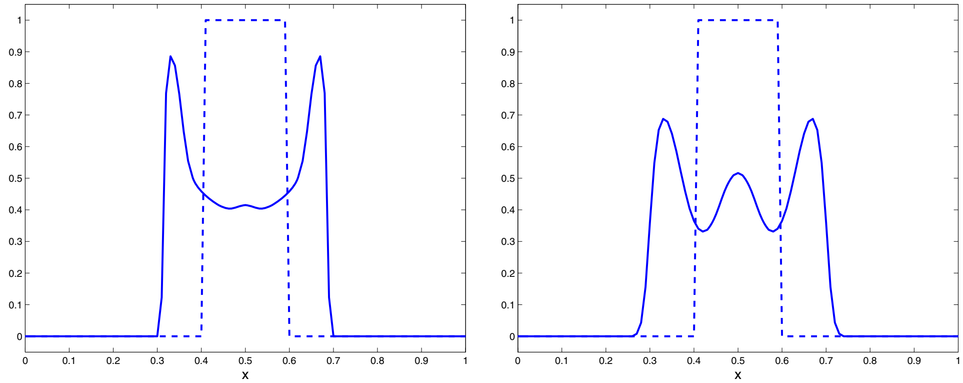


FIG. 5.2. $(a_\alpha, a_\beta) = (\alpha, \beta)$. *Solid line: Optimal density* r^{lim} *for* $(\alpha, \beta, d) = (1, 1.1, 10)$ *(left) and* $(\alpha, \beta, d) = (1, 4, 10)$ *(right). Dashed line: Optimal density* $r^{lim} = \mathcal{X}_{[0.4,0.6]}$ *for* $(\alpha, \beta, d) = (1, 1.1, 1)$ *and* $(\alpha, \beta, d) = (1, 4, 1)$.

s^{lim} and r^{lim}, respectively (obtained at the convergence of the descent algorithm). In agreement with [20] (case $\omega_2 = \emptyset$) and [22] (case $\omega_1 = \emptyset$), results depend qualitatively on the gap $\beta - \alpha$ and d. When $\beta - \alpha$ and d are small enough (function of the data of the problem), here $(\alpha, \beta, d) = (1, 1.1, 1)$, we observe that the optimal densities are characteristic functions. In this case, problem $(\widetilde{\mathrm{RP}})$ coincides with the original problem (P) (we check that when $s^{lim} \in L^\infty((0, T) \times \Omega, \{0, 1\})$, i.e., $s^{lim} = \mathcal{X}_{\omega_1}$, then $H(s^{lim}) = \alpha s^{lim} + \beta(1 - s^{\mathrm{lim}}) = \alpha \mathcal{X}_{\omega_1} + \beta(1 - \mathcal{X}_{\omega_1}))$. The original problem is therefore well posed in the class of characteristic function: $\mathcal{X}_{\omega_1} = s^{lim} \in L^\infty((0, T) \times \Omega; \{0, 1\})$ and $\mathcal{X}_{\omega_2} = r^{lim} \in L^\infty(\Omega; \{0, 1\})$.

Precisely, $r^{lim} = \mathcal{X}_{[1/2 - L_d/2, 1/2 + L_d/2]} = \mathcal{X}_{[0.4,0.6]}$, and the optimal position for the damping zone is—as expected according to the symmetry of u_0—the centered one: $\omega_2 = [0.4, 0.6]$. Moreover, the optimal distribution of (α, β)-material is time dependent (see Figure 5.1(top left)), and we observe that the weaker material α (black zone on the figure) is located, for each time t, on the point (x, t) where the amplitude of $u(x, t)$ is the lowest: on the extremities of Ω at time $t = 0$, and on the middle at time $t \approx 0.5$.

If now we consider a larger gap $\beta - \alpha$, for instance $(\alpha, \beta, d) = (1, 4, 1)$, the limit density s^{lim} is no longer a characteristic function and takes values in $(0, 1)$, highlighting microstructure (Figure 5.1(bottom left)). This suggests that the initial problem (P) is not well posed in the class of characteristic functions and does not coincide with the relaxed problem $(\widetilde{\mathrm{RP}})$. This also fully justifies the search and introduction of a relaxed well-posed formulation. We observe also that this gap is not enough larger to influence the density r^{lim}: we still have $r^{lim} = \mathcal{X}_{[0.4,0.6]}$.

Similarly, when we increase the value of the damping function d (and therefore the dissipation of the system), the limit density r^{lim} is no longer a characteristic function (see Figure 5.2 for $(\alpha, \beta, d) = (1, 1.1, 10)$ (left) and $(\alpha, \beta, d) = (1, 4, 10)$ (right)) but remains symmetric with respect to $x = 1/2$. The optimal domain is no longer the centered position but an infinite union of disjoint intervals (see section 5.2.3). This damping term with $d = 10$ changes significantly the dynamic of u and perturbs the optimal dynamical distribution of (α, β)-material (see Figure 5.1(right)). For $(\alpha, \beta, d) = (1, 1.1, 10)$, the function s^{lim} remains a characteristic function.

Finally, we plot the integrand of the cost function $\widetilde{I}$, i.e., the energy $E(t) \equiv \int_\Omega (|u_t|^2 + G(s^{lim})|u_x|^2) dx$ with respect to time (Figure 5.3). Although the system is

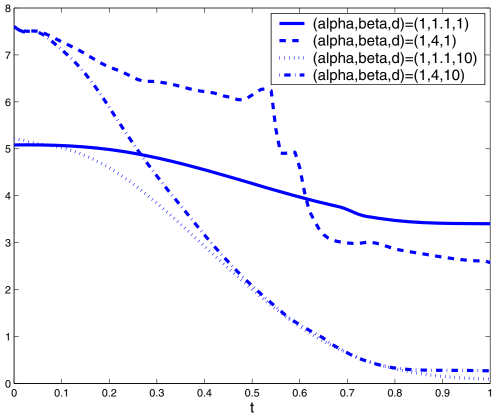


FIG. 5.3. $(a_\alpha, a_\beta) = (\alpha, \beta)$. *Evolution of* $\int_\Omega (|u_t|^2 + G(s^{lim})|u_x|^2)dx$ *versus* $t \in [0, T]$.

not necessarily dissipative when $\omega_2 = \emptyset$—we have the relation

$$\begin{aligned}\frac{dE(t)}{dt} &= 2\int_\Omega H_t(s)u_x^2 dx - 2\int_\Omega d(x)r(x)u_t^2 dx \\ &= 2\alpha\beta(\alpha-\beta)\int_\Omega \frac{s_t}{(\alpha(1-s)+\beta s)^2}u_x^2 dx - 2\int_\Omega d(x)r(x)u_t^2 dx\end{aligned} \tag{5.10}$$

—we observe that the optimal (α, β)-distribution leads to a dissipative system and that the dissipation is monotonous with respect to $(\beta - \alpha)$.

5.2.2. The quadratic case—$(a_\alpha, a_\beta) = (1, 1)$. This case falls in the arithmetic situation (see (4.6)), and the relaxed problem $(\widetilde{\mathrm{RP}})$ is then simply derived from the original one by replacing $(\mathcal{X}_{\omega_1}, \mathcal{X}_{\omega_2})$ by (s, r).

Once again, the optimal distribution of (α, β) and damping material strongly depends on the gap of the coefficients. Moreover, the numerical results still suggest that the original problem is not well posed if these gaps exceed critical values depending on the data (see Figure 5.4). The main difference with respect to the compliance case is observed for $(\alpha, \beta, d) = (1, 4, 10)$: it appears that the density r^{lim} is a characteristic function: $r^{lim} = \mathcal{X}_{[0.4,0.6]}$ (see Figure 5.5). A greater value of d (for instance, $d = 15$) is necessary to obtain values in $(0, 1)$. This phenomenon is due to the dissipative effect of the optimal (α, β)-distribution and highlights the interaction between s and r (or equivalently between ω_1 and ω_2).

Contrary to the compliance case where the density varies somewhat smoothly (see Figure 5.1), we observe in the bottom two panels in Figure 5.4 some high oscillations of the optimal density s with respect to both t and x (especially with $(\alpha, \beta, d) = (1, 4, 10)$). Due to the nonconvexity of the functional $\tilde{I}(s, r)$ with respect to s in the quadratic case, we recall that we do not know a priori whether the problem $(\widetilde{\mathrm{RP}})$ defined by (4.4) is well posed: we can only ensure that $\inf(\widetilde{\mathrm{RP}}) = \min(\mathrm{RP})$ (Lemma 4.2). The situation is different in the compliance case because $\tilde{I}$ is convex. Therefore, these oscillations may be related to the possible ill-posedness of $(\widetilde{\mathrm{RP}})$. These oscillations may also be caused, at least partially, by the numerical sensitivity of the approximation, as discussed above. Figure 5.6 depicts the evolution of the energy for the different values of α, β, and d.

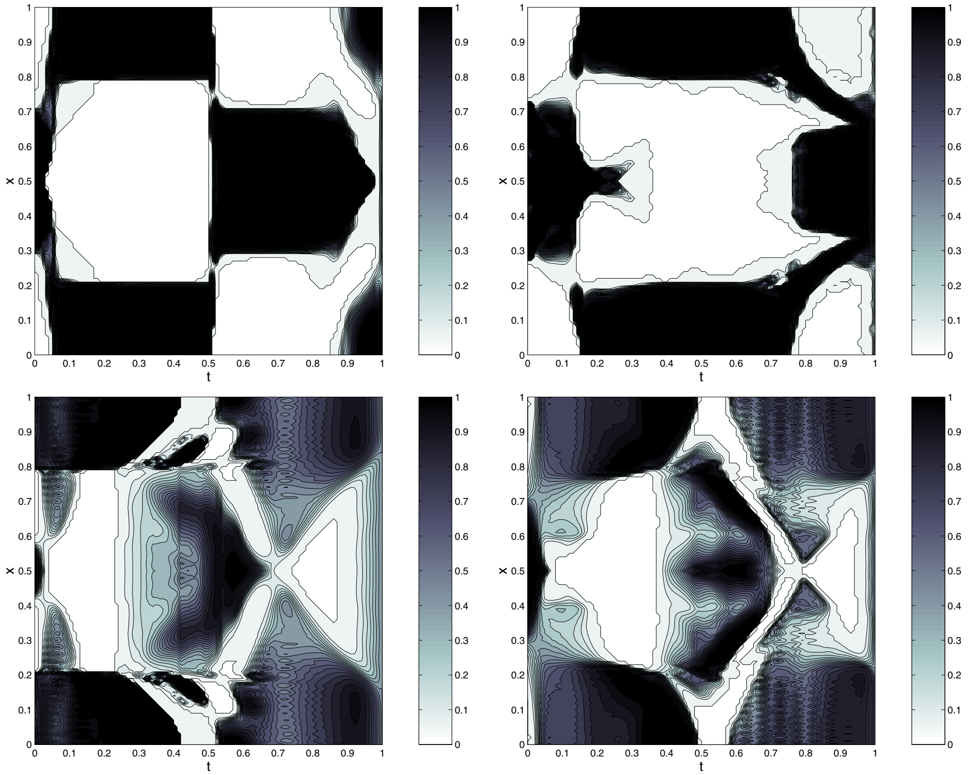


FIG. 5.4. $(a_\alpha, a_\beta) = (1, 1)$. *Optimal density* $s^{lim}(t, x)$ *on* $\Omega \times (0, T)$ *for* $(\alpha, \beta, d) = (1, 1.1, 1)$ *(top left),* $(\alpha, \beta, d) = (1, 1.1, 10)$ *(top right),* $(\alpha, \beta, d) = (1, 4, 1)$ *(bottom left), and* $(\alpha, \beta, d) = (1, 4, 10)$ *(bottom right).*

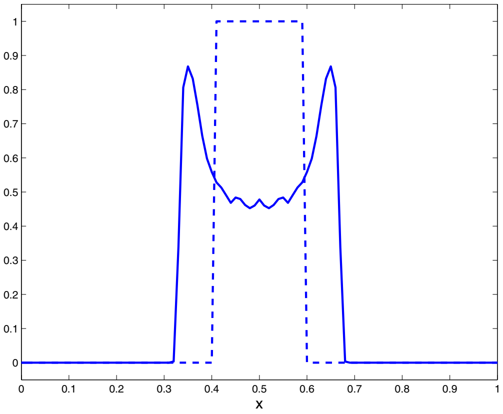


FIG. 5.5. $(a_\alpha, a_\beta) = (1, 1)$. *Solid line: Optimal density* r^{lim} *for* $(\alpha, \beta, d) = (1, 1.1, 10)$*. Dashed line: Optimal density* $r^{lim} = \mathcal{X}_{[0.4,0.6]}$ *for* $(\alpha, \beta, d) = (1, 1.1, 1)$*,* $(\alpha, \beta, d) = (1, 4, 1)$*, and* $(\alpha, \beta, d) = (1, 4, 10)$.

5.2.3. Extraction of a minimizing sequence $(\mathcal{X}_{\omega_1^k}, \mathcal{X}_{\omega_2^k})$ from the optimal density (s^{lim}, r^{lim}). Once we have the optimal microstructure of the (α, β)-material and damping material codified by the optimal density s and r, it remains

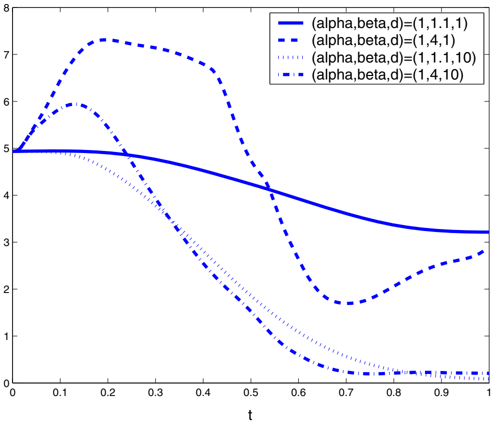


FIG. 5.6. $(a_\alpha, a_\beta) = (1, 1)$. *Evolution of* $\int_\Omega (|u_t|^2 + G(s^{lim})|u_x|^2)dx$ *versus* $t \in [0, T]$.

(from a practical viewpoint) to extract from (s^{lim}, r^{lim}), a sequence of characteristic functions $(\mathcal{X}_{\omega_1^k}, \mathcal{X}_{\omega_2^k})$ such that $\lim_{k\to\infty} \widetilde{I}(\mathcal{X}_{\omega_1^k}, \mathcal{X}_{\omega_2^k}) = \widetilde{I}(s^{lim}, r^{lim})$.

Recalling that $r^{lim}(x)$ is the volume fraction of the damping material at point x, we proceed as follows. Let us decompose the interval Ω into $M > 0$ nonempty subintervals such that $\Omega = \cup_{j=1,M}[x_j, x_{j+1}]$. Then, we associate with each interval $[x_j, x_{j+1}]$ the mean value $m_j \in [0, 1]$ defined by

$$m_j = \frac{1}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} r^{lim}(x)dx \tag{5.11}$$

and the division into two parts

$$[x_j, (1 - m_j)x_j + m_j x_{j+1}] \cup [(1 - m_j)x_j + m_j x_{j+1}, x_{j+1}]. \tag{5.12}$$

Finally, we introduce the function r_M^{pen} in $L^\infty(\Omega, \{0, 1\})$ by

$$r_M^{pen}(x) = \sum_{j=1}^{M} \mathcal{X}_{[x_j,(1-m_j)x_j+m_j x_{j+1}]}(x). \tag{5.13}$$

We easily check that $||r_M^{pen}||_{L^1(\Omega)} = ||r^{lim}||_{L^1(\Omega)}$ for all $M > 0$. The bivalued function r_M^{pen} takes more advantage of the information codified in the density r^{lim}. Similarly, using that $s(t, x)$ is the volume fraction of the α-material at point (t, x), we associate with s^{lim} a sequence of bivalued functions $s_N^{pen} \in L^\infty((0, T) \times \Omega, \{0, 1\})$ (see [20]).

For $(\alpha, \beta, d) = (1, 4, 10)$ and $(a_\alpha, a_\beta) = (\alpha, \beta)$, Figure 5.7 represents the function $r_{M=30}^{pen}$ associated with the density r^{lim} of Figure 5.2(right). Similarly, Figure 5.8 represents the function $s_{N=30}^{pen}$ associated with the optimal density s^{lim} of Figure 5.1(bottom right). Finally, we report in Table 5.1 values of $\widetilde{I}(s_N^{pen}, r_M^{pen})$ for several values of N and M. For $M = N = 40$, we obtain $\widetilde{I}(s_{40}^{pen}, r_{40}^{pen}) \approx 2.9803$ which is very near from the minimal value $I(s^{lim}, r^{lim}) \approx 2.9116$. These numerical results suggest the efficiency of this procedure to build optimal domains ω_1, ω_2 composed of a finite number of disjoints components and arbitrarily near the optimal distributions.

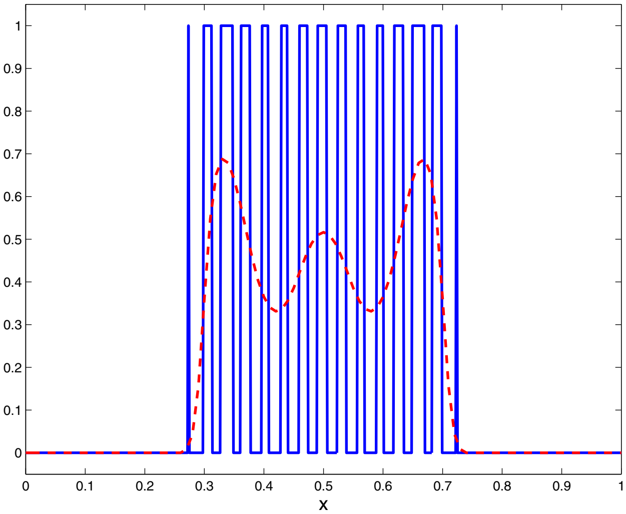


FIG. 5.7. $(a_\alpha, a_\beta) = (\alpha, \beta)$. *Characteristic function associated with the optimal density* r^{lim} *for* $(\alpha, \beta, d) = (1, 4, 10)$. $\widetilde{I}(s^{lim}, r^{lim}) \approx 2.9116$. $\widetilde{I}(s^{lim}, r^{pen}_{M=30}) \approx 3.0360$.

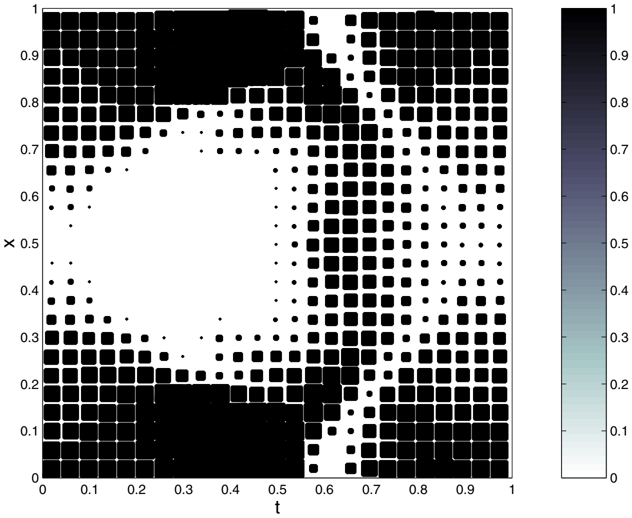


FIG. 5.8. $(a_\alpha, a_\beta) = (\alpha, \beta)$. *Characteristic function associated with the optimal density* s^{lim} *for* $(\alpha, \beta, d) = (1, 4, 10)$. $\widetilde{I}(s^{lim}, r^{lim}) \approx 2.9116$. $\widetilde{I}(s^{pen}_{N=30}, r^{lim}) \approx 3.0755$.

TABLE 5.1

$(a_\alpha, a_\beta) = (\alpha, \beta)$. $(\alpha, \beta, d) = (1, 4, 10)$—*Value of the cost function* $\widetilde{I}(s^{pen}_N, r^{pen}_M)$ *for* $M, N \in \{10, 20, 30, 40\}$.

$N\backslash M$	10	20	30	40
10	5.6181	5.2869	4.7629	4.4181
20	5.0940	4.4721	4.0761	3.6712
30	4.4910	3.8931	3.4612	3.1321
40	4.2192	3.4821	3.0712	2.9803

6. Concluding remarks and perspectives. We have analyzed the response of a 1-D damped string with respect to the spatio-temporal distribution of its longitudinal stiffness. The relaxed formulation highlights the smoothing effect of the damping term on the optimal spatio-temporal layout. Moreover, the numerical experiments indicate the strong dependence of the optimal distribution on the initial data (u_0, u_1). In order to get free of this dependence, it would be interesting to consider, for instance, an inf-sup problem of the form

$$\inf_{\mathcal{X}_{\omega_1},\mathcal{X}_{\omega_2}} \sup_{(u_0,u_1)\in H_0^1(\Omega)\times L^2(\Omega)} I(\mathcal{X}_{\omega_1}, \mathcal{X}_{\omega_2}, u_0, u_1), \tag{6.1}$$

where I designates the cost function (1.3). Another approach may consist of averaging the cost function over all initial data of unit energy (we refer to [10] in a similar context). Finally, at the numerical level, it seems important to investigate the numerical approximation of the fully relaxed problem (RP) and compare it with the simplified formulation $(\widetilde{\text{RP}})$. These aspects will be addressed in the near future.

REFERENCES

[1] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer, New York, 2002.

[2] E. Aranda and P. Pedregal, *Constrained envelope for a general class of design problems*, Discrete Contin. Dynam. Systems, Supplement (2003), pp. 30–41.

[3] J. C. Bellido and A. Donoso, *On an Optimal Design Problem in Wave-Propagation*, preprint 3-2007, Universidad Castilla-La-Mancha, Ciudad Real, Spain.

[4] M. P. Bendsøe and O. Sigmund, *Topology Optimization: Theory, Methods, and Applications,* Springer, Berlin, Heidelberg, New York, 2003.

[5] D. Bucur and G. Buttazzo, *Variational Methods in Shape Optimization Problems*, Progr. Nonlinear Differential Equations Appl. 65, Birkhäuser, Basel, 2005.

[6] M. Burger and S. J. Osher, *A survey on level set methods for inverse problems and optimal design*, European J. Appl. Math., 16 (2005), pp. 263–301.

[7] J. Cagnol and J-.P. Zolésio, *Shape control in hyperbolic problems. Optimal control of partial differential equations*, in Internat. Ser. Numer. Math. 133, Birkhäuser, Basel, 1999, pp. 77–88.

[8] A. Chambolle and F. Santosa, *Control of the wave equation by time-dependent coefficient*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 375–392.

[9] S. J. Cox, *Designing for optimal energy absorption* II*, The damped wave equation*, in Internat. Ser. Numer. Math. 126, Birkhäuser, Basel, 1998, pp. 103–109.

[10] S. J. Cox, I. Nakić, A. Rittmann, and K. Veselić, *Lyapunov optimization of a damped system*, Systems Control Lett., 53 (2004), pp. 187–194.

[11] B. Dacorogna, *Direct Method in the Calculus of Variations*, Springer, New York, 1989.

[12] A. Donoso and P. Pedregal, *Optimal design of* 2*-D conducting graded materials by minimizing quadratic functionals in the field*, Struct. Multidisc. Optim., 30 (2005), pp. 360–367.

[13] F. Fahroo and K. Ito, *Variational formulation of optimal damping designs*, Contemp. Math., 209 (1997), pp. 95–114.

[14] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes equations. Theory and Algorithms*, Springer, New York, 1986.

[15] P. Hebrard and A. Henrot, *Optimal shape and position of the actuators for the stabilization of a string*, Systems Control Lett., 48 (2003), pp. 199–209.

[16] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, Paris, 1968.

[17] K. Lurie, *Control of the coefficients of linear hyperbolic equations via spatio-temporal composites*, in Homogenization, V. Berdychevisky, V. Jikov, and G. Papanicolau, eds., World Scientific, Singapore, 1999, pp. 285–315.

[18] K. Lurie, *Some new advances in the theory of dynamic materials*, J. Elasticity, 72 (2003), pp. 229–239.

[19] F. Maestre and P. Pedregal, *Quasiconvexification in* 3*-D for a variational reformulation of an optimal design problem in conductivity*, Nonlinear Anal., 64 (2006), pp. 1962–1976.

[20] F. Maestre, A. Münch, and P. Pedregal, *Optimal design under the one-dimensional wave equation*, Interfaces Free Bound., to appear.

[21] A. Münch, P. Pedregal, and F. Periago, *A variational approach to a shape design problem for the wave equation*, C. R. Acad. Sci. Paris Sér. I, 343 (2006), pp. 371–376.

[22] A. Münch, P. Pedregal, and F. Periago, *Optimal design of the damping set for the stabilization of the wave equation*, J. Differential Equations, 231 (2006), pp. 331–358.

[23] A. Münch, P. Pedregal, and F. Periago, *Optimal Internal Design Stabilization of the Linear System of Elasticity*, preprint 2007/01, Université de Franche-Comté, Bensancon, France.

[24] F. Murat, *Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients*, Ann. Mat. Pura Appl. Ser., 4 (1977), pp. 49–68.

[25] P. Pedregal, *Parametrized measures and variational principles*, Birkhäuser, Berlin, 1997.

[26] P. Pedregal, *Vector variational problems and applications to optimal design*, ESAIM Control Optim. Calc. Var., 15 (2005), pp. 357–381.

[27] J. Sokołowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 133–153

ON POPULATION RESILIENCE TO EXTERNAL PERTURBATIONS[*]

LIONEL ROQUES[†] AND MICKAËL D. CHEKROUN[‡]

Abstract. We study a spatially explicit harvesting model in periodic or bounded environments. The model is governed by a parabolic equation with a spatially dependent nonlinearity of Kolmogorov–Petrovsky–Piskunov type, and a negative external forcing term $-\delta$. Using sub- and supersolution methods and the characterization of the first eigenvalue of some linear elliptic operators, we obtain existence and nonexistence results as well as results on the number of stationary solutions. We also characterize the asymptotic behavior of the evolution equation as a function of the forcing term amplitude. In particular, we define two critical values δ^* and δ_2 such that, if δ is smaller than δ^*, the population density converges to a "significant" state, which is everywhere above a certain small threshold, whereas if δ is larger than δ_2, the population density converges to a "remnant" state, everywhere below this small threshold. Our results are shown to be useful for studying the relationships between environmental fragmentation and maximum sustainable yield from populations. We present numerical results in the case of stochastic environments.

Key words. reaction-diffusion, heterogeneous media, harvesting models, stochastic environments, periodic environments

AMS subject classifications. 35K57, 35K55, 35J60, 35P05, 35P15, 92D25, 92D40, 60G60

DOI. 10.1137/060676994

1. Introduction.

Overexploitation has led to the extinction of many species [4]. Traditionally, models of ordinary differential equations (ODEs) or difference equations have been used to estimate the maximum sustainable yields from populations and to perform quantitative analysis of harvesting policies and management strategies [17]. Ignoring age or stage structures as well as delay mechanisms, which will not be treated by the present paper, the ODEs models are generally of the type

$$\frac{dU}{dt} = F(U) - Y(U), \tag{1.1}$$

where U is the population biomass at time t, $F(U)$ is the growth function, and $Y(U)$ corresponds to the harvest function. In these models, the most commonly used growth function is logistic, with $F(U) = U(\mu - \nu U)$ (see [5], [25], [35]), where $\mu > 0$ is the intrinsic growth rate of the population and $\nu > 0$ models its susceptibility to crowding effects.

Different harvesting strategies $Y(U)$ have been considered in the literature and are used in practical resource management. A very common one is the *constant-yield harvesting* strategy, where a constant number of individuals are removed per unit of time: $Y(U) = \delta$, with δ a positive constant. This harvesting function naturally appears when a quota is set on the harvesters [31], [32], [38]. Another frequently used harvesting strategy is the *proportional harvesting* strategy (also called *constant-effort*

[*] Received by the editors December 7, 2006; accepted for publication (in revised form) June 1, 2007; published electronically October 17, 2007.
http://www.siam.org/journals/siap/68-1/67699.html

[†] Unité Biostatistique et Processus Spatiaux (BioSP), INRA, Domaine St Paul - Site Agroparc, 84914 Avignon Cedex 9, France (lionel.roques@avignon.inra.fr, http://lionel.biosp.org).

[‡] Environmental Research and Teaching Institute, École Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France (chekro@lmd.ens.fr). The research of this author was supported by the European Commission's Project 12975 (NEST) "Extreme events: Causes and consequences (E2-C2)."

harvesting), where a constant proportion of the population is removed. It leads to a harvesting function of the type $Y(U) = \delta U$.

Much less has been done in this field using reaction-diffusion models (but see [23], [26], [29]). The aim of this paper is to perform an analysis of some harvesting models, within the framework of reaction-diffusion equations.

One of the most celebrated reaction-diffusion models was introduced by Fisher [15] and Kolmogorov, Petrovsky, and Piskunov [22] in 1937 (we call it the Fisher-KPP model). Since then, it has been widely used to model spatial propagation or spreading of biological species into homogeneous environments (see books [25], [28], and [40] for a review). The corresponding equation is

$$u_t = D\nabla^2 u + u(\mu - \nu u), \tag{1.2}$$

where $u = u(t, x)$ is the population density at time t and space position x, D is the diffusion coefficient, and μ and ν still correspond to the *constant* intrinsic growth rate and susceptibility to crowding effects. In the 1980s, this model was extended to heterogeneous environments by Shigesada, Kawasaki, and Teramoto [37]. The corresponding model (which we call the *SKT model* in this paper) is of the type

$$u_t = D\nabla^2 u + u(\mu(x) - \nu(x)u). \tag{1.3}$$

The coefficients $\mu(x)$ and $\nu(x)$ now depend on the space variable x and can therefore include some effects of environmental heterogeneity. More recently, this model revealed that the heterogeneous character of the environment plays an essential role in species persistence, in the sense that for different spatial configurations of the environment a population can survive or become extinct, depending on the habitat spatial structure [8], [12], [34], [36].

As mentioned above, the combination of a harvesting model with a Fisher-KPP population dynamics model, leading to an equation of the form $u_t = D\nabla^2 u + u(\mu - \nu u) - Y(x, u)$, has been considered in recent papers, either using a spatially dependent proportional harvesting term $Y(x, u) = q(x)u$ in [26], [29], or a spatially dependent and time-constant harvesting term $Y(x) = h(x)$ in [23]. In these papers, the models were considered in bounded domains with Dirichlet (lethal) boundary conditions.

Here we study a population dynamics model of the SKT type, with a spatially dependent harvesting term $Y(x, u)$:

$$u_t = D\nabla^2 u + u(\mu(x) - \nu(x)u) - Y(x, u). \tag{1.4}$$

We mainly focus on a "quasi-constant-yield" case, where the harvesting term depends on u only for very low population densities (ensuring the nonnegativity of u). We consider two types of domains and boundary conditions. In the first case, the domain is bounded with Neumann (reflective) boundary conditions; this framework is often more realistic for modeling species that cannot cross the domain boundary. In the second case, we consider the model (1.4) in the whole space $\mathbb{R}^N$ with periodic coefficients. This last situation, though technically more complex, is useful, for instance, for studying spreading phenomena [7], [9], and for studying the effects of environmental fragmentation, independently of the boundary effects. Lastly, note that the effects of variability in time of the harvesting function will be investigated in a forthcoming publication [13].

In section 2, we define a quasi-constant-yield harvesting reaction-diffusion model. We prove, on a firm mathematical basis, existence and nonexistence results for the equilibrium equations, as well as results on the number of possible stationary states. We also characterize the asymptotic behavior of the solutions of (1.4). In section 3, we illustrate the practical usefulness of the results of section 2, by studying the effects of the amplitude of the harvesting term on the population density in terms of environmental fragmentation. Lastly, in section 4, we give new results for the proportional harvesting case $Y(x,u) = q(x)u$.

2. Mathematical analysis of a quasi-constant-yield harvesting reaction-diffusion model.

For the sake of readability, the proofs of the results of section 2 are postponed to section 2.5.

2.1. Formulation of the model.

In this paper, we consider the model

$$u_t = D\nabla^2 u + u(\mu(x) - \nu(x)u) - \delta h(x)\rho_\varepsilon(u), \quad (t,x) \in \mathbb{R}_+ \times \Omega. \tag{2.1}$$

The function $u = u(t,x)$ denotes the population density at time t and space position x. The coefficient D, assumed to be positive, denotes the diffusion coefficient. The functions $\mu(x)$ and $\nu(x)$ respectively stand for the spatially dependent intrinsic growth rate of the population, and for its susceptibility to crowding effects. Two different types of domains Ω are considered: either $\Omega = \mathbb{R}^N$ or Ω is a smooth bounded and connected domain of $\mathbb{R}^N$ $(N \geq 1)$. We qualify the first case as the *periodic case*, and the second one as the *bounded case*. In the periodic case, we assume that the functions $\mu(x)$, $\nu(x)$, and $h(x)$ depend on the space variables in a periodic fashion. For that, let $L = (L_1, \ldots, L_N) \in (0,+\infty)^N$. We recall the following definition.

Definition 2.1. *A function g is said to be* L-periodic *if $g(x+k) = g(x)$ for all $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$ and $k \in L_1\mathbb{Z} \times \cdots \times L_N\mathbb{Z}$.*

Thus, in the periodic case, we assume that μ, ν, and h are L-periodic. In the bounded case we assume that Neumann boundary conditions hold: $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega$, where n is the outward unit normal to $\partial\Omega$. The period cell C is defined by

$$C := (0, L_1) \times \cdots \times (0, L_N)$$

in the periodic case, and in the bounded case we set

$$C := \Omega,$$

for the sake of simplicity of some forthcoming statements.

We furthermore assume that the functions μ and ν satisfy

$$\mu, \nu \in L^\infty(\Omega) \quad \text{and} \quad \exists\ \underline{\nu}, \overline{\nu} \in \mathbb{R} \text{ s.t. } 0 < \underline{\nu} < \nu(x) < \overline{\nu} \quad \forall\ x \in \Omega. \tag{2.2}$$

Regions with higher values of $\mu(x)$ and lower values of $\nu(x)$ will be qualified as being *more favorable*, while, on the other hand, regions with lower $\mu(x)$ and higher $\nu(x)$ values will be considered as being *less favorable* or, equivalently, *more hostile*.

The last term in (2.1), $\delta h(x)\rho_\varepsilon(u)$, corresponds to a quasi-constant-yield harvesting term. Indeed, the function ρ_ε satisfies

$$\rho_\varepsilon \in C^1(\mathbb{R}),\ \rho_\varepsilon' \geq 0,\ \rho_\varepsilon(s) = 0\ \forall s \leq 0 \quad \text{and} \quad \rho_\varepsilon(s) = 1\ \forall s \geq \varepsilon, \tag{2.3}$$

where ε is a nonnegative parameter. With such a harvesting function, the yield is constant in time whenever $u \geq \varepsilon$, while it depends on the population density when $u < \varepsilon$. In what follows, the parameter ε is taken to be very small. As we prove in the next sections, there are many situations where the solutions of the model always remain larger than ε. For these reasons, we qualify our model as a *quasi-constant-yield harvesting SKT model*, the "dominant" regime being the constant-yield one. Note that the function ρ_ε ensures the nonnegativity of the solutions of (2.1). From a biological point of view, ε can correspond to a threshold below which harvesting is progressively abandoned. Considering constant-yield harvesting functions without this threshold value would be unrealistic since it would lead to harvest on zero-populations.

Finally, we specify that $\delta \geq 0$ and that h is a function in $L^\infty(\Omega)$ such that

$$\exists\, \alpha > 0 \text{ with } \alpha \leq h(x) \leq 1\, \forall x \in \Omega. \tag{2.4}$$

We call h the *harvesting scalar field*, and δ designates in this way the amplitude of this field.

Before starting our analysis of this model, we consider the no-harvesting case, i.e., when $\delta = 0$. We recall the main known results in this case. These results will indeed be necessary for the analysis of the quasi-constant-yield harvesting SKT model.

2.2. The no-harvesting case. When $\delta = 0$ in (2.1), our model reduces to the SKT model described by (1.3). The behavior of the solutions of this model has been extensively studied in [8] and [9].

Results are formulated in terms of first (smallest) eigenvalue λ_1 of the Schrödinger operator $\mathcal{L}_\mu$ defined by

$$\mathcal{L}_\mu \phi := -D\nabla^2 - \mu(x)I,$$

with either periodic boundary conditions (on the period cell C) in the periodic case or Neumann boundary conditions in the bounded case. This operator is the linearized one of the full model around the trivial solution. Recall that λ_1 is defined as the unique real number such that there exists a function $\phi > 0$, the first eigenfunction, which satisfies

$$\begin{cases} -D\nabla^2\phi - \mu(x)\phi = \lambda_1\phi \quad \text{in } C, \\ \phi > 0 \quad \text{in } C, \qquad \|\phi\|_\infty = 1, \end{cases} \tag{2.5}$$

with either periodic or Neumann boundary conditions, depending on Ω. The function ϕ is uniquely defined by (2.5) [7] and belongs to $W^{2,\tau}(C)$ for all $1 \leq \tau < \infty$ (see [1] and [2] for further details). We set

$$\underline{\phi} := \min_{x \in C} \phi(x).$$

We recall that a stationary state p of (1.3) satisfies the equation

$$-D\nabla^2 p = p(\mu(x) - \nu(x)p). \tag{2.6}$$

The following result on the stationary states of (2.6) is proved in [8].

Theorem 2.2. (i) *If* $\lambda_1 < 0$, *then* (2.6) *admits a unique nonnegative*, *nontrivial*, *and bounded solution*, p_0.

(ii) *If* $\lambda_1 \geq 0$, *the only nonnegative and bounded solution of* (2.6) *is* 0.

Moreover, in the periodic case, the solution p_0 is L-periodic. Throughout this paper, p_0 always denotes the stationary solution given by Theorem 2.2.i.

In order to emphasize that this solution can be "far" from 0 (see Definition 2.5 and the commentary following (2.10)), we give a lower bound for p_0.

Proposition 2.3. *Assume that* $\lambda_1 < 0$; *then* $p_0 \geq \frac{-\lambda_1 \underline{\phi}}{\overline{\nu}}$ *in* Ω.

The asymptotic behavior of the solutions of (1.3) is also detailed in [8]. It is proved that $\lambda_1 < 0$ is a necessary and sufficient condition for species persistence, whatever the initial population u^0 is, as follows.

THEOREM 2.4. *Let u^0 be an arbitrary bounded and continuous function in Ω such that $u^0 \geq 0$, $u^0 \not\equiv 0$. Let $u(t,x)$ be the solution of* (1.3), *with initial datum $u(0,x) = u^0(x)$.*

(i) *If $\lambda_1 < 0$, then $u(t,x) \to p_0(x)$ in $W^{2,\tau}_{loc}(\Omega)$ for all $1 \leq \tau < \infty$ as $t \to +\infty$ (uniformly in the bounded case).*

(ii) *If $\lambda_1 \geq 0$, then $u(t,x) \to 0$ uniformly in Ω as $t \to +\infty$.*

The situation (i) corresponds to persistence, while in the case (ii) the population tends to extinction. In what follows, unless otherwise specified, we therefore always assume that $\lambda_1 < 0$, so that the population survives, at least when there is no harvesting. We are now in position to start our main analysis of steady states and related asymptotic behavior of the solutions of (2.1).

2.3. Stationary states analysis. As is classically demonstrated in finite dimensional dynamical systems theory and many problems in the infinite dimensional setting (see, e.g., [39]), the asymptotic behavior of the solutions of (2.1) is governed in part by the steady states and their relative stability properties. In that respect, we study in this section the positive stationary solutions of (2.1), namely the solutions of

$$-D\nabla^2 p_\delta = p_\delta(\mu(x) - \nu(x)p_\delta) - \delta h(x)\rho_\varepsilon(p_\delta), \quad x \in \Omega, \tag{2.7}$$

in the periodic and bounded cases. When needed, we may write $(2.7, \delta)$ instead of (2.7).

Note that, provided $p_\delta \geq \varepsilon$ in Ω, p_δ is equivalently a solution of the simpler equation

$$-D\nabla^2 p_\delta = p_\delta(\mu(x) - \nu(x)p_\delta) - \delta h(x), \quad x \in \Omega. \tag{2.8}$$

This last equation has been analyzed in the case of Dirichlet boundary conditions in [29], in the particular case of constant coefficients μ and ν.

Because of the type of harvesting function considered here, we are led to introduce the following definition.

DEFINITION 2.5. *Set $\varepsilon_0 := \frac{\varepsilon}{\underline{\phi}} \geq \varepsilon$. We say that a nonnegative function σ is* remnant *whenever $\max_C \sigma < \varepsilon_0$, whereas it is* significant *if it is a bounded function satisfying $\min_C \sigma \geq \varepsilon_0$.*

Remark 1. The concepts of remnant and significant solutions, as well as the harvesting term $\delta h(x)\rho_\varepsilon(u)$, are not classical. In order to clarify these notions, we present in Figure 1 a short graphical study of the nonspatial model

$$\frac{dU}{dt} = U(\mu - \nu U) - \delta\rho_\varepsilon(U) =: k(U), \quad t \in \mathbb{R}_+, \tag{2.9}$$

with constant coefficients $\mu, \nu > 0$.

Since ε_0 is assumed to be small in our model, the remnant solutions of (2.7) correspond to very low population densities. On the other hand, significant solutions are everywhere above ε_0. In particular, a constant yield is ensured in that case. In contrast to the ODE case, stationary solutions which are neither remnant nor significant may exist, as outlined in the next theorems. However, as we will see while studying the long-time behavior of the solutions of the model (2.1), they are of less importance (see Theorem 2.11 and section 3). The threshold ε_0 is different from ε

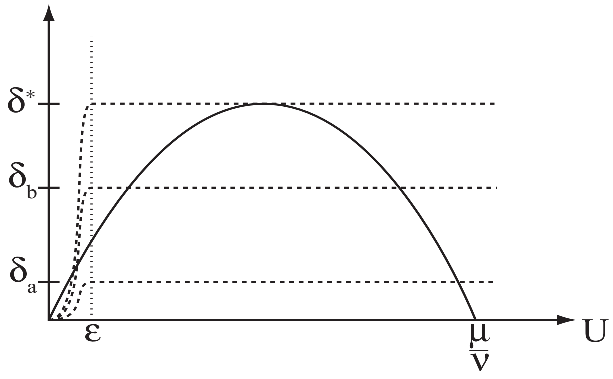


FIG. 1. *The logistic growth function $U \mapsto U(\mu - \nu U)$ (solid line), and the harvesting function $U \mapsto \delta\rho_\varepsilon(U)$ for three values of δ (dashed lines). The abscissae of the points of intersection of the solid and dashed lines correspond, respectively, to remnant (if smaller than ε) and significant (if strictly larger than ε) stationary solutions of* (2.9). *We observe that the number of significant solutions is as follows: one if $\delta < k(\varepsilon)$ (case $\delta = \delta_a$); two if $k(\varepsilon) \le \delta < \mu^2/(4\nu)$ (case $\delta = \delta_b$); one if $\delta = \mu^2/(4\nu)$ (case $\delta = \delta^*$); zero if $\delta > \mu^2/(4\nu)$. The number of nonzero remnant solutions is zero or more if $\delta \le k(\varepsilon)$ (depending on the shape of ρ_ε); one or more if $\delta > k(\varepsilon)$, since, from* (2.3), *$\rho'_\varepsilon(0) = 0$. We assumed here that $\varepsilon_0 = \varepsilon$.*

in general. We had to define remnant and significant functions using ε_0 for technical reasons (see the proof of Theorem 2.10.ii, equation (2.27)). Since ε is assumed to be very small, it has no implication on the biological interpretation of our results. Moreover, most of our results still work when ε_0 is replaced by ε.

Let us now start our analysis of (2.7). In what follows, we always assume that

$$\varepsilon_0 < \frac{-\lambda_1 \underline{\phi}}{4\overline{\nu}}, \tag{2.10}$$

so that, in particular, from Proposition 2.3, the solution p_0 of (2.6) is significant.

We begin by proving that there exists a threshold δ^* such that, if the amplitude δ is below δ^*, (2.7) admits significant solutions, while it does not in the other case.

THEOREM 2.6. *Assume that $\lambda_1 < 0$; then there exists $\delta^* \ge 0$ such that*

(i) *if $\delta \le \delta^*$, there exists at least a positive significant solution $p_\delta \le p_0$ of* (2.7);

(ii) *if $\delta > \delta^*$, there is no positive significant solution of* (2.7).

Remark 2. There is no positive bounded solution of (2.7) whenever $\lambda_1 \ge 0$.

Under stronger hypotheses, we are able to prove that (2.7) admits *at most* two significant solutions. In order to state this result, we need some definitions. Let G be the space defined by

$$G := H^1(C) \tag{2.11}$$

in the bounded case, and by

$$G := H^1_{per} = \left\{\psi \in H^1_{loc}(\mathbb{R}^N) \text{ such that } \psi \text{ is L-periodic}\right\} \tag{2.12}$$

in the periodic case. Let us define the standard Rayleigh quotient: for all $\psi \in G$, $\psi \not\equiv 0$, and for all $\sigma \in L^\infty(C)$,

$$\mathcal{R}_\sigma(\psi) := \frac{\int_C D|\nabla\psi|^2 - \sigma(x)\psi^2}{\int_C \psi^2}. \tag{2.13}$$

According to the Courant–Fischer theorem (see, e.g., [6]), the second smallest eigenvalue λ_2 of the operator $\mathcal{L}_\mu$ can be characterized by

$$\lambda_2 = \min_{E_k \subset G, \dim(E_k)=2} \max_{\psi \in E_k,\ \psi \not\equiv 0} \mathcal{R}_\mu(\psi). \tag{2.14}$$

This characterization is equivalent to the classical one given in [18].

We are now in position to state the following theorem.

Theorem 2.7. *Assume that* $\lambda_1 < 0 \leq \lambda_2$; *then, in the bounded case,* (2.7) *admits at most two significant solutions. In the periodic case,* (2.7) *admits at most two* L'*-periodic significant solutions for all* $L' \in (0, +\infty)^N$. *Moreover, under these hypotheses, if two solutions* $p_{1,\delta}$ *and* $p_{2,\delta}$ *exist, they are ordered in the sense that, for instance,* $p_{1,\delta} < p_{2,\delta}$ *in* Ω.

Remark 3. Similar methods also allow us to assess a result on the number of solutions of (2.8). Indeed, if $\lambda_1 < 0 \leq \lambda_2$, then we obtain that (2.8) admits at most two nonnegative bounded (and periodic in the periodic case) solutions. If these solutions exist, they are ordered.

In the periodic case, Theorem 2.7 also gives some information on the periodicity of the significant solutions of (2.7), which are actually found to have the same periodicity as the coefficients of (2.7), as seen in the next result.

Corollary 2.8. *Assume that* $\lambda_1 < 0 \leq \lambda_2$. *Then, in the periodic case, the significant periodic solutions of* (2.7) *are L-periodic.*

The fact that $\lambda_1 < 0$ is directly related to the instability of the trivial solution in the SKT model. The additional condition $\lambda_2 \geq 0$ in this theorem is linked to the existence of a stable manifold or center manifold of the steady state 0 of the SKT model, in some appropriate functional spaces (see [39]). Therefore, the assumptions of Theorem 2.7, and the Krein Rutmann theory, allow us to conclude that under these assumptions the unstable manifold of 0 is of dimension equal to *one* or equivalently the stable manifold is of codimension 1. Such results on multiplicity of solutions of elliptic nonlinear equations with a source or sink term have been investigated in the past and are known nowadays as being of Ambrosetti-problem type. These results also involve manifolds of codimension 1 (in the functional space of forcing) and first and second eigenvalues (for the Laplace operator only) (see [27] for a survey of these results).

In any event, Theorem 2.7 relies on the assumption that $\lambda_2 \geq 0$. In the next proposition, we give conditions under which λ_2 may become positive.

Proposition 2.9. (i) *In the bounded case, if C is a (smooth) domain with diameter* $d := \max_{x,y \in C} \|x - y\|_{\mathbb{R}^N}$, $\lambda_2(C) \geq D(\frac{\pi}{d})^2 - \max_C \mu$.

(ii) *In the periodic case,* $\lambda_2(C) \geq D(\frac{\pi}{L_d})^2 - \max_C \mu$, *where* L_d *denotes the length of the longest diagonal of the period cell* C.

For instance, when $C = [0,1] \times [0,1]$, we have $d = L_d = \sqrt{2}$; thus, for $D = 1$ and $\max_C \mu = 4$, we get $\lambda_2 > 0.9$. However, this lower bound is far from being optimal. Indeed, in all our computations of section 3, and under the same hypothesis on C and D, we always had $\lambda_2 > 0$, while $\max_C \mu = 10$. Sharper lower bounds for λ_2 can be found in [11]; however, those bounds are also more sensitive to the geometry of the domain and thus less general. They are therefore not detailed here.

We now introduce a result which is important for more applied ecological questions. Indeed, one of the main drawbacks of Theorem 2.6 is that it gives no computable bound for δ^*. Obtaining information on the value of δ^* is precious for ecological questions such as the study of the relationships between δ^* and the environmental heterogeneities. The next theorem states some computable estimates of δ^*.

Let us define

$$\delta_1 := \frac{\lambda_1^2 \underline{\phi}}{\overline{\nu}(1+\underline{\phi})^2} \quad \text{and} \quad \delta_2 := \frac{\lambda_1^2}{4\alpha\underline{\nu}}. \tag{2.15}$$

Note that neither δ_1 nor δ_2 depend on δ and ε.

THEOREM 2.10. (i) *If $\lambda_1 < 0$ and $\delta \leq \delta_1$, then there exists a positive significant (and L-periodic in the periodic case) solution p_δ of* (2.7) *such that $p_\delta \geq -\frac{\lambda_1 \underline{\phi}}{\overline{\nu}(1+\underline{\phi})}$.*

(ii) *If $\lambda_1 < 0$ and $\delta > \delta_2$, the only possible positive bounded solutions of* (2.7) *are remnant.*

The lower bound of part (i), for p_δ, does not depend on ε. Thus, there is a clear distinction between the remnant and significant solutions. Note that, of course, $\delta_1 \leq \delta_2$.

The formulae (2.15) allow numerical evaluations. An important quantity to compute is the size of the gap $\delta_2 - \delta_1$ and its fluctuations in terms of environmental configurations. This question is addressed in section 3 through a numerical study.

2.4. Asymptotic behavior. In this section, we prove that the quantity δ^* in fact corresponds to a maximum sustainable yield, in the sense that when δ is smaller than δ^*, the population density $u(t,x)$ converges to a significant stationary state of (2.1) as $t \to \infty$, whereas when δ is larger than δ^*, the population density converges to a stationary state which is not significant. In fact, when δ is larger than the quantity δ_2 defined by (2.15) we even prove that the population converges to a remnant stationary state of (2.1).

We assume here that the harvesting starts on a stabilized population governed by the standard SKT model with $\delta = 0$. From Theorem 2.4, this means that we study the behavior of the solutions $u(t,x)$ of our model (2.1), starting with the initial datum $u(0,x) = p_0(x)$. Since we have assumed that $\lambda_1 < 0$, it follows from Theorem 2.2, Proposition 2.3, and (2.10) that p_0 is well defined and significant.

Let us describe, with the next theorem, the long-time behavior of the population density.

THEOREM 2.11. *Let $u(t,x)$ be the solution of* (2.1) *with initial datum $u(0,x) = p_0(x)$. Then u is nonincreasing in t and the following hold:*

(i) *If $\delta \leq \delta^*$, $u(t,x) \to p_\delta(x)$ uniformly in Ω as $t \to +\infty$, where p_δ is the maximal significant solution of* (2.7). *Moreover, p_δ is L-periodic in the periodic case.*

(ii) *If $\delta > \delta^*$, then the function $u(t,\cdot)$ converges uniformly in Ω to a solution of* (2.7) *which is not significant.*

(iii) *If $\delta > \delta_2$, the function $u(t,\cdot)$ converges uniformly in Ω to a remnant solution of* (2.7).

Remark 4. If, in addition, we assume that $\lambda_2 \geq 0$, then Theorem 2.7 says that, whenever $\delta \leq \delta^*$, (2.1) admits at most two significant stationary states (which are periodic stationary states in the periodic case). In that case, the stationary state p_δ selected at large times is the higher one. If we do not assume that $\lambda_2 \geq 0$, this stationary state can still be defined as "the maximal one" that can be constructed by a sub- and supersolution method (see [3]).

From the above theorem, we observe that, whenever $\delta \leq \delta^*$, the solution $u(t,x)$ of (2.1), with initial datum p_0, remains significant for all times $t \geq 0$. This ensures a constant yield in time and justifies the name of the model.

Similar results could be obtained for a wider class of initial data. Indeed, with similar methods, the convergence of $u(t,x)$ to a significant solution of (2.7) can be obtained whenever $\delta \leq \delta^*$ for all bounded and continuous initial data $u(0,x)$ which are larger than the smallest significant solution of (2.7). In particular, when $u(0,x)$ is larger than the maximal significant solution of (2.7), $u(t,x)$ converges to this maximal significant solution as $t \to +\infty$. A more detailed analysis of the basin of attraction related to the maximal significant solution will be further investigated in the forthcoming paper [13].

Theorem 2.11 shows that the practical determination of δ^* is directly linked to the size of the gap $\delta_2 - \delta_1$. As we will see in section 3, this gap (δ_1, δ_2) can be very narrow in certain situations. In those cases, the numerical computation of δ_1 and δ_2 therefore gives a sharp localization of the maximum sustainable quota $\delta^* \in [\delta_1, \delta_2]$, which can be of nonnegligible ecological interest.

2.5. Proofs of the results of section 2.

Proof of Proposition 2.3. Let ϕ be defined by (2.5), with the appropriate boundary conditions. Set $\kappa_0 := \frac{-\lambda_1}{\nu}$. Then the function $\kappa_0\phi$ satisfies

$$\begin{aligned}-D\nabla^2(\kappa_0\phi) - \mu(x)\kappa_0\phi + \nu(x)(\kappa_0\phi)^2 &= \lambda_1\kappa_0\phi + \nu(x)(\kappa_0\phi)^2 \\ &= \kappa_0\phi(\lambda_1 + \nu(x)\kappa_0\phi) \leq 0.\end{aligned}$$

Thus $\kappa_0\phi$ is a subsolution of (2.6) satisfied by p_0. Since for $M \in \mathbb{R}$ large enough M is a supersolution of (2.6), it follows from the uniqueness of the positive bounded solution p_0 of (2.6) that $p_0 \geq \kappa_0\phi \geq \frac{-\lambda_1\phi}{\nu}$. □

Before proving Theorem 2.6, we begin with the following lemma.

LEMMA 2.12. *For all $\delta > 0$, if p_δ is a nonnegative bounded solution of* (2.7), *then* $p_\delta \leq p_0$.

Proof of Lemma 2.12. Assume that there exists $x_0 \in \Omega$ such that $p_\delta(x_0) > p_0(x_0)$. The function p_δ satisfies

$$-D\nabla^2 p_\delta - p_\delta(\mu(x) - \nu(x)p_\delta) = -\delta h(x)\rho_\varepsilon(p_\delta) \leq 0,$$

and thus p_δ is a subsolution of (2.6) satisfied by p_0. Since for $M \in \mathbb{R}$ large enough M is a supersolution of (2.6), we can apply a classic iterative method to infer the existence of a solution p_0' of (2.6) (with Neumann boundary conditions in the bounded case since both p_δ and M satisfy Neumann boundary conditions) such that $p_\delta \leq p_0' \leq M$. In particular, $p_0'(x_0) > p_0(x_0)$, which is in contradiction with the uniqueness of the positive bounded solution of (2.6). □

Proof of Theorem 2.6. Let us define

$$\delta^* := \sup\{\delta \geq 0, (2.7) \text{ admits a significant solution}\}.$$

For $\delta = 0$, we know from Proposition 2.3 that p_0 is a significant solution of (2.7). Moreover, for δ large enough, the nonexistence of significant solutions of (2.7) is a direct consequence of the maximum principle (it is also a consequence of the proof of Theorem 2.10.ii). Thus δ^* is well defined and bounded.

Assume that $\delta^* > 0$, and let us prove that $(2.7, \delta^*)$ admits a significant solution. By definition of δ^*, there exists a sequence $(p_{\delta_k})_{k\in\mathbb{N}}$ of solutions of $(2.7, \delta_k)$ with $0 < \delta_k \leq \delta^*$ and $\delta_k \to \delta^*$ as $k \to +\infty$. Moreover, from Lemma 2.12, $\varepsilon_0 \leq p_{\delta_k} \leq p_0$ for all $k \geq 0$. Thus, from standard elliptic estimates and Sobolev injections, the sequence $(p_{\delta_k})_{k\in\mathbb{N}}$ converges (up to the extraction of some subsequence) in $W^{2,\tau}_{loc}$, for all $1 \leq \tau < \infty$, to a significant solution p_{δ^*} of $(2.7, \delta^*)$.

Now, let $0 \leq \delta < \delta^*$. Then

$$-D\nabla^2 p_{\delta^*} - p_{\delta^*}(\mu(x) - \nu(x)p_{\delta^*}) + \delta h(x) = (\delta - \delta^*)h(x) < 0,$$

and thus p_{δ^*} is a subsolution of $(2.7, \delta)$. Since p_0 is a supersolution of $(2.7, \delta)$, and $p_{\delta^*} \leq p_0$, a classical iterative method gives the existence of a significant solution p_δ of $(2.7, \delta)$ (with Neumann boundary conditions in the bounded case since both p_0 and p_δ satisfy Neumann boundary conditions). This concludes the proof of Theorem 2.6. □

Proof of Theorem 2.7. As a preliminary, we prove that if two solutions exist, then they cannot intersect. Let $p_{1,\delta}$ and $p_{2,\delta}$ be two significant solutions of (2.7). In the bounded case, we assume that $p_{1,\delta}$ and $p_{2,\delta}$ satisfy Neumann boundary conditions. In the periodic case, we assume that there exists $\mathrm{L}' \in (0, +\infty)^N$ such that $p_{1,\delta}$ and $p_{2,\delta}$ are L'-periodic, and then denote the period cell by C'. Let us set $q_\delta := p_{2,\delta} - p_{1,\delta}$. Then q_δ verifies

$$-D\nabla^2 q_\delta - [\mu(x) - \nu(x)(p_{1,\delta} + p_{2,\delta})]q_\delta = 0; \tag{2.16}$$

thus, setting $\rho(x) := \mu(x) - \nu(x)(p_{1,\delta} + p_{2,\delta})$, we obtain

$$-D\nabla^2 q_\delta - \rho(x)q_\delta = 0, \tag{2.17}$$

with the same boundary conditions that were satisfied by $p_{1,\delta}$ and $p_{2,\delta}$.

Let $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$ be respectively the first and second eigenvalues of the operator $\mathcal{L}_\rho := -D\nabla^2 - \rho I$. Let $\mathcal{R}_\sigma(\phi)$, be defined by (2.13). Since $\rho(x) < \mu(x) - 2\underline{\nu}\varepsilon_0$ for all $x \in \Omega$, we get

$$\mathcal{R}_\rho(\varphi) \geq \mathcal{R}_\mu(\varphi) + 2\underline{\nu}\varepsilon_0$$

for all $\varphi \in G'$, where $G' := H^1(C)$ in the bounded case and

$$G' := H^1_{per} = \left\{\varphi \ \in \ H^1_{loc}(\mathbb{R}^N) \text{ such that } \varphi \text{ is L}'\text{-periodic}\right\}$$

in the periodic case. Thus, by the classical min-max formula (2.14), it follows that

$$\widehat{\lambda_2} \geq \lambda_2 + 2\underline{\nu}\varepsilon_0 > 0. \tag{2.18}$$

Furthermore, from (2.17), 0 is an eigenvalue of the operator $\mathcal{L}_\rho$. Thus, (2.18) implies that $\widehat{\lambda_1} = 0$. As a consequence, q_δ is a principal eigenfunction of the operator $\mathcal{L}_\rho$. The principal eigenfunction characterization thus implies that q_δ has a constant sign. Finally, we get that $p_{1,\delta}$ and $p_{2,\delta}$ do not intersect each other.

Let us now prove that (2.7) admits at most two significant solutions. Arguing by contradiction, we assume that there exist three significant (L'-periodic in the periodic case, for some $\mathrm{L}' \in (0, +\infty)^N$) solutions $p_{1,\delta}$, $p_{2,\delta}$, and $p_{3,\delta}$ of (2.7). From the above result, we may assume, without loss of generality, that $p_{3,\delta} > p_{2,\delta} > p_{1,\delta} > \varepsilon_0$. Set $q_{2,1} := p_{2,\delta} - p_{1,\delta}$ and $q_{3,2} := p_{3,\delta} - p_{2,\delta}$; then these functions satisfy the equations

$$-D\nabla^2 q_{2,1} - \rho_{2,1}(x)q_{2,1} = 0 \tag{2.19}$$

and

$$-D\nabla^2 q_{3,2} - \rho_{3,2}(x)q_{3,2} = 0, \tag{2.20}$$

with $\rho_{2,1} := \mu(x) - \nu(x)(p_{1,\delta} + p_{2,\delta})$ and $\rho_{3,2} := \mu(x) - \nu(x)(p_{2,\delta} + p_{3,\delta})$. Moreover, $q_{2,1} > 0$ and $q_{3,2} > 0$. Thus 0 is the first eigenvalue of the operators $\mathcal{L}_{\rho_{2,1}} := -D\nabla^2 - \rho_{2,1}I$ and $\mathcal{L}_{\rho_{3,2}} := -D\nabla^2 - \rho_{3,2}I$ with either Neumann or L′-periodic boundary conditions.

From the strong maximum principle (see, e.g., [18]) (together with Hopf's lemma in the bounded case, and using the L′-periodicity of $q_{3,2}$ in the periodic case), we obtain the existence of $\theta > 0$ such that $q_{3,2} > \theta$. Since the operator $\mathcal{L}_{\rho_{3,2}}$ is self-adjoint, we have the following formula for its first eigenvalue $\widehat{\lambda_1}^{3,2}$:

$$\widehat{\lambda_1}^{3,2} = \min_{\varphi \in G'} \mathcal{R}_{\rho_{3,2}}(\varphi).$$

Thus

$$\begin{aligned}\widehat{\lambda_1}^{3,2} = \min_{\varphi \in G'} \left\{ \mathcal{R}_{\rho_{2,1}}(\varphi) + \frac{\int_C \nu(p_{3,\delta} - p_{1,\delta})\varphi^2}{\int_C \varphi^2} \right\} &\geq \min_{\varphi \in G'} \left\{ \mathcal{R}_{\rho_{2,1}}(\varphi) \right\} + \underline{\nu}\theta \\ &\geq \widehat{\lambda_1}^{2,1} + \underline{\nu}\theta,\end{aligned}$$

where $\widehat{\lambda_1}^{2,1}$ is the first eigenvalue of the operator $\mathcal{L}_{\rho_{2,1}}$. Since the first eigenvalues of the operators $\mathcal{L}_{\rho_{2,1}}$ and $\mathcal{L}_{\rho_{3,2}}$ are both 0, we deduce that $0 \geq 0 + \underline{\nu}\theta > 0$, hence a contradiction. □

Proof of Corollary 2.8. Let p_δ be a significant L′-periodic solution of (2.7), and let $k \in \prod_{i=1}^N L_i\mathbb{Z}$. From the L-periodicity of (2.7), $p_\delta(\cdot + k)$ is also a solution of (2.7). By periodicity of p_δ, the functions p_δ and $p_\delta(\cdot + k)$ intersect each other. Thus, from Theorem 2.7, since p_δ and $p_\delta(\cdot + k)$ are both L′-periodic, $p_\delta \equiv p_\delta(\cdot + k)$. Therefore, p_δ is an L-periodic function. □

Proof of Proposition 2.9. In the bounded case, let $\tilde{C}$ be the convex hull of the set C. It was proved in [30] that the second Neumann eigenvalue of the Laplace operator $-D\nabla^2$ on $\tilde{C}$ was larger than $D(\frac{\pi}{d})^2$. Since $C \subset \tilde{C}$, we have $H^1(C) \subset H^1(\tilde{C})$. Using formula (2.14), we thus obtain that the second eigenvalue of $\mathcal{L}_\mu$ in the bounded case satisfies $\lambda_2 \geq D(\frac{\pi}{d})^2 - \max_C \mu$. This proves part (i) of Proposition 2.9.

In the periodic case, since H^1_{per} can be seen as a subset of $H^1(C)$, it follows from (2.14) that

$$\lambda_2 \geq \min_{E_k \subset H^1(C), \dim(E_k)=2} \ \max_{\psi \in E_k,\ \psi \not\equiv 0} \mathcal{R}_\mu(\psi). \tag{2.21}$$

The period cell C is convex but not smooth enough to assert that the right-hand side of (2.21) is equal to the second eigenvalue in the bounded case. Let L_d be the longest diagonal of C. Then C is included in a ball B_{L_d} of diameter L_d. Thus, from formula (2.14), the right-hand side of (2.21) is larger than the second eigenvalue of $\mathcal{L}_\mu$ on B_{L_d}. From (i), the conclusion of (ii) follows. □

Proof of Theorem 2.10*, part* (i). Let λ_1 and ϕ be defined by (2.5), and let κ be a nonnegative real number such that $\kappa > \varepsilon_0$. Then we have

$$\begin{aligned}-D\nabla^2(\kappa\phi) - \kappa\phi(\mu(x) - \kappa\phi\nu(x)) + \delta h(x)\rho_\varepsilon(\kappa\phi) &\leq \lambda_1\kappa\phi + \kappa^2\phi^2\nu(x) + \delta \\ &\leq \kappa\phi(\lambda_1 + \kappa\phi\nu(x)) + \delta \\ &\leq \max_{\tau \in I}\{\tau(\lambda_1 + \tau\overline{\nu})\} + \delta,\end{aligned} \tag{2.22}$$

where $I = \{\kappa\phi(x),\, x \in C\}$. Setting $g(\tau) := \tau(\lambda_1 + \tau\overline{\nu})$, since $\|\phi\|_\infty = 1$, and since g is a convex function, it follows from (2.22) that

$$-D\nabla^2(\kappa\phi) - \kappa\phi(\mu(x) - \kappa\phi\nu(x)) + \delta h(x)\rho_\varepsilon(\kappa\phi) \leq \max\{g(\kappa), g(\kappa\underline{\phi})\} + \delta. \tag{2.23}$$

Let us take κ_0 be such that $g(\kappa_0) = g(\kappa_0\underline{\phi})$, namely $\kappa_0 = -\frac{\lambda_1}{\underline{\nu}(1+\underline{\phi})}$ (note that $\kappa_0\phi > \varepsilon$). We get

$$-D\nabla^2(\kappa_0\phi) - \kappa_0\phi(\mu(x) - \kappa_0\phi\nu(x)) + \delta h(x) \leq -\frac{\lambda_1^2\underline{\phi}}{\underline{\nu}(1+\underline{\phi})^2} + \delta \leq 0, \tag{2.24}$$

from the hypothesis on δ of Theorem 2.10.i. Therefore, $\kappa_0\phi$ is a subsolution of (2.7) with either L-periodic or Neumann boundary conditions. Moreover, if M is a large enough constant, M is a supersolution of (2.7) with L-periodic or Neumann boundary conditions. Thus, it follows from a classical iterative method that there exists a solution $\underline{p_\delta}$ of (2.7), with the required boundary conditions, and which satisfies $\kappa_0\phi \leq \underline{p_\delta \leq M}$ in Ω. Moreover, in the periodic case, since $\kappa_0\phi$ and M are L-periodic and since (2.7) is also L-periodic, it follows that p_δ is L-periodic. Theorem 2.10.i is proved. □

Proof of Theorem 2.10, *part* (ii). Assume that $\lambda_1 < 0$, $\delta > \delta_2$, and that there exists a positive bounded solution p_δ of (2.7) which is not remnant; i.e.,

$$\exists\ x_0 \text{ with } p_\delta(x_0) \geq \varepsilon_0. \tag{2.25}$$

Since ϕ is bounded from below away from 0 and p_δ is bounded, we can define

$$\gamma^* = \inf\{\gamma > 0,\ \gamma\phi > p_\delta \text{ in } \Omega\} > 0. \tag{2.26}$$

It follows from the definition of γ^* that $\gamma^*\phi \geq p_\delta$ in Ω, and in particular, $\gamma^*\phi(x_0) \geq p_\delta(x_0) \geq \varepsilon_0$. Since $\|\phi\|_\infty = 1$, we get $\gamma^* \geq \varepsilon_0$. Thus,

$$\gamma^*\phi \geq \varepsilon_0\underline{\phi} = \varepsilon, \tag{2.27}$$

which implies $\rho_\varepsilon(\gamma^*\phi) = 1$. Thus, $h(x)\rho_\varepsilon(\gamma^*\phi) \geq \alpha$, and we get

$$-D\nabla^2(\gamma^*\phi) - \gamma^*\phi(\mu(x) - \gamma^*\phi\nu(x)) + \delta h(x)\rho_\varepsilon(\gamma^*\phi) \geq \gamma^*\phi(\lambda_1 + \gamma^*\phi\nu(x)) + \delta\alpha$$

on Ω. Moreover, since $\gamma^*\phi > 0$ and $\nu \geq \underline{\nu}$, we have $\gamma^*\phi(\lambda_1 + \gamma^*\phi\nu(x)) \geq -\frac{\lambda_1^2}{4\underline{\nu}}$. Using the fact that $\delta > \delta_2 = \frac{\lambda_1^2}{4\alpha\underline{\nu}}$, we thus get

$$-D\nabla^2(\gamma^*\phi) - \gamma^*\phi(\mu(x) - \gamma^*\phi\nu(x)) + \delta h(x)\rho_\varepsilon(\gamma^*\phi) \geq -\frac{\lambda_1^2}{4\underline{\nu}} + \delta\alpha > 0 \tag{2.28}$$

on Ω. Therefore, $\gamma^*\phi$ is a supersolution of (2.7). Set $z := \gamma^*\phi - p_\delta$. From the definition of γ^*, we know that $z \geq 0$ and that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in Ω such that $z(x_n) \to 0$ as $n \to +\infty$.

In the bounded case, up to the extraction of some subsequence, $x_n \to \overline{x} \in \Omega$ as $n \to +\infty$. By continuity, $z(\overline{x}) = 0$. Moreover, subtracting (2.7) from (2.28), we get

$$-D\nabla^2 z + [\nu(x)(\gamma^*\phi + p_\delta) + \chi(x) - \mu(x)]z > 0 \quad \text{in } \Omega, \tag{2.29}$$

where the function χ is defined by $\chi(x) = \delta h(x)\frac{\rho_\varepsilon(\gamma^*\phi(x)) - \rho_\varepsilon(p_\delta(x))}{\gamma^*\phi(x) - p_\delta(x)}$ whenever $\gamma^*\phi(x) - p_\delta(x) \neq 0$, and $\chi(x) = \rho_\varepsilon'(p_\delta(x))$ otherwise. Since ρ_ε is C^1, χ is bounded. Thus $b(x) := \nu(x)(\gamma^*\phi + p_\delta) + \chi(x) - \mu(x)$ is a bounded function. Using the strong elliptic maximum principle, we deduce from (2.29) that $z \equiv 0$. Thus $\gamma^*\phi \equiv p_\delta$ is a positive solution of (2.7). It is in contradiction with (2.28).

In the periodic case, we must also consider the situation where the sequence $(x_n)_{n\in\mathbb{N}}$ is not bounded. Let $(\overline{x}_n) \in \overline{C}$ be such that $x_n - \overline{x}_n \in \prod_{i=1}^N L_i\mathbb{Z}$. Up to the extraction of some subsequence, we can assume that there exists $\overline{x}_\infty \in \overline{C}$ such that $\overline{x}_n \to \overline{x}_\infty$ as $n \to +\infty$. Set $\phi_n(x) = \phi(x + x_n)$ and $p_{\delta,n}(x) = p_\delta(x + x_n)$. From standard elliptic estimates and Sobolev injections, it follows that (up to the extraction of some subsequence) $p_{\delta,n}$ converge in $W^{2,\tau}_{loc}$, for all $1 \leq \tau < \infty$, to a function $p_{\delta,\infty}$ satisfying

$$-\nabla^2(Dp_{\delta,\infty}) - p_{\delta,\infty}(\mu(x + \overline{x}_\infty) - p_{\delta,\infty}\nu(x + \overline{x}_\infty)) + \delta h(x + \overline{x}_\infty)\rho_\varepsilon(p_{\delta,\infty}) = 0$$

in $\mathbb{R}^N$, while $\gamma^*\phi_n$ converges to $\gamma^*\phi_\infty := \gamma^*\phi(\cdot + \overline{x}_\infty)$, and

$$-\nabla^2(D\gamma^*\phi_\infty) - \gamma^*\phi_\infty(\mu(x + \overline{x}_\infty) - \gamma^*\phi_\infty\nu(x + \overline{x}_\infty)) + \delta h(x + \overline{x}_\infty)\rho_\varepsilon(\gamma^*\phi_\infty) > 0$$

in $\mathbb{R}^N$. Let us set $z_\infty(x) := \gamma^*\phi_\infty(x) - p_{\delta,\infty}(x)$. Then $z_\infty(x) = \lim_{n\to+\infty} z(x + x_n)$, and therefore $z_\infty \geq 0$ and $z_\infty(0) = 0$. Moreover, there exists a bounded function b_∞ such that

$$-D\nabla^2 z_\infty + b_\infty z_\infty > 0 \quad \text{in } \mathbb{R}^N. \tag{2.30}$$

It then follows from the strong maximum principle that $z_\infty \equiv 0$, and we again obtain a contradiction. Finally, we necessarily have $p_\delta \leq \varepsilon_0$, and the proof of Theorem 2.10.ii is complete. □

Proof of Theorem 2.11, *part* (i). Assume that $\delta \leq \delta^*$. Let p_δ be the unique maximal significant solution defined in the proof of Theorem 2.10.i. Then, from Lemma 2.12,

$$p_\delta(x) \leq p_0(x) = u(0, x) \quad \forall x \in \Omega, \tag{2.31}$$

which implies

$$p_\delta(x) \leq u(t, x) \quad \text{in } \mathbb{R}_+ \times \Omega, \tag{2.32}$$

since p_δ is a stationary solution of (2.1). Moreover, since p_0 is a supersolution of (2.7), u is nonincreasing in time t, and standard parabolic estimates imply that u converges in $W^{2,\tau}_{loc}(\Omega)$, for all $1 \leq \tau < \infty$, to a bounded stationary solution u_∞ of (2.1). Furthermore, from (2.32) we deduce that $p_\delta \leq u_\infty \leq p_0$. Since p_δ is the maximal positive solution of (2.7), it follows that $u_\infty \equiv p_\delta$. Moreover, in the periodic case, since p_0 and (2.1) are L-periodic, $u(t, x)$ is also L-periodic in x. Therefore the convergence is uniform in Ω. Part (i) of Theorem 2.11 is proved. □

Proof of Theorem 2.11, *parts* (ii) *and* (iii). Assume that $\delta > \delta^*$. Since 0 is a stationary solution of (2.1) and $u(0, x) = p_0 > 0$, we obtain that $u(t, x) > 0$ in $\mathbb{R}^+ \times \Omega$, and again, from standard parabolic estimates, we know that u converges in $W^{2,\tau}_{loc}(\Omega)$ (for all $1 \leq \tau < \infty$) to a bounded stationary solution $\underline{u_\infty} \geq 0$ of (2.1) as $t \to +\infty$. Moreover, in the periodic case, from the L-periodicity of the initial data and of (2.1), we know that $u(t, \cdot)$ and $\underline{u_\infty}$ are L-periodic. Therefore the convergence is uniform in Ω. It follows from Theorem 2.6.ii that $\underline{u_\infty}$ cannot be a significant solution of (2.7). Moreover, if $\delta > \delta_2$, Theorem 2.10.ii ensures that $\underline{u_\infty}$ is a remnant solution of (2.7). □

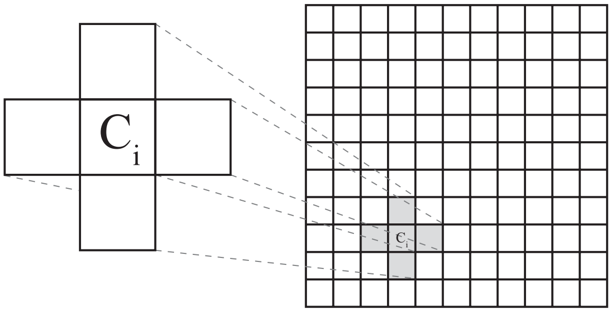


FIG. 2. *The* 4*-neighborhood system*: *an element* C_i *of* C *and its four neighbors.*

3. Numerical investigation of the effects of environmental fragmentation. We propose here to apply the results of section 2, on the estimation of the maximum sustainable yield, to the study of the effects of environmental fragmentation. A theoretical investigation of the relationships between maximum sustainable yield and fragmentation is difficult to achieve (see Remark 5). To overcome this difficulty, we propose a numerical study in the case of stochastic environments. First, we show that the gap $\delta_2 - \delta_1$, obtained from (2.15) and Theorem 2.10, remains small whatever the degree of fragmentation is. This gap corresponds to the numerical values of the harvesting quota δ for which we do not know whether the population density will converge to a significant or a remnant solution of the stationary equation (2.7). Second, we show that there is a monotone increasing relationship between the maximal sustainable yield δ^* and the habitat aggregation.

Remark 5. In a periodic environment, a simple way of changing the degree of fragmentation without changing the relative spatial pattern (favorable area/unfavorable area ratio) is to modify the size of the period cell C. Assume that $\mu(x) = \eta(\frac{x}{L})$, for some 1-periodic function η with positive integral and for some $L > 0$. This means that the environment consists of square cells of side L. Setting $\lambda_{1,L} := \lambda_1$ and $\phi_L := \phi$, we then have $-D\Delta\phi_L - \eta\left(\frac{x}{L}\right)\phi_L = \lambda_{1,L}\phi_L$ on $[0,L]^N$. The function $\psi_L(x) := \phi_L(Lx)$ thus satisfies $-D\Delta\psi_L - L^2\eta(x)\psi_L = L^2\lambda_{1,L}\psi_L$ in $[0,1]^N$, with 1-periodicity. From the Rayleigh formula we thus obtain

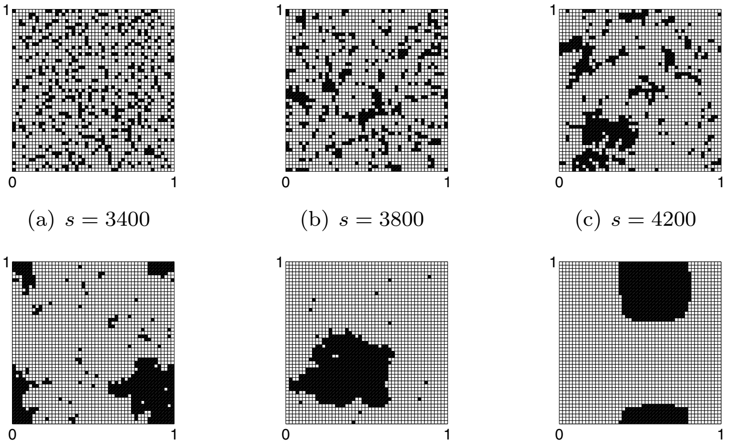
$$\lambda_{1,L} = \min_{\psi \in H^1_{per}} \frac{D}{L^2} \frac{\int_{[0,1]^N} |\nabla\psi|^2}{\int_{[0,1]^N} \psi^2} - \frac{\int_{[0,1]^N} \eta\psi^2}{\int_{[0,1]^N} \psi^2};$$

therefore $\lambda_{1,L} < 0$ (since $\psi \equiv 1 \in H^1_{per}$), and $\lambda_{1,L}$ decreases with L. It implies that δ_2 increases with L. The relationship between δ_1 and L is less clear since $\underline{\phi_L} = \min_C \phi_L$ may not always be an increasing function of L.

In order to lessen the boundary effects and to focus on fragmentation, we place ourselves in the periodic case. For our numerical computations, we assume that the environment is made of two components, favorable and unfavorable regions. This is expressed in the model (2.1) through the coefficient $\mu(x)$, which takes two values μ^+ or μ^-, depending on the space variable x. We also assume that

$$\mu^+ > \mu^-, \quad \nu(x) \equiv 1, \quad h(x) \equiv 1, \text{ and } D = 1.$$

Using a stochastic model for landscape generation [34], we built 2000 samples of binary environments, on the two-dimensional period cell $C = [0,1]^2$, with different



(a) $s = 3400$ (b) $s = 3800$ (c) $s = 4200$

(d) $s = 4600$ (e) $s = 4800$ (f) $s = 4900$

FIG. 3. *Some samples of the landscapes used for the computations of δ_1 and δ_2, with different values of the habitat aggregation index s. The black areas correspond to more favorable environment, where $\mu(x) = \mu^+$.*

degrees of fragmentation. In all these environments, the favorable region, where $\mu(x) = \mu^+$, occupies 20% of the period cell. The environmental fragmentation is defined as follows. We discretize the cell C into $n_C = 50 \times 50$ equal squares C_i. The lattice made of the cells C_i is equipped with a 4-neighborhood system $V(C_i)$ (see Figure 2), with toric conditions. On each cell C_i, we assume that the function μ takes either the value μ^+ or μ^-, while the number $n_+ = \text{card}\{i,\ \mu \equiv \mu^+ \text{ on } C_i\}$ is fixed to $n_C \times \frac{20}{100} = 500$. For each landscape sample $\omega = (\mu(C_i))_{i=1,\dots,n_C}$, we set $s(\omega) = \frac{1}{2}\sum_{C_i \subset C}\sum_{C_j \in V(C_i)} \mathbb{1}\{\mu(C_j) = \mu(C_i)\}$, the number of pairs of neighbors (C_i, C_j) such that μ takes the same value on C_i and C_j ($\mathbb{1}\{\cdot\}$ is the indicator function). The number $s(\omega)$ is directly linked to the environmental fragmentation: a landscape pattern is all the more aggregated as $s(\omega)$ is high, and all the more fragmented as $s(\omega)$ is small (Figure 3). Thus, we shall refer to s as the "habitat aggregation index."

Remark 6. There exist several ways of obtaining hypothetical landscape distributions. The commonest are neutral landscape models, originally introduced by Gardner et al. [16]. They can include parameters which regulate the fragmentation [20]. We preferred to use a stochastic landscape model presented in [34], since it allows an exact control of the favorable and unfavorable surfaces and is therefore well adapted for analyzing the effects of fragmentation per se. This model is inspired from statistical physics. The number of pairs of similar neighbors s is controlled during the process of landscape generation. This quantity can be measured a posteriori on the landscape samples. Other measures of fragmentation could have been used, such as fractal dimension (see [24]). For a discussion on the different ways of measuring habitat fragmentation in real-world situations, the interested reader can refer to [14].

For our computations, we took $\mu^+ = 10$ and $\mu^- = 0$, and we computed the corresponding values of λ_1^i, δ_1^i, and δ_2^i on each landscape sample ω^i of aggregation index s^i, for $i = 1, \dots, 2000$. The eigenvalues λ_1^i were computed with a finite elements method. We fitted the data sets $\{(s^i, \delta_1^i)\}_{i=1,\dots,2000}$ and $\{(s^i, \delta_2^i)\}_{i=1,\dots,2000}$ using ninth degree polynomials (it is enough to assess whether the relations between s and δ_1, δ_2 tend to be monotonic or not). The resulting fitted curves $\delta_{1,f}$ and $\delta_{2,f}$ are presented

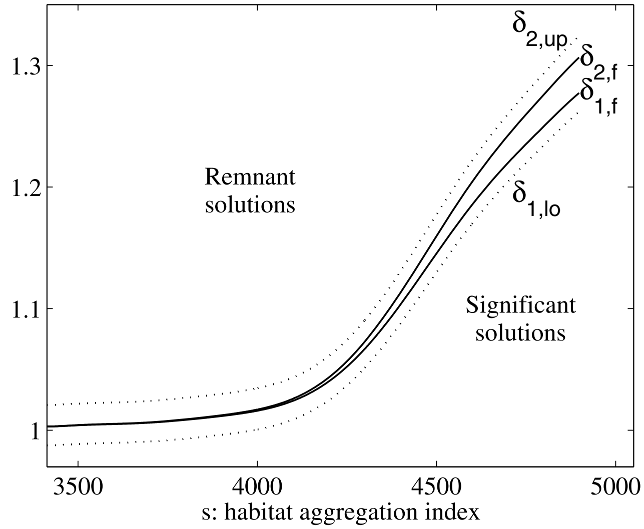


FIG. 4. *Solid lines*: $\delta_{1,f}$ *and* $\delta_{2,f}$ *correspond respectively to the data sets* $\{(s^i,\delta_1^i)\}_{i=1,\dots,2000}$ *and* $\{(s^i,\delta_2^i)\}_{i=1,\dots,2000}$, *fitted with ninth degree polynomials. Dashed lines*: $\delta_{1,lo}$ *is a lower prediction bound for new observations of* δ_1, *and* $\delta_{2,up}$ *an upper prediction bound for new observations of* δ_2, *with in both cases a certainty level of* 99%.

in Figure 4. Under the assumption of normally distributed values of δ_1 and δ_2 for fixed s values, we computed a lower prediction bound ($\delta_{1,lo}$) for new observation of δ_1 and an upper prediction bound for δ_2 ($\delta_{2,up}$), with a level of certainty of 99%. Thus, given a configuration ω, with a fixed value of s, when δ is smaller than $\delta_{1,lo}$ we take a 0.5% chance of being above δ_1, while when δ is larger than $\delta_{2,up}$ we take a 0.5% chance of being below δ_2. The small thickness of the intervals $(\delta_{1,lo},\delta_{2,up})$ emphasizes the quality of the relationship between the habitat aggregation index s and the maximum sustainable yield $\delta^* \in [\delta_1,\delta_2]$. This also indicates that the criteria of Theorems 2.10 and 2.11 are close to being optimal, at least in some situations.

Furthermore, as we can observe, the values of δ_1 and δ_2 tend to increase as s increases, and thus as the environment aggregates. Since $\delta^* \in [\delta_1,\delta_2]$, we deduce from the computations presented in Figure 4 that δ^* tends to increase with environmental aggregation.

These tests were performed for particular values of μ^+ and μ^-. However, the thickness of the interval (δ_1,δ_2) can be determined for all values of μ^+,μ^- without further numerical computations, provided that $\mu^+-\mu^-=10$. Indeed, let us set $B:=\mu^+-\mu^-$. For a fixed value of B, let $\mu_0(x)$ be a given L-periodic function in $L^\infty(\mathbb{R}^N)$ taking only the two values $\mu_0^+=B$ and $\mu_0^-=0$. Let $\lambda_{1,0}$ be the first eigenvalue of the operator $-\nabla^2-\mu_0 I$ on C, with L-periodicity conditions, ϕ_0 the associated eigenfunction with minimal value $\underline{\phi_0}$, and

$$\delta_{1,0}:=\frac{\lambda_{1,0}^2\underline{\phi_0}}{(1+\underline{\phi_0})^2} \quad \text{and} \quad \delta_{2,0}:=\frac{\lambda_{1,0}^2}{4}.$$

We have the following proposition.

PROPOSITION 3.1. *Assume that* $\mu(x)=\mu_0(x)+\mu^-$, *with* $\mu^->\lambda_{1,0}$. *Let* δ_1 *and* δ_2 *be defined by* (2.15). *Then we have* $\delta_2-\delta_1=(1-\frac{\mu^-}{\lambda_{1,0}})^2(\delta_{2,0}-\delta_{1,0})$.

This result also indicates that the information on δ^* is all the more precise as the growth rate function takes low values. However, the "relative thickness" of the interval (δ_1, δ_2), compared to δ_1, $\frac{\delta_2-\delta_1}{\delta_1}$, does not depend on μ^-, as can be easily seen.

Proof of Proposition 3.1. The relation $\lambda_1[\mu(x)] = \lambda_{1,0} - \mu^-$ is a direct consequence of the uniqueness of the first eigenvalue λ_1. We assume that $\mu^- > \lambda_{1,0}$, so that $\lambda_1[\mu(x)] < 0$. From the uniqueness of the eigenfunction ϕ associated with λ_1, ϕ does not depend on μ^-. Therefore, δ_1 and δ_2 satisfy $\delta_1 = \frac{(\lambda_{1,0}-\mu^-)^2 \underline{\phi_0}}{(1+\underline{\phi_0})^2}$ and $\delta_2 = \frac{(\lambda_{1,0}-\mu^-)^2}{4}$. The result immediately follows. □

4. A few comments on the proportional harvesting model. In this model, the population density u is governed by the equation

$$u_t = D\nabla^2 u + u(\mu(x) - \nu(x)u) - q(x)u, \quad x \in \Omega, \tag{4.1}$$

with L-periodicity of the functions $\mu(x)$, $\nu(x)$, and $q(x)$ in the periodic case, and with Neumann or Dirichlet boundary conditions in the bounded case. Setting

$$\tau(x) := \mu(x) - q(x),$$

this model becomes equivalent to the SKT model (1.3). Hence, many properties of the solutions of this model are described in the existing literature. In particular the existence, nonexistence, and uniqueness results of Theorems 2.2 and 2.4 apply. The condition $\lambda_1[\mu(x) - q(x)] < 0$ is therefore necessary and sufficient for species persistence. Furthermore, the theoretical results of [8], [12], [33], [34] on the effects of habitat arrangement on species persistence are also true for this model.

For instance, when the function $\mu(x)$ is constant, with $\mu(x) \equiv \mu_1 > 0$, and if the domain Ω is convex and symmetric with respect to each axis $\{x_1 = 0\}, \dots, \{x_N = 0\}$, the next result is a straightforward consequence of the paper [8].

THEOREM 4.1. (i) *In the periodic case,* $\lambda_1[\mu_1 - q_k^*(x)] \le \lambda_1[\mu_1 - q(x)]$.

(ii) *In the bounded Dirichlet case,* $\lambda_1[\mu_1 - q_k^*(x)] \le \lambda_1[\mu_1 - q(x)]$.

(iii) *In the bounded Neumann case, if* Ω *is a rectangle,* $\lambda_1[\mu_1 - q_k^\sharp(x)] \le \lambda_1[\mu_1 - q(x)]$.

Here q_k^* denotes the symmetric decreasing Steiner rearrangement of the function q with respect to the variable x_k, and $q_k^\sharp$ denotes the monotone rearrangement of q with respect to x_k (see [8] and [10] for the definition of these rearrangements). These rearrangements of a function q preserve not only its mean value, but also its distribution function. This means that if, for instance, q corresponds to a "patch" function taking the values q_1, q_2, and q_3 in some regions A_1, A_2, and A_3, respectively, with $A_1 + A_2 + A_3 = |C|$, then the areas of the regions where the rearranged functions q^* and $q^\sharp$ take the values q_1, q_2, and q_3 remain equal to A_1, A_2, and A_3, respectively.

Theorem 4.1 combined with Theorem 2.4 says that the spatially rearranged harvesting strategies are better for species survival. This result can be helpful from a resource management point of view. Indeed, the authorities can rearrange the position of the harvested areas in order to improve the chances of population persistence. The result of Theorem 4.1 shows that, in the framework of these models, the creation of a large reserve gives persistence more chances than the creation of several small reserves, and is in accordance with the former results of [23] and [26] in the Dirichlet case. See Figure 5 for some illustrations in the bounded case with Dirichlet and Neumann boundary conditions.

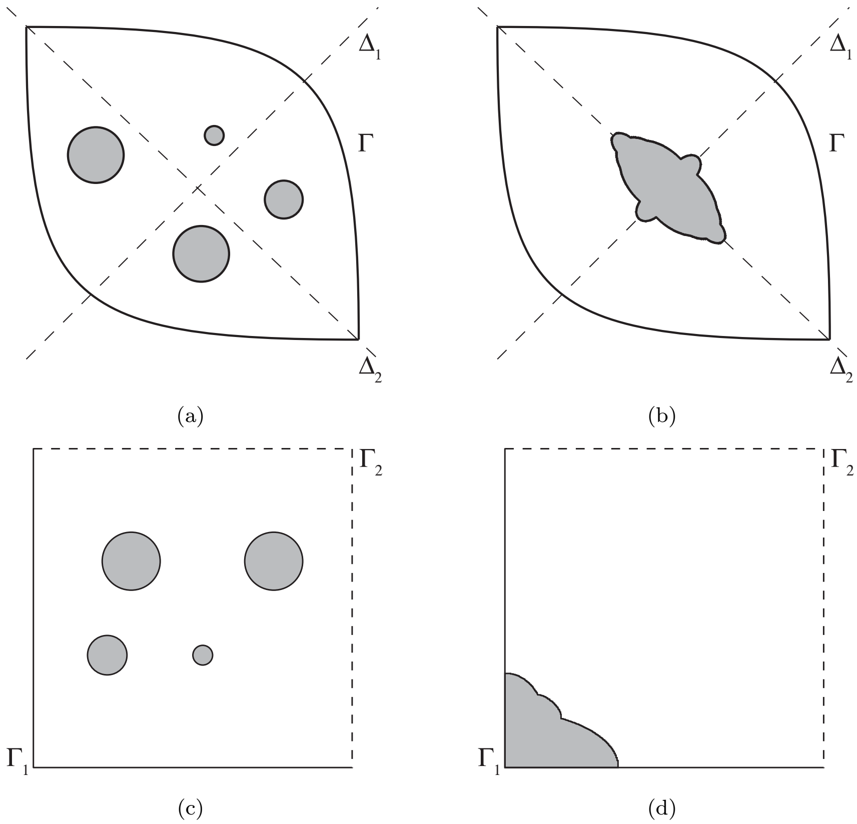


FIG. 5. *Examples of applications of Theorem* 4.1.ii–iii *to reserves management. In panels* (a) *and* (b), *the boundary* Γ *of* Ω *is lethal (Dirichlet boundary conditions).* (a) *The initial effort function* $q(x)$ *takes two values,* $q^+ > 0$ *in the white area, and* $q^- = 0$ *in the shadowed regions, which correspond to reserves.* (b) *Position of the reserves after a symmetric decreasing Steiner rearrangement along the* Δ_1 *and* Δ_2 *axes, successively. The rearranged configuration* (b) *always give more chances of species persistence. In panels* (c) *and* (d), *the boundary* Γ *is divided into two parts*: $\Gamma = \Gamma_1 \cup \Gamma_2$. Γ_1 *is represented with a solid line and can correspond to a coast, while* Γ_2 *is represented with a dashed line and can correspond to a nonphysical limit that the species cannot cross (Neumann boundary conditions).* (c) *The effort function* $q(x)$ *again takes two values,* $q^+ > 0$ *in the white area, and* $q^- = 0$ *in the reserves.* (d) *Position of the reserves after monotone rearrangement along the horizontal and vertical axes, successively. The chances of persistence are better in the rearranged configuration* (d).

5. Discussion. We have proposed a model for the study of populations in heterogeneous environments, for populations submitted to an external negative forcing term. This forcing term could be regarded as a "quasi-constant-yield" harvesting, depending only on the population density u when u is below a certain small threshold ε. The introduction of such a threshold ε was necessary for ensuring the nonnegativity of the solutions of our model, and therefore its actuality.

We carried out new mathematical results on the elliptic equation satisfied by the stationary states of the model, and on the associated parabolic equation. Both qualitative and quantitative results were obtained.

From the qualitative point of view, we described the behavior of the model solutions in terms of the harvesting amplitude δ. Two main types of stationary solutions were found: the remnant solutions, always below a small threshold ε_0 and therefore close to 0, and the significant solutions, always above this threshold, thus ensuring a time-constant yield. We discussed the maximum number of significant stationary solutions, which we found equal to 2, under a hypothesis of positivity of the second eigenvalue λ_2 of a linear operator. We further investigated the long-time behavior of the solution of our model, starting from a nonharvested population at equilibrium. We found a critical value δ^* of the harvesting term amplitude, below which the population density tends over time to a significant stationary solution, and above which it converges to a stationary solution which is not significant. We also established quantitative formulae for some lower and upper bounds for δ^*: δ_1 and δ_2, respectively. The threshold δ_2 has the additional property that, whenever the amplitude δ is above δ_2, the population density decreases to a remnant stationary solution.

The quantitative aspects of our study mainly consisted of discussing the effect of environmental fragmentation on these thresholds δ_1 and δ_2, and therefore on the interactions between environmental fragmentation and maximum sustainable yield. Namely, when computing the values of δ_1 and δ_2 on 2000 samples of stochastically obtained patchy environments, with different levels of fragmentation, we found an increasing relationship between these two coefficients and an environmental aggregation index s. This indicates that, for given areas of favorable and unfavorable regions, the harvesting quota that a species can sustain, while ensuring a time-constant yield, is higher when the favorable regions are aggregated.

The reader may note that, in our model, the species mobility was not affected by the environmental heterogeneity. Such a dependence could be modeled by using a more general dispersion term, of the form $\nabla \cdot (A(x)\nabla u)$, instead of $D\nabla^2 u$, where $A(x)$ stands for the diffusion matrix (see [8], [36]). In fact, most of our results still work when the matrix A is of class $C^{1,\alpha}$ (with $\alpha > 0$) and uniformly elliptic, i.e., when there exists $\tau > 0$ such that $A(x) \geq \tau I_N$ for all $x \in \Omega$. Indeed, Theorems 2.2, 2.4, 2.7, 2.10, and 2.11 remain true under this more general assumption. However, the effects of environmental heterogeneity may differ, depending on the way $A(x)$ and $\mu(x)$ are correlated (see [21]). In the proportional harvesting case, the results of section 4 on the effects of the arrangements of the harvested regions may also not be valid with this dispersion term. However, in situations where $A(x)$ takes low values (slow motion) when $q(x)$ is low ("reserves"; see section 4), as underlined in [33], a simultaneous rearrangement of the functions $A(x)$ and $q(x)$ would lead to lower λ_1 values and therefore to higher chances of species survival.

Acknowledgment. The authors would like to thank the anonymous referees for their valuable suggestions and insightful comments.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[3] H. Amann, *Supersolution, monotone iteration and stability*, J. Differential Equations, 21 (1976), pp. 367–377.

[4] J. E. M. Baillie, C. Hilton-Taylor, and S. N. Stuart, eds., 2004 *IUCN Red List of Threatened Species. A Global Species Assessment*, IUCN, Gland, Switzerland, Cambridge, UK, 2004.

[5] J. R. Beddington and R. M. May, *Harvesting natural populations in a randomly fluctuating environment*, Science, 197 (1977), pp. 463–465.

[6] Z. Belhachmi, D. Bucur, G. Buttazzo, and J.-M. Sac-Epée, *Shape optimization problems for eigenvalues of elliptic operators*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 171–184.

[7] H. Berestycki and F. Hamel, *Front propagation in periodic excitable media*, Comm. Pure Appl. Math., 55 (2002), pp. 949–1032.

[8] H. Berestycki, F. Hamel, and L. Roques, *Analysis of the periodically fragmented environment model:* I—*Species persistence*, J. Math. Biol., 51 (2005), pp. 75–113.

[9] H. Berestycki, F. Hamel, and L. Roques, *Analysis of the periodically fragmented environment model:* II—*Biological invasions and pulsating travelling fronts*, J. Math. Pures Appl., 84 (2005), pp. 1101–1146.

[10] H. Berestycki and T. Lachand-Robert, *On the monotone rearrangement in cylinders and applications*, Math. Nachr., 266 (2004), pp. 3–19.

[11] J. H. Bramble and L. E. Payne, *Bounds in the Neumann problem for second order uniformly elliptic operators*, Pacific J. Math., 12 (1962), pp. 823–833.

[12] R. S. Cantrell and C. Cosner, *Spatial Ecology via Reaction-Diffusion Equations*, Ser. Math. Comput. Biol., John Wiley and Sons, Chichester, UK, 2003.

[13] M. D. Chekroun and L. Roques, *Spatially-Explicit Harvesting Models. The Influence of Seasonal Variations*, in preparation.

[14] L. Fahrig, *Effects of habitat fragmentation on biodiversity*, Ann. Rev. Ecol. Syst., 34 (2003), pp. 487–515.

[15] R. A. Fisher, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 335–369.

[16] R. H. Gardner, B. T. Milne, M. G. Turner, and R. V. O'Neill, *Neutral models for the analysis of broad-scale landscape pattern*, Landscape Ecol., 1 (1987), pp. 19–28.

[17] W. M. Getz and R. G. Haight, *Population Harvesting: Demographic Models of Fish, Forests and Animal Resources*, Princeton Monographs in Population Biology, Princeton University Press, Princeton, NJ, 1989.

[18] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[19] B. Kawohl, *On the isoperimetric nature of a rearrangement inequality and its consequences for some variational problems*, Arch. Ration. Mech. Anal., 94 (1986), pp. 227–243.

[20] T. H. Keitt, *Spectral representation of neutral landscapes*, Landscape Ecol., 15 (2000), pp. 479–494.

[21] N. Kinezaki, K. Kawasaki, and N. Shigesada, *Spatial dynamics of invasion in sinusoidally varying environments*, Population Ecol., 48 (2006), pp. 263–270.

[22] A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov, *Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bulletin Université d'État à Moscou (Bjul. Moskowskogo Gos. Univ.), Série internationale A, 1 (1937), pp. 1–26.

[23] K. Kurata and J. Shi, *Optimal Spatial Harvesting Strategy and Symmetry-Breaking*, preprint.

[24] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1982.

[25] J. D. Murray and R. P. Sperb, *Minimum domains for spatial patterns in a class of reaction-diffusion equations*, J. Math. Biol., 18 (1983), pp. 169–184.

[26] M. G. Neubert, *Marine reserves and optimal harvesting*, Ecol. Lett., 6 (2003), pp. 843–849.

[27] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, Courant Lecture Notes 6, AMS, Providence, RI, 2001.

[28] A. Okubo and S. A. Levin, *Diffusion and Ecological Problems—Modern Perspectives*, 2nd ed., Springer-Verlag, New York, 2002.

[29] S. Oruganti, R. Shivaji, and J. Shi, *Diffusive logistic equation with constant effort harvesting,* I: *Steady states*, Trans. Amer. Math. Soc., 354 (2002), pp. 3601–3619.

[30] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292.

[31] J. G. Robinson and R. E. Bodmer, *Towards wildlife management in tropical forests*, J. Wildlife Management, 63 (1999), pp. 1–13.

[32] J. G. Robinson and K. H. Redford, *Sustainable harvest of neo-tropical mammals*, in Neo-Tropical Wildlife Use and Conservation, J. G. Robinson and K. H. Redford, eds., Chicago University Press, Chicago, IL, 1991, pp. 415–429.

[33] L. Roques and F. Hamel, *Mathematical analysis of the optimal habitat configurations for species persistence*, Math. Biosci., to appear; DOI 10.1016/j.mbs.2007.05.007.

[34] L. Roques and R. Stoica, *Species persistence decreases with habitat fragmentation: An analysis in periodic stochastic environments*, J. Math. Biol., 55 (2007), pp. 189–205.

[35] M. B. Schaefer, *Some considerations of population dynamics and economics in relation to the management of the commercial marine fisheries*, J. Fish. Res. Board Can., 14 (1957), pp. 669–681.

[36] N. Shigesada and K. Kawasaki, *Biological Invasions: Theory and Practice*, Oxford Series in Ecology and Evolution, Oxford University Press, Oxford, UK, 1997.

[37] N. Shigesada, K. Kawasaki, and E. Teramoto, *Traveling periodic waves in heterogeneous environments*, Theoret. Population Biol., 30 (1986), pp. 143–160.

[38] P. A. Stephens, F. Frey-Roos, W. Arnold, and W. J. Sutherland, *Sustainable exploitation of social species: A test and comparison of models*, J. Appl. Ecol., 39 (2002), pp. 629–642.

[39] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Appl. Math. Sci., 68, Springer-Verlag, New York, 1997.

[40] P. Turchin, *Quantitative Analysis of Movement: Measuring and Modeling Population Redistribution in Animals and Plants*, Sinauer Associates, Sunderland, MA, 1998.

SIAM J. APPL. MATH.
Vol. 68, No. 1, pp. 154–170

EFFECTS OF GENERAL INCIDENCE AND POLYMER JOINING ON NUCLEATED POLYMERIZATION IN A MODEL OF PRION PROLIFERATION*

MEREDITH L. GREER†, P. VAN DEN DRIESSCHE‡, LIN WANG§, AND G. F. WEBB¶

Abstract. Two processes are incorporated into a new model for transmissible prion diseases. These are general incidence for the lengthening process of infectious polymers attaching to and converting noninfectious monomers, and the joining of two polymers to form one longer polymer. The model gives rise to a system of three ordinary differential equations, which is shown to exhibit threshold behavior dependent on the value of the parameter combination giving the basic reproduction number $\mathcal{R}_0$. For $\mathcal{R}_0 < 1$, infectious polymers die out, whereas for $\mathcal{R}_0 > 1$, the system is locally asymptotic to a positive disease equilibrium. The effect of both general incidence and joining is to decrease the equilibrium value of infectious polymers and to increase the equilibrium value of normal monomers. Since the onset of disease symptoms appears to be related to the number of infectious polymers, both processes may significantly inhibit the course of the disease. With general incidence, the equilibrium distribution of polymer lengths is obtained and shows a sharp decrease in comparison to the distribution resulting from mass action incidence. Qualitative global results on the disease free and disease equilibria are proved analytically. Numerical simulations using parameter values from experiments on mice (reported in the literature) provide quantitative demonstration of the effects of these two processes.

Key words. prion diseases, nucleated polymerization, general incidence, polymer joining, prion proliferation

AMS subject classifications. 92D30, 34D23, 35Q80

DOI. 10.1137/06066076X

1. Introduction.

Prion diseases, though widely studied at many levels, continue to challenge understanding. A prion is an infectious protein. Several prion diseases are known, or suspected, to be transmissible, both via ingestion and iatrogenically; as a group, they are thus referred to as transmissible spongiform encephalopathies (TSEs). Examples include scrapie, which affects sheep and goats; bovine spongiform encephalopathy (BSE), which affects cows; chronic wasting disease (CWD), which affects mule deer and elk; and variant Cruetzfeldt–Jakob disease (vCJD), which affects humans [8, 9]. Additionally, mice and hamsters in laboratory experiments can be infected with scrapie [30].

Though incidence of vCJD in humans has declined to just a few new cases per year [36] and BSE incidence also appears to be declining [23], prion diseases warrant ongoing study for reasons that include the following. First, there may be previously unrecognized routes of infection: new research shows that prions can bind to some soils and cause infection via inoculation with those soils [17], indicating that graz-

*Received by the editors May 24, 2006; accepted for publication (in revised form) June 6, 2007; published electronically October 17, 2007. This work was partially supported by NSERC, MITACS, PIMS, and NSF-DMS 0516737.
http://www.siam.org/journals/siap/68-1/66076.html

†Corresponding author. Department of Mathematics, Bates College, 213 Hathorn Hall, Lewiston, ME 04240 (mgreer@bates.edu).

‡Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8W 3P4, Canada (pvdd@math.uvic.ca).

§Department of Mathematics and Statistics, University of New Brunswick, Fredericton, NB E3B 5A3, Canada (lwang2@unb.ca).

¶Vanderbilt University, Department of Mathematics, 1326 Stevenson Center, Nashville, TN 37240-0001 (glenn.f.webb@vanderbilt.edu).

ing animals may acquire TSEs despite having safely prepared feed. Second, prions are extremely difficult to destroy, remaining infective despite heat or radiation that would inactivate other known infectious agents [1, 5]. Third, prion replication offers a new frontier in scientific understanding: protein-only replication cannot depend on nucleic acids, but must occur somehow for TSEs to spread. Comprehending how this replication works may provide great insight to other biological processes.

A specific naturally occurring protein is vulnerable to infection by prions; it is therefore known as prion protein. In its noninfectious form prion protein is denoted by $\mathrm{PrP^C}$, and in its infectious form it is denoted by $\mathrm{PrP^{Sc}}$; see, for example, [22] for discussion of this notation. The forms differ only in the folding of the protein [27]. Humans, cows, sheep, and other animals susceptible to TSEs produce $\mathrm{PrP^C}$ normally [4]. There is evidence both that an accumulation of $\mathrm{PrP^{Sc}}$ may be toxic [24, 21] and that a lack of $\mathrm{PrP^C}$ may leave the brain overly susceptible to stress [29]. Either or both of these may lead to symptoms associated with TSEs. In the case of transmissible prion disease, some portion of $\mathrm{PrP^{Sc}}$ is introduced into the system, and this $\mathrm{PrP^{Sc}}$ can cause more infectious protein to be made. Though the mechanism for such protein replication is not fully understood, nucleated polymerization is a likely candidate [15, 18].

Nucleated polymerization involves $\mathrm{PrP^{Sc}}$ attaching to $\mathrm{PrP^C}$ and converting it to $\mathrm{PrP^{Sc}}$. While proteins usually exist as individual units, also known as monomers, it appears that $\mathrm{PrP^{Sc}}$ benefits from aggregating in some way [11, 16]. Aggregation confers greater stability, and may even be necessary to maintain the alternate protein folding. We assume within this paper that these aggregates have a linear form [18, 26], and we typically refer to the aggregates as polymers. In our nucleated polymerization model, each polymer may attach at either end to a $\mathrm{PrP^C}$ monomer, quickly converting it to the infectious form of $\mathrm{PrP^{Sc}}$. Since the polymer has thus increased its length by one unit of protein, we refer to this process as *lengthening*. Nucleated polymerization also involves polymer *splitting*. We assume a minimum viable polymer length, so that when polymer splitting results in pieces below the minimum length, these pieces must break apart into their component units of $\mathrm{PrP^C}$. Additionally, our model includes polymer *joining*, in which two $\mathrm{PrP^{Sc}}$ polymers join together to form one longer polymer.

Models of nucleated polymerization for $\mathrm{PrP^C}$ monomers and $\mathrm{PrP^{Sc}}$ polymers containing a discrete number of monomers are formulated and analyzed in [20] and [22]. Based on these, a model with continuous polymer length is introduced in [13] and further analyzed in [12, 14, 28]. All these models assume mass action incidence for the lengthening process of infectious polymers attaching to $\mathrm{PrP^C}$ units. We generalize this form of incidence in a way that reduces lengthening when the total amount of infectious protein becomes large in proportion to the number of polymers. Some research [24] has indicated that only truncated forms of polymers are able to lengthen this way; it is also possible that polymers within a specific range of lengths are able to lengthen at the fastest rate, but that all polymers are capable of lengthening [31]. Our general incidence term captures these features by reducing the rate of lengthening when total $\mathrm{PrP^{Sc}}$ mass is large relative to the total number of $\mathrm{PrP^{Sc}}$ polymers. That is, we reduce the rate of lengthening as the average polymer length becomes greater. In addition, our model is the first to include polymer joining. Joining is implied by the fact that large fibrils or aggregates of $\mathrm{PrP^{Sc}}$ are observed in late stages of disease [2, 10].

We start in section 2 by incorporating the processes of general incidence and polymer joining into an ordinary differential equation (ODE) for the number of monomers, coupled with a partial integro-differential equation for the density of polymers depending on polymer length. Under some assumptions, this system is converted to a system of three ODEs, which is analyzed in section 3. Numerical simulations for parameters obtained from experimental data on mice [30] are presented in section 4, and we conclude in section 5 with a discussion.

2. Model formulation. A core model of nucleated polymerization exists in [12, 13, 14, 28] with some extensions in [32]. We continue to use the same variables and parameters and introduce two new parameters, ω and η, to account for general incidence and polymer joining, respectively:

- $V(t)$ is the number of $\mathrm{PrP^C}$ monomers at time t;
- $u(x,t)$ is the density of $\mathrm{PrP^{Sc}}$ polymers of length x at time t;
- x_0 is the lower bound for polymer length; that is, polymers have length x with $x_0 < x < \infty$;
- $U(t) = \int_{x_0}^{\infty} u(x,t)dx$ is the number of $\mathrm{PrP^{Sc}}$ polymers at time t;
- $P(t) = \int_{x_0}^{\infty} xu(x,t)dx$ is the number of $\mathrm{PrP^{Sc}}$ monomers comprising polymers at time t;
- $W(t) = P(t) - x_0U(t)$ is the number of $\mathrm{PrP^{Sc}}$ units not accounted for within the minimal polymer lengths;
- λ is the source rate for naturally produced $\mathrm{PrP^C}$ monomers;
- γ is the metabolic degradation rate for $\mathrm{PrP^C}$;
- τ is a rate associated with lengthening of $\mathrm{PrP^{Sc}}$ polymers by attaching to and converting $\mathrm{PrP^C}$ monomers;
- ω is a parameter associated with polymer lengthening;
- $\beta(x)$ is the length-dependent rate of polymer breakage;
- $\kappa(x,y)$ is the probability, when a polymer of length y breaks, that one of the two resulting polymers has length x;
- $\mu(x)$ is the length-dependent metabolic degradation rate of $\mathrm{PrP^{Sc}}$ polymers;
- η is the rate at which $\mathrm{PrP^{Sc}}$ polymers join together.

All parameters are assumed to be positive with the exception of ω and η, which may also be zero.

2.1. PDE model. Our model, incorporating both general incidence and polymer joining into the model formulated and discussed in [12, 13, 14, 28], has monomer dynamics governed by

$$V'(t) = \lambda - \gamma V(t) - \frac{\tau V(t)U(t)}{1+\omega P(t)} + 2\int_0^{x_0} x \int_{x_0}^{\infty} \beta(y)\kappa(x,y)u(y,t)dydx \tag{2.1}$$

with $V'(t) = \frac{dV}{dt}$, and polymer dynamics given by

$$\begin{aligned} u_t(x,t) + \frac{\tau V(t)}{1+\omega P(t)}u_x(x,t) &= -(\mu(x)+\beta(x))u(x,t) + 2\int_x^{\infty} \beta(y)\kappa(x,y)u(y,t)dy \\ &\quad + \eta\int_{x_0}^{x} u(x-y,t)u(y,t)dy - 2\eta u(x,t)\int_{x_0}^{\infty} u(y,t)dy, \end{aligned} \tag{2.2}$$

subject to nonnegative initial conditions and the boundary condition

$$u(x_0,t) = 0. \tag{2.3}$$

We write the polymer lengthening term in (2.1) in the general form $\frac{\tau V(t)U(t)}{1+\omega P(t)}$. Note that in the case $\omega = 0$ this is a mass action term. Otherwise, as $P(t)$ becomes large there is a saturation effect, with the result that less lengthening occurs overall. This matches the in vitro observations of [24, 31].

The polymer joining term $\eta\int_{x_0}^{x} u(x-y,t)u(y,t)dy$ introduces the joining parameter η and indicates that a new polymer of length x results from the joining of two smaller polymers of lengths $x-y$ and y. Note that the upper integration limit can be written as x or $x-x_0$ with identical results, as there are zero polymers of length less than x_0. Changing the form of the integration limit does not affect analysis of the model. The last term $2\eta u(x,t)\int_{x_0}^{\infty} u(y,t)dy$ describes the loss of a polymer of length x when it joins with another polymer, of any length, to create a larger polymer. Symmetry mandates the factor 2.

Note that with mass action incidence and no polymer joining, i.e., $\omega = 0$ and $\eta = 0$, our model reduces to that in [12, 13, 14, 28]. For this case, a model with bounded $\beta(x)$, $\mu(x)$, and a general kernel $\kappa(x,y)$ is analyzed in [32].

2.2. Conversion to ODEs. Under an assumption of equidistributed splitting, a system of three ODEs in V, U, and P can be obtained from (2.1) and (2.2). Equidistributed splitting means that splitting is equally likely wherever two protein units have joined together; hence the splitting rate $\beta(x)$ is proportional to polymer length x, i.e., $\beta(x) = \beta x$. The accompanying splitting kernel is then

$$\kappa(x,y) = \begin{cases} 1/y & \text{if} \quad y > x_0 \ \text{ and } \ 0 < x < y, \\ 0 & \text{if} \quad y \le x_0 \ \text{ or } \ y \le x. \end{cases}$$

We make the additional assumption that polymer metabolic degradation occurs at a constant rate, i.e., $\mu(x) = \mu$. A form of the PDEs that assumes mass action and no polymer joining was converted to ODEs in [28] by integrating (2.2), and integrating the product of x and (2.2), over $[x_0,\infty)$. Proceeding similarly, the general incidence term is independent of x and converts analogously. The joining integral $\int_{x_0}^{\infty} u(x,t)\int_{x_0}^{\infty} u(y,t)dydx$ simplifies to $U^2(t)$. The remaining joining integral from (2.2) gives

$$\begin{aligned}\int_{x_0}^{\infty}\int_{x_0}^{x} u(y,t)u(x-y,t)\,dy\,dx &= \int_{x_0}^{\infty}\int_{0}^{x-x_0} u(x-z,t)u(z,t)\,dz\,dx \\ &= \int_{0}^{\infty}\int_{z+x_0}^{\infty} u(x-z,t)u(z,t)\,dx\,dz \\ &= \int_{0}^{\infty}\int_{x_0}^{\infty} u(w,t)u(z,t)\,dw\,dz \\ &= U^2(t).\end{aligned}$$

The resulting system of equations is

$$\begin{aligned} U' &= \beta P - \mu U - 2\beta x_0 U - \eta U^2, \\ V' &= \lambda - \gamma V - \frac{\tau V U}{1+\omega P} + \beta x_0^2 U, \\ P' &= \frac{\tau V U}{1+\omega P} - \mu P - \beta x_0^2 U. \end{aligned} \tag{2.4}$$

It is useful to have equations for infectious polymers, noninfectious monomers, and infectious monomers comprising polymers. To use analysis appropriate for compartmental models, we replace the U equation with an equation for $x_0 U$, and the

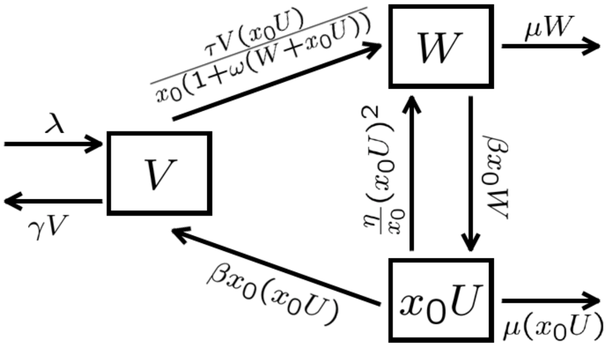


FIG. 1. *Compartmental diagram of system* (2.5).

P equation with an equation for $W = P - x_0U$. The x_0U compartment contains all PrP^{Sc} units that make up the minimum lengths of the polymers. The W compartment contains all additional PrP^{Sc} units. The resulting system of equations is

$$\begin{aligned}
(x_0U)' &= \beta x_0 W - \mu(x_0U) - \beta x_0(x_0U) - \frac{\eta}{x_0}(x_0U)^2,\\
V' &= \lambda - \gamma V - \frac{\tau V(x_0U)}{x_0(1+\omega(W+x_0U))} + \beta x_0^2 U,\\
W' &= \frac{\tau V(x_0U)}{x_0(1+\omega(W+x_0U))} - (\mu+\beta x_0)W + \frac{\eta}{x_0}(x_0U)^2.
\end{aligned} \tag{2.5}$$

The compartmental diagram of this system appears in Figure 1.

3. Model analysis.

3.1. Nondimensionalization.

To facilitate analysis, rewrite the ODE system (2.5) in a nondimensionalized form. Let $\alpha = \mu + \beta x_0$ and $T = \alpha t$. Rewrite $U(t) = \frac{\alpha}{\tau}\mathcal{X}(T)$, $V(t) = \frac{\alpha^2}{\beta\tau}\mathcal{Y}(T)$, and $W(t) = \frac{\alpha^2}{\beta\tau}\mathcal{Z}(T)$. Define $\sigma = \frac{\beta\lambda\tau}{\alpha^3}$, $\rho = \frac{\gamma}{\alpha}$, $\delta = \frac{\beta x_0}{\alpha}$, $\nu = \frac{\omega\alpha^2}{\beta\tau}$, and $\phi = \frac{\eta}{\tau}$. Then

$$\begin{aligned}
\mathcal{X}' &= \mathcal{Z} - \mathcal{X} - \phi\mathcal{X}^2,\\
\mathcal{Y}' &= \sigma - \rho\mathcal{Y} - \frac{\mathcal{X}\mathcal{Y}}{1+\nu(\mathcal{Z}+\delta\mathcal{X})} + \delta^2\mathcal{X},\\
\mathcal{Z}' &= \frac{\mathcal{X}\mathcal{Y}}{1+\nu(\mathcal{Z}+\delta\mathcal{X})} - \mathcal{Z} + \delta\phi\mathcal{X}^2,
\end{aligned} \tag{3.1}$$

with $\mathcal{X}' = \frac{d\mathcal{X}}{dT}$. The nondimensionalization process reduces the number of parameters from eight to five. Note that $\delta = \frac{\beta x_0}{\beta x_0 + \mu} \in (0,1)$. Setting $\nu = 0$ simplifies the incidence term to mass action, whereas setting $\phi = 0$ simplifies the model to the case with no polymer joining.

In all that follows, disease is assumed to be initially present; thus the nonnegative initial conditions for the nondimensional system are $\mathcal{X}(0) \geq 0$, $\mathcal{Y}(0) \geq 0$, $\mathcal{Z}(0) \geq 0$, with $\mathcal{X}(0) + \mathcal{Z}(0) > 0$.

PROPOSITION 3.1. *Let $\nu, \phi \geq 0$, $\sigma, \rho > 0$, and $\delta \in (0,1)$. For each $(\mathcal{X}(0), \mathcal{Y}(0), \mathcal{Z}(0)) \in \mathbb{R}^3_+$ the system* (3.1) *has a unique bounded solution in $\mathbb{R}^3_+$ defined for all $T \geq 0$.*

Proof. Let $F : \mathbb{R}^3_+ \to \mathbb{R}^3_+$ be given by

$$\begin{aligned} F((\mathcal{X},\mathcal{Y},\mathcal{Z})) &= (F_1, F_2, F_3) \\ &= \left(\mathcal{Z} - \mathcal{X} - \phi\mathcal{X}^2, \sigma - \rho\mathcal{Y} - \frac{\mathcal{X}\mathcal{Y}}{1+\nu(\mathcal{Z}+\delta\mathcal{X})} + \delta^2\mathcal{X}, \frac{\mathcal{X}\mathcal{Y}}{1+\nu(\mathcal{Z}+\delta\mathcal{X})} - \mathcal{Z} + \delta\phi\mathcal{X}^2\right), \end{aligned}$$

and observe that F is Lipschitz continuous on bounded sets of $\mathbb{R}^3_+$. For $T \geq 0$ and $(\mathcal{X},\mathcal{Y},\mathcal{Z}) \in \mathbb{R}^3_+$, it follows that $F_1 \geq 0$ when $\mathcal{X} = 0$, $F_2 \geq 0$ when $\mathcal{Y} = 0$, and $F_3 \geq 0$ when $\mathcal{Z} = 0$. Thus by Corollary A.5 in [35] there exists a unique nonnegative solution to (3.1) in $\mathbb{R}^3_+$ for $T \in [0,\infty)$. Since

$$\begin{aligned} \frac{d}{dT}(\delta\mathcal{X}(T) + \mathcal{Y}(T) + \mathcal{Z}(T)) &= \sigma - (1-\delta)\delta\mathcal{X}(T) - \rho\mathcal{Y}(T) - (1-\delta)\mathcal{Z}(T) \\ &\leq \sigma - \theta(\delta\mathcal{X}(T) + \mathcal{Y}(T) + \mathcal{Z}(T)), \end{aligned}$$

where $\theta = \min\{1-\delta, \rho\} > 0$, it follows that $\delta\mathcal{X}(T) + \mathcal{Y}(T) + \mathcal{Z}(T) \leq \max\{\frac{\sigma}{\theta}, \delta\mathcal{X}(0) + \mathcal{Y}(0) + \mathcal{Z}(0)\} = M$. Thus the existence of a unique global nonnegative bounded solution is proved. □

3.2. Computing and interpreting $\mathcal{R}_0$. The disease free equilibrium (DFE) for this nondimensionalized general model of nucleated polymerization is $(\bar{\mathcal{X}}, \bar{\mathcal{Y}}, \bar{\mathcal{Z}}) = (0, \frac{\sigma}{\rho}, 0)$. Note that in the absence of disease, $\bar{\mathcal{Y}}$ is stable. The DFE may be used to find the basic reproduction number $\mathcal{R}_0$, which indicates the average number of new infections caused by a single infective introduced to an entirely susceptible population. One technique [37] examines the infective compartments, in this case the equations within (3.1) for $\mathcal{X}$ and $\mathcal{Z}$. The Jacobian J of the $(\mathcal{X},\mathcal{Z})$ system about the DFE is apportioned into two matrices F and G such that $J = F - G$, where F contains all elements resulting from new infections and G contains all remaining movement between compartments. Then $\mathcal{R}_0$ is the spectral radius of the matrix FG^{-1}. For the model given in (3.1),

$$F = \begin{bmatrix} 0 & 0 \\ \frac{\sigma}{\rho} & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix},$$

and the spectral radius of FG^{-1} is $\frac{\sigma}{\rho}$. Hence $\mathcal{R}_0 = \frac{\sigma}{\rho}$. The next result follows from Theorem 2 of [37].

LEMMA 3.2. *If* $\mathcal{R}_0 < 1$, *then the DFE of* (3.1) *is locally asymptotically stable*; *if* $\mathcal{R}_0 > 1$, *then the DFE is unstable.*

In the biological variables,

$$\mathcal{R}_0 = \frac{\beta\lambda\tau}{\gamma(\beta x_0 + \mu)^2}. \tag{3.2}$$

The same $\mathcal{R}_0$ results from the model of nucleated polymerization in which the lengthening mechanism proceeds according to mass action and polymer joining does not occur [28]. This result makes it clear that general incidence and joining do not affect the potential success of infection via nucleated polymerization. However, as shown later, the inclusion of a generalized incidence term and polymer joining does affect the distribution of polymer lengths as time progresses during disease and alters the disease equilibrium.

3.3. Global stability of the DFE. Assuming that $\rho \geq 1$, the DFE of the general model of nucleated polymerization given in (3.1) is globally attractive for $\mathcal{R}_0 \leq 1$ and globally asymptotically stable (GAS) for $\mathcal{R}_0 < 1$. The assumption that $\rho \geq 1$ is justified biologically; see data in section 4. To show the DFE results, first consider the model in the case with $\nu = 0$, that is, where polymer lengthening occurs via mass action.

LEMMA 3.3. *If $\mathcal{R}_0 \leq 1$, then the DFE $(\bar{\mathcal{X}}, \bar{\mathcal{Y}}, \bar{\mathcal{Z}}) = (0, \frac{\sigma}{\rho}, 0)$ of the system*

$$\begin{aligned} \mathcal{X}' &= \mathcal{Z} - \mathcal{X} - \phi\mathcal{X}^2, \\ \mathcal{Y}' &= \sigma - \rho\mathcal{Y} - \mathcal{X}\mathcal{Y} + \delta^2\mathcal{X}, \\ \mathcal{Z}' &= \mathcal{X}\mathcal{Y} - \mathcal{Z} + \delta\phi\mathcal{X}^2 \end{aligned} \tag{3.3}$$

is globally attractive.

Proof. Consider the Liapunov function

$$\Phi = \frac{1}{2}(\mathcal{Y} - \bar{\mathcal{Y}})^2 + k_1(\mathcal{X} + \mathcal{Z})$$

with $k_1 = (2 - \delta^2 - \bar{\mathcal{Y}})$. Since both $\delta < 1$ and $\bar{\mathcal{Y}} = \mathcal{R}_0 \leq 1$, then $k_1 > 0$. This Liapunov function is the same as that used by [28] for the nondimensionalized model with mass action and no joining. Its derivative given by

$$\Phi' = -\rho(\mathcal{Y} - \bar{\mathcal{Y}})^2 - \phi(1-\delta)k_1\mathcal{X}^2 - \mathcal{X}[(\mathcal{Y} - 1)^2 + (1 - \delta^2)(1 - \bar{\mathcal{Y}})]$$

is nonpositive for $\mathcal{R}_0 \leq 1$. Also $\Phi' = 0$ only if $\mathcal{Y} = \bar{\mathcal{Y}}$ and $\mathcal{X} = 0$. Thus by LaSalle's invariance principle [19] the DFE $(0, \frac{\sigma}{\rho}, 0)$ of (3.3) is globally attractive. □

THEOREM 3.4. *Assume $\rho \geq 1$. If $\mathcal{R}_0 \leq 1$, then the DFE $(\bar{\mathcal{X}}, \bar{\mathcal{Y}}, \bar{\mathcal{Z}}) = (0, \frac{\sigma}{\rho}, 0)$ of the system* (3.1) *is globally attractive. If $\mathcal{R}_0 < 1$, then the DFE is GAS.*

Proof. From systems (3.1) and (3.3), create the equivalent respective systems

$$\begin{aligned} \mathcal{X}' &= \mathcal{Z} - \mathcal{X} - \phi\mathcal{X}^2, \\ \mathcal{Z}' &= \frac{\mathcal{X}(\mathcal{Y} + \mathcal{Z}) - \mathcal{X}\mathcal{Z}}{1 + \nu(\mathcal{Z} + \delta\mathcal{X})} - \mathcal{Z} + \delta\phi\mathcal{X}^2, \\ (\mathcal{Y} + \mathcal{Z})' &= \sigma - \rho(\mathcal{Y} + \mathcal{Z}) + (\rho - 1)\mathcal{Z} + \delta^2\mathcal{X} + \delta\phi\mathcal{X}^2, \end{aligned} \tag{3.4}$$

and (with $\nu = 0$)

$$\begin{aligned} \mathcal{X}' &= \mathcal{Z} - \mathcal{X} - \phi\mathcal{X}^2, \\ \mathcal{Z}' &= \mathcal{X}(\mathcal{Y} + \mathcal{Z}) - \mathcal{X}\mathcal{Z} - \mathcal{Z} + \delta\phi\mathcal{X}^2, \\ (\mathcal{Y} + \mathcal{Z})' &= \sigma - \rho(\mathcal{Y} + \mathcal{Z}) + (\rho - 1)\mathcal{Z} + \delta^2\mathcal{X} + \delta\phi\mathcal{X}^2, \end{aligned} \tag{3.5}$$

subject to the same nonnegative initial conditions. Since $\rho \geq 1$, system (3.5) is K-monotone. Then $\nu \geq 0$, $\delta \in (0,1)$, $\mathcal{X} \geq 0$, and $\mathcal{Z} \geq 0$ imply by a standard comparison theorem given in [34, Appendix B1] that $(\mathcal{X}, \mathcal{Z}, \mathcal{Y} + \mathcal{Z})_{(3.4)} \leq (\mathcal{X}, \mathcal{Z}, \mathcal{Y} + \mathcal{Z})_{(3.5)}$. Let $\mathcal{R}_0 \leq 1$. Since by Lemma 3.3, the DFE of (3.5) is globally attractive, and by Proposition 3.1, $\mathcal{X} \geq 0$ and $\mathcal{Z} \geq 0$, it follows that $\mathcal{X} \to 0$ and $\mathcal{Z} \to 0$ for system (3.4). From the second equation of (3.1), the theory of asymptotically autonomous systems [6] shows that $\mathcal{Y} \to \frac{\sigma}{\rho}$. The global asymptotic stability result then follows from Lemma 3.2. □

In order to use the comparison theorem to show that the DFE of system (3.1) is GAS, it is required that $\rho \geq 1$ in Theorem 3.4. Next we apply the Liapunov method to establish that the DFE is GAS without assuming that $\rho \geq 1$ but at a cost: $\mathcal{R}_0$ is required to be less than $1-\delta^2$.

THEOREM 3.5. *The DFE of system* (3.1) *is GAS if* $\mathcal{R}_0 \leq 1-\delta^2$.

Proof. Define

$$\Phi = \mathcal{Y} - \bar{\mathcal{Y}} \ln(\mathcal{Y}/\bar{\mathcal{Y}}) + \mathcal{Z} + \mathcal{X}.$$

Notice that $\sigma = \rho\bar{\mathcal{Y}}$. The derivative of Φ along the solution of system (3.1) is given by

$$\begin{aligned}\Phi' &= -\rho\frac{(\mathcal{Y}-\bar{\mathcal{Y}})^2}{\mathcal{Y}} - (1-\delta)\phi\mathcal{X}^2 - \delta^2\bar{\mathcal{Y}}\frac{\mathcal{X}}{\mathcal{Y}} + \mathcal{X}\left(\delta^2 + \frac{\bar{\mathcal{Y}}}{1+\nu(\mathcal{Z}+\delta\mathcal{X})} - 1\right)\\ &\leq -\rho\frac{(\mathcal{Y}-\bar{\mathcal{Y}})^2}{\mathcal{Y}} - (1-\delta)\phi\mathcal{X}^2 - \delta^2\bar{\mathcal{Y}}\frac{\mathcal{X}}{\mathcal{Y}} + \mathcal{X}\left(\delta^2 + \bar{\mathcal{Y}} - 1\right).\end{aligned}$$

This shows that Φ' is nonpositive for $\mathcal{R}_0 = \frac{\sigma}{\rho} = \bar{\mathcal{Y}} < 1-\delta^2$ and $\Phi' = 0$ only if $\mathcal{Y} = \bar{\mathcal{Y}}$ and $\mathcal{X} = 0$. Again, by LaSalle's invariance principle and Lemma 3.2, the DFE is GAS. □

3.4. Existence and stability of the EE. We now consider an endemic equilibrium (EE) with disease present, i.e., $\mathcal{X} > 0$, $\mathcal{Y} > 0$, $\mathcal{Z} > 0$.

LEMMA 3.6. *If* $\mathcal{R}_0 > 1$, *then system* (3.1) *has a unique EE. If* $\phi = 0$, *then that EE is given by* $(\mathcal{X}^*, \mathcal{Y}^*, \mathcal{Z}^*) = \left(\frac{\sigma-\rho}{\rho\nu(1+\delta)+(1-\delta^2)}, \frac{\sigma\nu(1+\delta)+(1-\delta^2)}{\rho\nu(1+\delta)+(1-\delta^2)}, \frac{\sigma-\rho}{\rho\nu(1+\delta)+(1-\delta^2)}\right)$. *If* $\mathcal{R}_0 < 1$, *then* (3.1) *has no EE.*

Proof. If $\phi > 0$, then system (3.1) cannot be solved explicitly for the EE. However, for $\phi \geq 0$, at equilibrium, $\mathcal{Z}$ and $\mathcal{Y}$ can be expressed in terms of $\mathcal{X}$ by

$$\begin{aligned}\mathcal{Z} &= \mathcal{X} + \phi\mathcal{X}^2,\\ \mathcal{Y} &= \frac{1}{\rho}\left[\sigma - \mathcal{X}(1+\phi\mathcal{X}) + \delta\mathcal{X}(\phi\mathcal{X}+\delta)\right]\\ &= \left[1+\nu\mathcal{X}(1+\delta+\phi\mathcal{X})\right]\left[1+\phi\mathcal{X}(1-\delta)\right],\end{aligned}\tag{3.6}$$

and $\mathcal{X}$ satisfies the cubic equation

$$\begin{aligned}0 &= \rho\nu\phi^2(1-\delta)\mathcal{X}^3 + \left[\rho\nu\phi + \rho\nu\phi(1-\delta^2) + \phi(1-\delta)\right]\mathcal{X}^2\\ &\quad + \left[\rho\nu(1+\delta) + \rho\phi(1-\delta) + (1-\delta^2)\right]\mathcal{X} + \rho - \sigma.\end{aligned}\tag{3.7}$$

Since the first three coefficients of (3.7) are positive and the constant term is negative for $\mathcal{R}_0 > 1$, there is a unique positive root. The expressions in (3.6) show that unique positive equilibrium values for $\mathcal{Y}$ and $\mathcal{Z}$ result from the unique positive $\mathcal{X}$; hence there is a unique EE $(\mathcal{X}^*, \mathcal{Y}^*, \mathcal{Z}^*)$ for $\mathcal{R}_0 > 1$. If $\phi = 0$, then the solution of (3.7) is given explicitly as $\mathcal{X}^* = \frac{\sigma-\rho}{\rho\nu(1+\delta)+(1-\delta^2)}$, giving $\mathcal{Y}^*$ and $\mathcal{Z}^*$ from (3.6) as in the lemma statement. If $\mathcal{R}_0 < 1$, then (3.7) has no positive root, and hence there is no EE. □

The proof of the following result is standard, using the Routh–Hurwitz conditions. For details, see Appendix A.

THEOREM 3.7. *If* $\mathcal{R}_0 > 1$, *then the unique EE of system* (3.1) *is locally asymptotically stable.*

Remark 3.8. If $\mathcal{R}_0 > 1$ and $\rho \geq 1$, then every solution of (3.1) approaches either the EE or the DFE.

Proof. Consider the equivalent system $(\mathcal{X}, \mathcal{Z}, \mathcal{Y}+\mathcal{Z})$ of (3.1), namely (3.4). This equivalent system's matrix of partial derivatives has the sign pattern

$$\begin{bmatrix} - & + & 0 \\ + & - & + \\ + & * & - \end{bmatrix}$$

in the case when $\rho \geq 1$ (where $*$ is $+$ or 0), indicating an irreducible cooperative system. Then by Theorems 2.3.2 and 4.1.2 on respective pages 18 and 57 of [33], the system exhibits monotone dynamical flow and solutions must approach an equilibrium. □

The system equivalent to (3.1) has only two possible equilibria: the DFE $(0, 0, \frac{\sigma}{\rho})$ and the EE $(\mathcal{X}^*, \mathcal{Z}^*, \mathcal{Y}^* + \mathcal{Z}^*)$ from Lemma 3.6. If $\mathcal{R}_0 \geq 1$, then by Lemma 3.2, the DFE is unstable, and by Theorem 3.7, the EE is locally asymptotically stable. These facts together with numerical simulations (see section 4) indicate that the EE is GAS if $\mathcal{R}_0 > 1$ and $\rho \geq 1$, but we do not have a proof.

A Liapunov function argument is used in section 3.4 of [28] to prove a global asymptotic stability result in the case $\omega = \eta = 0$.

3.5. Effects of ν and ϕ on the EE. The nucleated polymerization model with mass action and without polymer joining has been well studied in earlier work [12, 13, 14, 28]. It is useful to understand the effects of positive values of ν and ϕ on the EE of model (3.1). By taking partial derivatives of (3.6) and (3.7) and using parameter relationships at the EE, the following signs are determined. (For selected details, see Appendix B.)

Proposition 3.9. *At the EE of* (3.1), *for* $\mathcal{R}_0 > 1$, $\frac{\partial \mathcal{X}^*}{\partial \nu} < 0$, $\frac{\partial \mathcal{Y}^*}{\partial \nu} > 0$, *and* $\frac{\partial \mathcal{Z}^*}{\partial \nu} < 0$.

Proposition 3.10. *At the EE of* (3.1), *for* $\mathcal{R}_0 > 1$, $\frac{\partial \mathcal{X}^*}{\partial \phi} < 0$ *and* $\frac{\partial \mathcal{Y}^*}{\partial \phi} > 0$.

3.6. Summary of ODE results for biological variables. The previous results are now summarized in terms of the original biological variables in system (2.5). Recall that $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ in (3.1) are respectively proportional to $x_0 U$ (the number of PrP$^{\mathrm{Sc}}$ units in the minimum lengths of the polymers), V (the number of PrP$^{\mathrm{C}}$ monomers), and W (the PrP$^{\mathrm{Sc}}$ units not accounted for within the minimum lengths of the polymers), all satisfying system (2.5). With nonnegative initial conditions, system (2.5) has a unique bounded solution in $\mathbb{R}^3_+$ defined for all $t \geq 0$. The basic reproduction number $\mathcal{R}_0$ is given by (3.2). The DFE $(x_0 U, V, W) = (0, \frac{\lambda}{\gamma}, 0)$ is globally attractive if $\gamma \geq \beta x_0 + \mu$ and $\mathcal{R}_0 \leq 1$ and is GAS if $\mathcal{R}_0 < 1$. If $\mathcal{R}_0 > 1$ and $\gamma \geq \beta x_0 + \mu$, then the unique EE $(x_0 U^*, V^*, W^*)$ demonstrated in Lemma 3.6 is locally asymptotically stable in $\mathbb{R}^3_+ \setminus [\{0\} \times \mathbb{R}_+ \times \{0\}]$. For $\mathcal{R}_0 > 1$, at the EE, $\frac{\partial (x_0 U^*)}{\partial \omega} < 0$, $\frac{\partial V^*}{\partial \omega} > 0$, $\frac{\partial W^*}{\partial \omega} < 0$, $\frac{\partial (x_0 U^*)}{\partial \eta} < 0$, and $\frac{\partial V^*}{\partial \eta} > 0$. The sign of $\frac{\partial W^*}{\partial \eta}$ is undetermined in general, since the sign of $\frac{\partial \mathcal{Z}^*}{\partial \phi}$ is unknown.

We can also interpret the results in terms of P (the number of PrP$^{\mathrm{Sc}}$ monomer units comprising the polymers) from (2.4). At the DFE, $P = 0$, and at the EE, $P^* = W^* + x_0 U^*$. By adding the second and third equations in (2.4), it follows that $\frac{\partial P^*}{\partial \omega} < 0$ and $\frac{\partial P^*}{\partial \eta} < 0$ at the EE. The ratio $\frac{P}{U}$ gives the mean polymer length. Dividing the last equation of (2.4) by U and differentiating with respect to η, it is seen that $\frac{d}{d\eta}\left(\frac{P^*}{U^*}\right) > 0$ at the EE.

3.7. A solution of the PDE system in the case of general incidence. Returning to (2.1) and (2.2), consider the case of general incidence but no joining, i.e., $\eta = 0$. The corresponding system of ODEs given in system (2.4), again with $\eta = 0$, has EE from Lemma 3.6 given by the following:

$$\begin{aligned} U^* &= \frac{\beta\gamma(\beta x_0+\mu)^2 - \beta^2\lambda\tau}{(2\beta x_0+\mu)[\omega\gamma(\beta x_0+\mu)^2+\beta\mu\tau]}, \\ V^* &= \frac{(\beta x_0+\mu)^2(\omega\lambda+\mu)}{\omega\gamma(\beta x_0+\mu)^2+\beta\mu\tau}, \\ P^* &= \frac{\gamma(\beta x_0+\mu)^2-\beta\lambda\tau}{\omega\gamma(\beta x_0+\mu)^2+\beta\mu\tau}. \end{aligned} \tag{3.8}$$

To find an equilibrium distribution of polymer lengths, set $\frac{\partial}{\partial t}u(x,t) = 0$ in (2.2). Compute the derivative with respect to x of the rest of (2.2), substituting in values of U^*, V^*, and P^* from (3.8) to obtain

$$\frac{d^2}{dx^2}[u(x)] + \frac{\beta(\beta x+\mu)}{(\beta x_0+\mu)^2}\,\frac{d}{dx}[u(x)] + \frac{3\beta^2}{(\beta x_0+\mu)^2}u(x) = 0. \tag{3.9}$$

The boundary condition $u(x_0) = 0$, first given in (2.3), can be used to find solutions to (3.9) of the form

$$u(x) = Ce^{-\frac{\beta(x-x_0)(\beta x+\beta x_0+2\mu)}{2(\beta x_0+\mu)^2}}(x-x_0)(\beta x+\beta x_0+2\mu). \tag{3.10}$$

Note that from (2.2) with $x = x_0$,

$$\frac{d}{dx}[u(x)] = 2\beta U^*\left(\frac{1+\omega P^*}{\tau V^*}\right). \tag{3.11}$$

Substitute into (3.11) values of U^*, V^*, and P^* from (3.8). Then compute the derivative of (3.10) and set it equal to (3.11) to find

$$C = \frac{\beta^3(\beta\lambda\tau - \gamma(\beta x_0+\mu)^2)}{(\beta x_0+\mu)^3(2\beta x_0+\mu)[\omega\gamma(\beta x_0+\mu)^2+\mu\beta\tau]}.$$

The equilibrium solution from (3.10), denoted by $u^*(x)$, is thus

$$u^*(x) = \left(e^{-\frac{\beta(x-x_0)(\beta x+\beta x_0+2\mu)}{2(\beta x_0+\mu)^2}}\right)\frac{\beta^3(x-x_0)(\beta x+\beta x_0+2\mu)[\beta\lambda\tau(1-1/\mathcal{R}_0)]}{(\beta x_0+\mu)^3(2\beta x_0+\mu)[\omega\gamma(\beta x_0+\mu)^2+\mu\beta\tau]}, \tag{3.12}$$

where $\mathcal{R}_0$ is given in (3.2). Note that the numerator of $u^*(x)$ requires $x > x_0$ and $\mathcal{R}_0 > 1$. The denominator of $u^*(x)$ shows that an increase in ω decreases the number of polymers of length x at steady state for all viable lengths x.

From (3.10), it can be seen that the value of x at which $u^*(x)$ achieves its maximum is independent of ω and is given by

$$x = (\sqrt{3}-1)\frac{\mu}{\beta} + \sqrt{3}\,x_0. \tag{3.13}$$

However, from (3.12), the magnitude of this maximum decreases as ω increases.

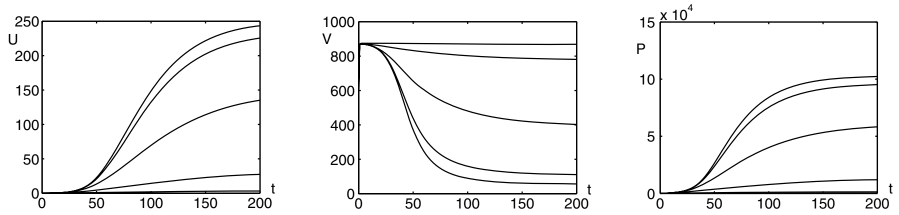


FIG. 2. *Varying $\omega = 0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$ for populations $U(t)$, $V(t)$, and $P(t)$ with $x_0 = 6$, $\lambda = 4400$, $\gamma = 5$, $\tau = 0.3$, $\mu = 0.04$, $\beta = 10^{-4}$, $\eta = 0$. Range of ω runs top to bottom on U and P graphs, bottom to top on V graph.*

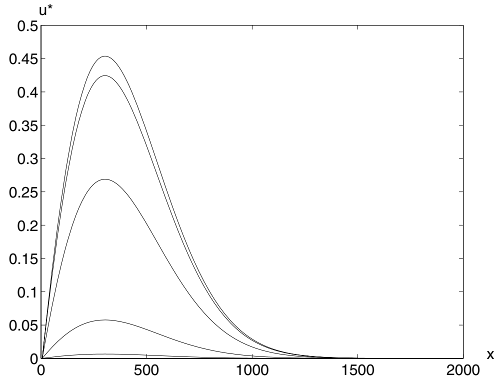


FIG. 3. *Steady-state polymer distribution $u^*(x)$ with $x_0 = 6$, $\lambda = 4400$, $\gamma = 5$, $\tau = 0.3$, $\mu = 0.04$, $\beta = 10^{-4}$, $\eta = 0$, and $\omega = 0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$. Range of ω runs from top curve to bottom on graph.*

4. Numerical simulations. To complement the previous analytical results, we present some numerical simulations. All simulations shown, unless otherwise noted, use the same parameters as in [14], namely $x_0 = 6/(\text{SAF/sq})$, $\lambda = 4400/\text{day}$, $\gamma = 5/\text{day}$, $\tau = 0.3/(\text{SAF/sq} * \text{day})$, $\mu = 0.04/\text{day}$, and $\beta = 10^{-4}(\text{SAF/sq})/\text{day}$, giving $\rho \approx 2 \times 10^5 > 1$. These parameters follow from data and observations in [3, 7, 20, 25, 30]. Some broader ranges include that $x_0 \approx 6$–30 [20], PrP$^{\text{C}}$ has a half-life of 3–6 hours [3, 7, 25] and hence $\gamma \approx 3$–$5/\text{day}$, $\mu \ll \gamma$ [20, 25], and $\lambda \approx 10^3$–$10^4/\text{day}$ [20]. The units SAF/sq are a measure of scrapie-associated fibrils counted in spleens of Compton white mice that had been given intracerebral injections of the 139A scrapie strain [30]. Note that the above parameter set gives $\mathcal{R}_0 \approx 16$. We vary values of ω and η to investigate the changes introduced by these parameters.

First consider general incidence. The effects of the parameter ω on U, V, and P, discussed in section 3.6, are shown in Figure 2. Additionally, the equilibrium solution for $u(x)$ found in (3.12) allows a comparison of steady-state polymer distributions, given differing values of ω. This appears in Figure 3, computed from (3.12). Note that, for all values of ω, the maximum value of $u(x)$ occurs at $x \approx 303$, as can be computed from (3.13).

Next consider joining. Section 3.6 describes the effects of η on the EE of system (2.4), shown numerically in Figure 4. As discussed in section 3.6, the sign of $\frac{\partial \mathcal{Z}^*}{\partial \phi}$ is undetermined; hence the sign of $\frac{\partial W^*}{\partial \eta}$ is also undetermined. It turns out that most parameter combinations, but not all, support $\frac{\partial \mathcal{Z}^*}{\partial \phi} < 0$. The opposite can occur in the

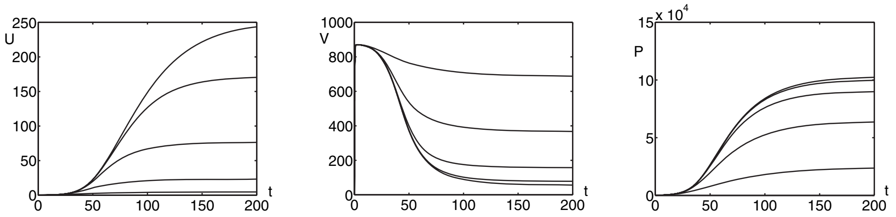


FIG. 4. *Varying $\eta = 0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ for populations $U(t)$, $V(t)$, and $P(t)$ with $x_0 = 6$, $\lambda = 4400$, $\gamma = 5$, $\tau = 0.3$, $\mu = 0.04$, $\beta = 10^{-4}$, $\omega = 0$. Range of η runs top to bottom on U and P graphs, bottom to top on V graph.*

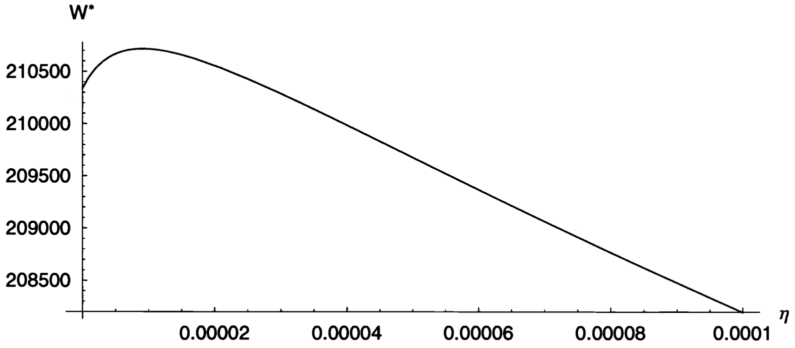


FIG. 5. *Dependence of W^* on η with $x_0 = 6$, $\lambda = 4400$, $\gamma = 5$, $\tau = 0.3$, $\mu = 0.02$, $\beta = 10^{-4}$, $\omega = 0$.*

case that $\delta \to 1$, which is possible in the case that μ is small or βx_0 is large. A brief explanation appears in Appendix C. Even so, it appears that $\frac{\partial \mathcal{Z}^*}{\partial \phi} > 0$ for only small values of ϕ. This effect is demonstrated in Figure 5 for $\frac{\partial W^*}{\partial \eta}$, which is proportional to $\frac{\partial \mathcal{Z}^*}{\partial \phi} > 0$. The parameters used in Figure 5 are the same as those listed above, but with smaller μ, namely $\mu = 0.02$, and $\omega = 0$.

Last, combine general incidence with joining. Lemma 3.2, Theorem 3.7, and Remark 3.8 together suggest that the EE of system (3.1) is GAS. Numerical simulations such as those shown in Figure 6 support this suggestion. The pair of surfaces in this figure show long-term equilibrium values of U and P, denoted U_∞ and P_∞, as both ω and η vary. For all shown pairs of ω and η values, both U_∞ and P_∞ remain positive, indicating (as a consequence of Remark 3.8) that they correspond to U^* and P^*. The shown ranges for ω and η correspond to the lower range of values used in Figures 2, 3, 4, and 5. Similar graphs generated using higher values of ω and η also result in positive values of U_∞ and P_∞. The parameter values used for x_0, λ, γ, τ, μ, and β are the same as those given at the beginning of this section.

Figures 2, 3, and 4 were computed using MATLAB, with `ode15s` for Figures 2 and 4. Figures 5 and 6 were computed using Mathematica.

5. Biological interpretation and discussion. We now discuss the analytical results and numerical simulations (for the assumed parameter values) in terms of prion biology. From sections 3.2 and 3.3, the system (2.4) always has a DFE $(U, V, P) = (0, \frac{\lambda}{\gamma}, 0)$, which attracts all solutions if $\gamma \geq \mu + \beta * x_0$ and $\mathcal{R}_0 = \frac{\beta\lambda\tau}{\gamma(\beta x_0 + \mu)^2} \leq 1$. This is the only equilibrium for $\mathcal{R}_0 < 1$, but for $\mathcal{R}_0 > 1$ there is a unique EE (U^*, V^*, P^*), with $P^* = W^* + x_0 U^*$ and $V^* \leq \lambda/\gamma$, as can be seen from (3.6). This equilibrium

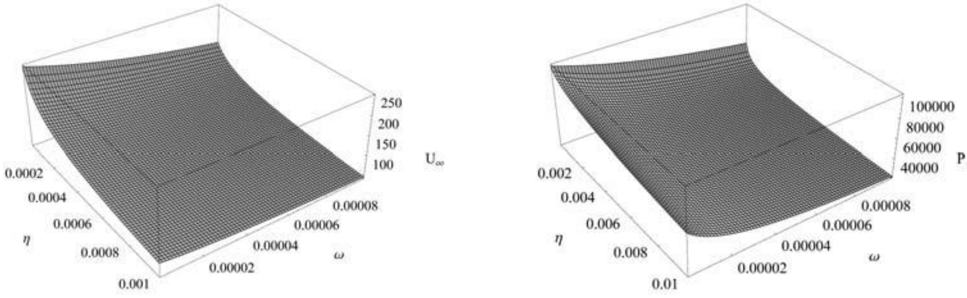


FIG. 6. *Long-term values U_∞ and P_∞ as η and ω vary. In both graphs $x_0 = 6$, $\lambda = 4400$, $\gamma = 5$, $\tau = 0.3$, $\mu = 0.04$, and $\beta = 10^{-4}$. On the U_∞ graph, $10^{-6} < \eta < 10^{-3}$, $10^{-6} < \omega < 10^{-4}$, and $50 < U_\infty < 250$. On the P_∞ graph, $10^{-6} < \eta < 10^{-2}$, $10^{-6} < \omega < 10^{-4}$, and $20,000 < P_\infty < 100,000$.*

is given explicitly by (3.8) in the case of no joining, and with joining it can be found from the solution of a cubic (see Lemma 3.6). If $\mathcal{R}_0 > 1$, then this EE is locally asymptotically stable.

From section 3.6, both increased ω and increased η cause U^* to decrease. The change related to ω indicates that as the total population of PrP$^{\text{Sc}}$ has a greater effect on general incidence, the total number of polymers at the EE decreases. For the parameters used, if $\omega \geq 10^{-3}$, then the values of (U^*, V^*, P^*) are close to those at the DFE, as seen in Figure 2. The change related to η indicates that a higher rate of polymer joining results in fewer total polymers at the EE. Increased ω and increased η cause V^* to increase. Hence the same biological changes cause both a decrease in PrP$^{\text{Sc}}$ polymers and an increase in PrP$^{\text{C}}$ at the EE. Additionally, increased ω and increased η cause the equilibrium value P^* of total PrP$^{\text{Sc}}$ to decrease. If η increases, then at the EE the mean polymer length $\frac{P^*}{U^*}$ increases, with $\ln P^*$ decreasing more slowly than $\ln U^*$.

The effects on W^* are more complicated. Increased ω, that is, increased dependence of incidence on the total PrP$^{\text{Sc}}$ population, decreases W^*. On the other hand, an increased rate of polymer joining has a variable effect on W^*. Differing parameter combinations can cause W^* to either increase or decrease with a positive change in η; see Figure 5. That noted, it is also true that most viable parameter combinations cause W^* to decrease when η increases.

Recall that the form of $\mathcal{R}_0$ given in (3.2) is the same with either mass action or our general incidence term, and with or without polymer joining. Despite the inability of ω and η to affect disease persistence, however, each of these parameters has a demonstrable effect on the steady-state values of U, V, and P. Also, increasing ω clearly decreases the number of polymers of each possible length, with the maximum for $\omega = 10^{-4}$ being about half the maximum for $\omega = 0$; see Figure 3. From data given by Rubenstein et al. [30], the onset of symptoms of scrapie can be estimated [14] to occur as $U(t)$ reaches a critical value of 130 SAF/sq. From Figures 2 and 4, the inclusion of general incidence or joining may result in U^* less than this critical value, while V^* remains closer to its DFE value. Thus, if the effects of prion diseases are caused by either an excess of PrP$^{\text{Sc}}$ or a lack of PrP$^{\text{C}}$ [24, 29], then changing the EE by increasing ω or η may be enough to delay or prevent the onset of disease symptoms.

Appendix A. Proof of Theorem 3.7. Consider a system equivalent to (3.1), namely, the system given by (3.4). Setting each of the derivatives to zero gives

$$\begin{aligned}\mathcal{Z}^* &= \mathcal{X}^* + \phi(\mathcal{X}^*)^2,\\ \frac{(\mathcal{Y}+\mathcal{Z})^* - \mathcal{Z}^*}{1+\nu(\mathcal{Z}^*+\delta\mathcal{X}^*)} &= 1 + \phi(1-\delta)\mathcal{X}^*.\end{aligned}$$

Set $q = 1 + \phi(1-\delta)\mathcal{X}^*$ and $r = 1 + \nu\mathcal{X}^*(1+\delta+\phi\mathcal{X}^*)$. Then the Jacobian of (3.4) at the unique EE is given by

$$\begin{bmatrix} -1-2\phi\mathcal{X}^* & 1 & 0 \\ 2\phi\delta\mathcal{X}^* + q - \dfrac{\nu\delta q\mathcal{X}^*}{r} & -1 - \dfrac{\mathcal{X}^*(1+\nu q)}{r} & \dfrac{\mathcal{X}^*}{r} \\ \delta^2 + 2\phi\delta\mathcal{X}^* & \rho - 1 & -\rho \end{bmatrix}.$$

The Jacobian yields the characteristic equation

$$z^3 + c_1 z^2 + c_2 z + c_3 = 0$$

with

$$\begin{aligned} c_1 &= \rho + 2 + 2\phi\mathcal{X}^* + \frac{\mathcal{X}^*(1+\nu q)}{r},\\ c_2 &= \phi\mathcal{X}^*(1-\delta) + 2\rho(1+\phi\mathcal{X}^*) + \frac{[2(1+\phi\mathcal{X}^*) + (\rho+1+\delta+2\phi\mathcal{X}^*)\nu q]\mathcal{X}^*}{r},\\ c_3 &= (1+2\phi\mathcal{X}^*)\left(\rho + \frac{\rho\nu q\mathcal{X}^*}{r} + \frac{\mathcal{X}^*}{r}\right) - \rho\left(1 + \phi(1+\delta)\mathcal{X}^* - \frac{\delta\nu q\mathcal{X}^*}{r}\right)\\ &\quad - \frac{\mathcal{X}^*}{r}(\delta^2 + 2\phi\delta\mathcal{X}^*). \end{aligned}$$

Notice that $0 < \delta < 1$ and $\mathcal{X}^* > 0$ when $\mathcal{R}_0 > 1$. Clearly $c_1 > 0$ and $c_2 > 0$. Additionally,

$$c_3 > (1+2\phi\mathcal{X}^*)\left(\rho + \frac{\mathcal{X}^*}{r}\right) - \rho(1+\phi(1+\delta)\mathcal{X}^*) - \frac{\mathcal{X}^*}{r}(\delta^2 + 2\phi\delta\mathcal{X}^*) > 0.$$

Rewriting c_3 as

$$c_3 = \frac{\mathcal{X}^*}{r}[1 + 2\phi\mathcal{X}^* - (\delta^2 + 2\phi\delta\mathcal{X}^*)] + \frac{\rho\nu q\mathcal{X}^*}{r}(1 + 2\phi\mathcal{X}^* + \delta) + \rho\phi\mathcal{X}^*(1-\delta),$$

it can be shown that

$$c_1\, c_2 > \frac{\mathcal{X}^*}{r}(2 + 2\phi\mathcal{X}^*) + \frac{\rho\nu q\mathcal{X}^*}{r}(2 + 2\phi\mathcal{X}^*) + \rho(2 + 2\phi\mathcal{X}^*) > c_3.$$

Hence the Routh–Hurwitz conditions are satisfied and the proof is complete. □

Appendix B. Selected proofs of Propositions 3.9 and 3.10. Differentiate (3.7) implicitly to give $\frac{\partial\mathcal{X}^*}{\partial\nu} < 0$, then compute from (3.6) that

$$\frac{\partial\mathcal{Y}^*}{\partial\nu} = -\frac{\partial\mathcal{X}^*}{\partial\nu}(1-\delta^2) - 2\phi\mathcal{X}^*\frac{\partial\mathcal{X}^*}{\partial\nu}(1-\delta) > 0.$$

Differentiate (3.7) implicitly to obtain

$$\frac{\partial \mathcal{X}^*}{\partial \phi} = \frac{-2\nu\rho\phi(1-\delta)(\mathcal{X}^*)^3 - [\nu\rho + \nu\rho(1-\delta^2) + 1 - \delta](\mathcal{X}^*)^2 - \rho(1-\delta)\mathcal{X}^*}{3\nu\rho\phi^2(1-\delta)(\mathcal{X}^*)^2 + 2[\nu\rho\phi + \nu\rho\phi(1-\delta^2) + \phi(1-\delta)]\mathcal{X}^* + \nu\rho(1+\delta) + \rho\phi(1-\delta) + 1 - \delta^2},$$

which shows that $\frac{\partial \mathcal{X}^*}{\partial \phi} < 0$. Next compute $\rho\frac{\partial \mathcal{Y}^*}{\partial \phi}$ and divide by $(1-\delta) > 0$ to find

$$\frac{\rho}{1-\delta}\frac{\partial \mathcal{Y}^*}{\partial \phi} = -2\phi\mathcal{X}\frac{\partial \mathcal{X}^*}{\partial \phi} - (\mathcal{X}^*)^2 - \frac{\partial \mathcal{X}^*}{\partial \phi}(1+\delta).$$

Substitute $\frac{\partial \mathcal{X}^*}{\partial \phi}$ from above, and write the full right-hand side over a common denominator. The resulting numerator can be simplified to give

$$\begin{aligned}&\nu\rho\phi^2(1-\delta)(\mathcal{X}^*)^4 + 2\nu\rho\phi(1-\delta^2)(\mathcal{X}^*)^3 + \rho\phi(1-\delta)(\mathcal{X}^*)^2\\ &\quad + \nu\rho(1-\delta^2)(1+\delta)(\mathcal{X}^*)^2 + \rho(1-\delta^2)\mathcal{X}^*.\end{aligned}$$

The numerator is seen to be strictly positive, over a positive denominator, and hence $\frac{\partial \mathcal{Y}^*}{\partial \phi} > 0$.

Appendix C. Computing values of ϕ for which $\frac{\partial \mathcal{Z}^*}{\partial \phi} > 0$. Given the EE expressions for $\mathcal{X}^*$ and $\mathcal{Z}^*$ in (3.6) and (3.7), clearly $\frac{\partial \mathcal{Z}^*}{\partial \phi} > 0$ requires that

$$\frac{\partial \mathcal{X}^*}{\partial \phi} > \frac{-(\mathcal{X}^*)^2}{2\phi\mathcal{X}^* + 1},$$

where $\frac{\partial \mathcal{X}^*}{\partial \phi}$ is given by

$$\frac{\partial \mathcal{X}^*}{\partial \phi} = \frac{-2\rho\nu\phi(1-\delta)(\mathcal{X}^*)^3 - [\rho\nu + \rho\nu(1-\delta^2) + (1-\delta)](\mathcal{X}^*)^2 - \rho(1-\delta)\mathcal{X}^*}{3\rho\nu\phi^2(1-\delta)(\mathcal{X}^*)^2 + 2[\rho\nu\phi + \rho\nu\phi(1-\delta^2) + \phi(1-\delta)]\mathcal{X}^* + \rho\nu(1+\delta) + \rho\phi(1-\delta) + 1 - \delta^2}.$$

Letting $\delta \approx 1$,

$$\frac{\partial \mathcal{X}^*}{\partial \phi} \approx -\frac{\rho\nu(\mathcal{X}^*)^2}{2\rho\nu\phi\mathcal{X}^* + 2\rho\nu} > -\frac{\rho\nu(\mathcal{X}^*)^2}{2\rho\nu\phi\mathcal{X}^* + \rho\nu} = -\frac{(\mathcal{X}^*)^2}{2\phi\mathcal{X}^* + 1}.$$

Hence for δ near 1, there are likely to be ranges of ϕ values for which $\frac{\partial \mathcal{Z}^*}{\partial \phi} > 0$.

REFERENCES

[1] T. Alper, D. A. Haig, and M. C. Clarke, *The exceptionally small size of the scrapie agent*, Biochem. Biophys. Res. Comm., 22 (1966), pp. 278–284.

[2] I. V. Baskakov, G. Legname, M. A. Baldwin, S. B. Prusiner, and F. E. Cohen, *Pathway complexity of prion protein assembly into amyloid*, J. Biol. Chem., 277 (2002), pp. 21140–21148.

[3] D. R. Borchelt, M. Scott, A. Taraboulos, N. Stahl, and S. B. Prusiner, *Scrapie and cellular prion proteins differ in their kinetics of synthesis and topology in cultured cells*, J. Cell Biol., 110 (1990), pp. 743–752.

[4] S. Brandner, S. Isenmann, A. Raeber, M. Fischer, A. Sailer, Y. Kobayashi, S. Marino, C. Weissmann, and A. Aguzzi, *Normal host prion protein necessary for scrapie-induced neurotoxicity*, Nature, 379 (1996), pp. 339–343.

[5] P. Brown, P. P. Liberski, A. Wolff, and D. C. Gajdusek, *Resistance of scrapie infectivity to steam autoclaving after formaldehyde fixation and limited survival after ashing at* 360°*: Practical and theoretical implications*, J. Infect. Diseases, 161 (1990), pp. 467–472.

[6] C. Castillo-Chavez and H. R. Thieme, *Asymptotically autonomous epidemic models*, in Mathematical Population Dynamics: Analysis of Heterogeneity, I. Theory of Epidemics, O. Arino, D. Axelrod, M. Kimmel, and M. Langlais, eds., Wuerz, Winnipeg, Canada, 1995, pp. 33–50.

[7] B. Caughey, R. E. Race, D. Ernst, M. J. Buchmeier, and B. Chesebro, *Prion protein biosynthesis in scrapie-infected and uninfected neuroblastoma cells*, J. Virol., 63 (1989), pp. 175–181.

[8] J. Chin, ed., *Control of Communicable Diseases Manual*, 17th ed., American Public Health Association, Washington, DC, 2000.

[9] Chronic Wasting Disease Alliance, project website at http://www.cwd-info.org/index.php, May 17, 2006.

[10] J. H. Come, P. E. Fraser, and P. T. Lansbury, Jr., *A kinetic model for amyloid formation in the prion diseases: Importance of seeding*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 5959–5963.

[11] M. Eigen, *Prionics or the kinetic basis of prion diseases*, Biophys. Chem., 63 (1996), pp. 11–18.

[12] H. Engler, J. Prüss, and G. F. Webb, *Analysis of a Model for the Dynamics of Prions* II, J. Math. Anal. Appl., 324 (2006), pp. 98–117.

[13] M. L. Greer, *A Population Model of Prion Dynamics*, Ph.D. Thesis, Department of Mathematics, Vanderbilt University, Nashville, TN, 2002.

[14] M. L. Greer, L. Pujo-Menjouet, and G. F. Webb, *A Mathematical analysis of the dynamics of prion proliferation*, J. Theoret. Biol., 242 (2006), pp. 598–606.

[15] J. S. Griffith, *Self-replication and scrapie*, Nature, 215 (1967), pp. 1043–1044.

[16] M. Horiuchi and B. Caughey, *Prion protein interconversions and the transmissible spongiform encephalopathies*, Structure, 7 (1999), pp. R231–R240.

[17] C. J. Johnson, K. E. Phillips, P. T. Schramm, D. McKenzie, J. M. Aiken, and J. A. Pedersen, *Prions adhere to soil minerals and remain infectious*, Public Library of Science (PLoS) Pathogens, 2 (2006), pp. 296–302.

[18] P. T. Lansbury and B. Caughey, *The chemistry of scrapie infection: Implications of the 'ice 9' metaphor*, Proc. Natl. Acad. Sci. USA, 92 (1995), pp. 1–5.

[19] J. P. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 25, SIAM, Philadelphia, 1976.

[20] J. Masel, V. A. A. Jansen, and M. S. Nowak, *Quantifying the kinetic parameters of prion replication*, Biophys. Chem., 77 (1999), pp. 139–152.

[21] V. Novitskaya, O. V. Bocharova, I. Bronstein, and I. V. Baskakov, *Amyloid fibrils of mammalian prion protein are highly toxic to cultured cells and primary neurons*, J. Biol. Chem., 281 (2006), pp. 13828–13836.

[22] M. A. Nowak, D. C. Krakauer, A. Klug, and R. M. May, *Prion infection dynamics*, Integr. Biol., 1 (1998), pp. 3–15.

[23] Office International des Epizooties (World Organization for Animal Health), available online at http://www.oie.int/eng/info/en_esb.html, May 17, 2006.

[24] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, F. E. Cohen, and S. B. Prusiner, *Conversion of α-helices into β-sheets features in the formation of the scrapie prion proteins*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 10962–10966.

[25] R. J. H. Payne and D. C. Krakauer, *The paradoxical dynamics of prion disease latency*, J. Theoret. Biol., 191 (1998), pp. 345–352.

[26] N. Pöschel, V. Brilliantov, and C. Frommel, *Kinetics of prion growth*, Biophys. J., 85 (2003), pp. 3460–3474.

[27] S. B. Prusiner, *Prions*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 13363–13383.

[28] J. Prüss, L. Pujo-Menjouet, G. F. Webb, and R. Zacher, *Analysis of a model for the dynamics of prions*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 215–225.

[29] X. Roucou, M. Gains, and A. C. Leblanc, *Neuroprotective functions of prion protein*, J. Neurosci. Res., 75 (2004), pp. 153–161.

[30] R. Rubenstein, P. A. Merz, R. J. Kascsak, C. L. Scalici, M. C. Papini, R. I. Carp, and R. H. Kimberlin, *Scrapie-infected spleens: Analysis of infectivity, scrapie-associated fibrils, and protease-resistant proteins*, J. Infect. Diseases, 164 (1991), pp. 29–35.

[31] J. R. Silveira, G. J. Raymond, A. G. Hughson, R. E. Race, V. L. Sim, S. F. Hayes, and B. Caughey, *The most infectious prion protein particles*, Nature, 437 (2005), pp. 257–261.

[32] G. Simonett and C. Walker, *On the solvability of a mathematical model for prion proliferation*, J. Math. Anal. Appl., 324 (2006), pp. 580–603.

[33] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Mathematical Surveys and Monographs 41, American Mathematical Society, Providence, RI, 1995.

[34] H. L. Smith and P. Waltman, *The Theory of the Chemostat: Dynamics of Microbial Competition*, Cambridge University Press, Cambridge, UK, 1995.

[35] H. R. Thieme, *Mathematics in Population Biology*, Princeton Series in Theoret. Comput. Biol., Princeton University Press, Princeton, NJ, 2003.

[36] UK Creutzfeldt-Jakob Disease Surveillance Unit, available online at http://www.cjd.ed.ac.uk/figures.html, May 17, 2006.

[37] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 171–198

A HIERARCHY OF DIFFUSIVE HIGHER-ORDER MOMENT EQUATIONS FOR SEMICONDUCTORS*

ANSGAR JÜNGEL[†], STEFAN KRAUSE[‡], and PAOLA PIETRA[§]

Abstract. A hierarchy of diffusive partial differential equations is derived by a moment method and a Chapman–Enskog expansion from the semiconductor Boltzmann equation assuming dominant collisions. The moment equations are closed by employing the entropy maximization principle of Levermore. The new hierarchy contains the well-known drift-diffusion model, the energy-transport equations, and the six-moments model of Grasser et al. It is shown that the diffusive models are of parabolic type. Two different formulations of the models are derived: a drift-diffusion formulation, allowing for a numerical decoupling, and a symmetric formulation in generalized dual-entropy variables, inspired by nonequilibrium thermodynamics. An entropy inequality (or H-theorem) follows from the latter formulation.

Key words. semiconductor Boltzmann equation, moment method, Chapman–Enskog expansion, entropy maximization, energy-transport model, higher-order moments

AMS subject classifications. 35Q35, 76P05, 82C35, 82D37

DOI. 10.1137/070683313

1. Introduction.

The semiconductor Boltzmann equation is of fundamental importance for the modeling of classical transport of charged carriers in solids. Its solution is the microscopic distribution function $f(x, p, t)$ depending on the spatial variable x, the (crystal) momentum p, and the time t. Macroscopic quantities, such as the particle density, current density, and energy density, can be computed from certain integrals over the momentum space, which are called moments. Since the numerical solution of the Boltzmann equation, by direct or Monte Carlo methods, is extremely time consuming and not suitable to simulate real problems in semiconductor production mode, approximate models have been derived, consisting of evolution equations for a certain number of moments of the distribution function.

The idea of the moment method is to multiply the Boltzmann equation by certain weight functions depending only on the momentum variable and to integrate over the momentum space. This leads (for a finite number of weight functions) to the so-called moment equations which are generally not closed; i.e., there are more moments than equations. This is called the closure problem. In order to obtain a closed set of equations, additional information is needed. Here we use a diffusion scaling and follow the approach of Müller and Ruggeri [45] or Levermore [41], who closed the

*Received by the editors February 21, 2007; accepted for publication (in revised form) July 19, 2007; published electronically October 17, 2007. This research has been partially supported by the bilateral DAAD-Vigoni Program. This research is part of the ESF program "Global and geometrical aspects of nonlinear partial differential equations (GLOBAL)."

http://www.siam.org/journals/siap/68-1/68331.html

[†]Institut für Analysis und Scientific Computing, Technische Universität Wien, Wiedner Hauptstr. 8-10, 1040 Wien, Austria (juengel@anum.tuwien.ac.at). The first author acknowledges partial support from the Wissenschaftskolleg "Differential Equations," funded by the Fonds zur Förderung der wissenschaftlichen Forschung, and from the Deutsche Forschungsgemeinschaft, grants JU359/5 and JU359/7.

[‡]Institut für Mathematik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany (krause@mathematik.uni-mainz.de).

[§]Istituto di Matematica Applicata e Tecnologie Informatiche, C.N.R., Via Ferrata 1, 27100 Pavia, Italy (pietra@dragon.ian.pv.cnr.it). The third author acknowledges partial support from the Progetto Intergruppo INDAM "Non linear waves and applications to compressible and incompressible fluids and to charged particle transport."

set of equations (essentially) by taking that distribution function in the definition of the moments, which maximizes the kinetic entropy under the constraints of given moments. This approach was first used in [18]. For a more recent reference, see [12]. In the context of semiconductor problems, entropy maximization has been introduced in [3] (see also [2] for a complete list of references). We derive for the first time diffusive moment models of arbitrary order and for collision operators under abstract hypotheses.

Depending on the number of moments, one obtains a hierarchy of macroscopic equations. The lowest-order model is the standard drift-diffusion model, consisting of the mass conservation equation and a constitutive equation for the current density [44]. This model is often used in device simulations at an industrial level, but it cannot cope with hot-electron or high-field phenomena, occurring in modern ultrasmall devices. Hence, higher-order moments of the distribution function need to be included leading to hydrodynamic or diffusive systems of equations.

First we review the hydrodynamic-type models which are mathematically hyperbolic conservation laws [41]. These models are derived from the Boltzmann equation in the hydrodynamic scaling. As a closure condition, an expansion of the distribution function around a heated Maxwellian using Hermite polynomials [24, 46] or using Grad's expansion [42] has been employed, which gives the so-called hydrodynamic equations [8], consisting of conservation laws for mass, momentum, and energy. The equations may also be closed using the entropy maximization principle. When 13 moments are taken into account, so-called extended hydrodynamic models have been derived [1, 4]. Hydrodynamic models of arbitrary order have been obtained in [49, 53, 54]. Finally, we mention that recently this approach has been generalized to (extended) quantum hydrodynamic models, which are obtained starting from the Wigner equation [17, 37].

Performing the diffusion limit in the Boltzmann equation, combined with the moment method, leads to diffusion-type moment equations. With the moments 1 and $\varepsilon(p)$, where $\varepsilon(p)$ is the carrier kinetic energy, energy-transport models [52] can be derived [6, 7]. These models consist of conservation laws of mass and energy and constitutive relations for particle and energy fluxes. They have been widely studied in the engineering as well as in the mathematical literature (see, e.g., [5, 11, 30, 43, 48, 55] for some engineering and [6, 13, 15, 22, 31, 32, 34] for some mathematical references). Energy-transport equations allow for the modeling of hot-electron effects. However, for ultrasmall devices, the numerical results are not sufficiently accurate compared to Monte Carlo simulations of the Boltzmann equation.

Improved accuracy has been obtained by including further moments of the distribution function leading, for instance, to the six-moments model of Grasser et al. [27] (also see [51]). The six-moments model consists of conservation laws for mass, energy, and the so-called kurtosis and constitutive equations for the corresponding three fluxes. Compared to the extended hydrodynamic models, the advantage of this model is that it constitutes a system of parabolic equations instead of hyperbolic ones, which simplifies the numerical discretization and solution considerably. Up to now, the employed closure in the literature is only heuristic, and the determination of the flux relations is based on approximations [30]. Our approach does not need any approximation and works for general collision operators (under some conditions) and general nonparabolic band structures.

More precisely, we derive, under suitable assumptions (see (H1)–(H4) below), diffusive higher-order moment models of the form

$$\partial_t m_i + \mathrm{div} J_i - i J_{i-1} \cdot \nabla V = W_i, \quad i = 0, \ldots, N,$$

where m_i are the moments (m_0 being the particle density and m_1 the energy density), J_i are the fluxes, V is the electric potential, and W_i are the averaged inelastic scattering terms (with $W_0 = 0$). The fluxes are given by

$$J_i = -\sum_{j=0}^{N} \big(D_{ij} \nabla \lambda_j + j D_{i,j-1} \nabla V \lambda_j \big),$$

where D_{ij} are the diffusion coefficients, coming from the dominant scattering processes, and λ_i are the Lagrange multipliers, coming from the constrained entropy maximization problem. The moments m_i depend nonlinearly on the Lagrange multipliers λ_j. Besides our derivation, the main results of this paper are as follows:

- The diffusion matrix (D_{ij}) is symmetric and positive definite under some topological assumptions on the semiconductor band structure, and the dependence of the moments m_i on λ_j is monotone in the sense of operators. Thus, the evolution problem is of parabolic type.
- The flux equations can be written equivalently in the drift-diffusion form

  $$J_i = -\nabla d_i - F_i(d) d_i \nabla V, \quad i = 0, \ldots, N,$$

  where $d_i = D_{i0}$ and $F_i(d)$ are nonlinear functions of $d = (d_0, \ldots, d_N)$ (see section 4.1 for details). This formulation allows for a numerical decoupling and the use of local Slotboom variables for designing a discretization scheme (see [15] and Remark 4.2 below).
- The convective parts including the electric field $-\nabla V$ can be eliminated by introducing generalized dual-entropy variables $\nu = (\nu_0, \ldots, \nu_N)$, depending on the Lagrange multipliers and the electric potential, such that

  $$\partial_t \rho_i(\nu) + \mathrm{div} F_i = g_i, \quad F_i = -\sum_{j=0}^{N} C_{ij} \nabla \nu_j,$$

  where ρ_i depends on ν, g_i depends on W_j and $\partial_t V$, and the new diffusion matrix (C_{ij}) is symmetric and positive definite (see section 4.2 for details). This formulation is useful for the numerical discretization of the equations employing standard (mixed) finite elements [23]. Moreover, it extends the dual-entropy notion known in nonequilibrium thermodynamics [19, 40].
- We are able to recover many well known diffusion models, such as the drift-diffusion, energy-transport, and six-moments models of Grasser et al. Compared to [29], no approximation of the highest-order moment is needed.

The originality of this paper consists in the facts (i) that we present for the first time a complete hierarchy of diffusion moment models for general collision operators, (ii) that we present a unifying approach of the derivation of these models, and (iii) that the derived models have very pleasant features useful for the mathematical analysis and the numerical discretization of the equations.

The paper is organized as follows. In section 2 we state our assumptions on the band structure and the collision operator, and we derive the model equations by a Chapman–Enskog expansion. Furthermore, some properties and several examples of the diffusion matrix are given. In section 3 we show that the drift-diffusion, energy-transport, and six-moments models can be recovered from the general theory. Section 4 is devoted to the drift-diffusion and dual-entropy formulation. We conclude in section 5. Finally, in the appendix some technical results are proved.

2. Derivation of the model equations. Let $B \subset \mathbb{R}^3$ be the first Brillouin zone of the semiconductor crystal under consideration. The set B is symmetric with respect to the origin; hence, we can identify it with the three-dimensional torus. We assume throughout this paper that all variables and functions are scaled. The evolution of the charged particles in the semiconductor is described by a distribution function $f(x,p,t) \geq 0$ depending on time $t > 0$ and space-crystal momentum variables $(x,p) \in \Omega \times B$, where $\Omega \subset \mathbb{R}^3$ is the semiconductor domain. The distribution function $f = f_\alpha$ is assumed to satisfy the (dimensionless) semiconductor Boltzmann equation in diffusion scaling:

$$\alpha^2 \partial_t f_\alpha + \alpha\big(u \cdot \nabla_x f_\alpha + \nabla_x V \cdot \nabla_p f_\alpha\big) = Q(f_\alpha); \tag{2.1}$$

i.e., we change the space and time scale according to $x \to x/\alpha$ and $t \to t/\alpha^2$, where the Knudsen number α is the ratio of the (optical) phonon energy and the typical kinetic energy of an electron (see [6] for details of the scaling). The Knudsen number is assumed to be small compared to one (like in [6]). The group velocity $u = u(p)$ is defined by $u = \nabla_p \varepsilon(p)$, where $\varepsilon(p)$ is the kinetic carrier energy given by the band structure of the semiconductor crystal. The function $V = V(x,t)$ denotes the electric potential which is assumed to be given or to be determined from the Poisson equation

$$\lambda^2 \Delta V = \int_B f dp - C(x),$$

where $\lambda > 0$ is the (scaled) Debye length and $C(x)$ the doping profile, modeling fixed charged background ions in the semiconductor crystal.

Below, we will perform the (formal) asymptotic limit $\alpha \to 0$. This limit avoids any assumption on the distribution function (unlike in [25]), but, on the other hand, we need some hypotheses on the collision operator. More specifically, we assume that the collision operator can be decomposed into two parts: a dominant part and a small part

$$Q(f) = Q_1(f) + \alpha^2 Q_2(f).$$

This decomposition has been justified in [6, 20], for instance. We will suppose (see section 2.2) that the kernel of Q_1 consists of generalized Maxwellians introduced in section 2.1 and that the moments of $Q_1(f)$ vanish. These assumptions are well known in this context, and they are necessary to perform the diffusion limit $\alpha \to 0$.

In order to specify our assumptions on the collision operator, we need the so-called generalized Maxwellian introduced in the following subsection.

2.1. Entropy maximization. We define the (scaled) relative entropy for $f(x, p, t)$ by

$$H(f)(x,t) = -\int_B f(\log f - 1 + \varepsilon(p))dp.$$

Here and in the following, we consider only scaled quantities. The generalized Maxwellian is defined as the maximizer of a certain constrained extremal problem. In order to define this problem, let scalar *weight functions* $\kappa(p) = (\kappa_0(p), \ldots, \kappa_N(p))$ and *moments* $m(x,t) = (m_0(x,t), \ldots, m_N(x,t))$ be given. In [54], also vector-valued weight functions are considered. We impose the following assumptions on κ_i and ε.

(H1) Let $N \geq 1$. The weight functions $\kappa_i(p)$ $(i = 0, \ldots, N)$ and the kinetic energy $\varepsilon(p)$ are smooth and even in p. Moreover, $\kappa_0 = 1$ and $\kappa_1 = \varepsilon$.

The case $N = 0$ is treated in section 3.

Example 2.1. Examples for the weight functions are

$$\kappa^{(1)} = (1, \varepsilon, \varepsilon^2, \varepsilon^3 \dots), \quad \kappa^{(2)} = (1, \varepsilon, |u|^2, \varepsilon|u|^2, |u|^4, \varepsilon|u|^4, \dots). \tag{2.2}$$

The kinetic energy may be given, for instance, in the parabolic band approximation, by $\varepsilon(p) = \frac{1}{2}|p|^2$. Clearly, in this case $\kappa^{(1)}$ and $\kappa^{(2)}$ coincide (up to multiplicative factors). A more refined model is the Kane dispersion relation, which takes into account the nonparabolicity at higher energies, $\varepsilon(1 + \delta\varepsilon) = \frac{1}{2}|p|^2$, where $\delta > 0$ is the nonparabolicity parameter. In terms of ε, we have

$$\varepsilon(p) = \frac{|p|^2}{1 + \sqrt{1 + 2\delta|p|^2}} = \frac{1}{2\delta}\big(\sqrt{1 + 2\delta|p|^2} - 1\big). \tag{2.3}$$

If $\delta = 0$, we recover the parabolic band approximation. The above examples for $\kappa^{(i)}$ and ε satisfy (H1).

We recall that, instead of Kane's dispersion relation, also the approximation $a\varepsilon(p)^b = |p|^2/2$ has been suggested, where the parameters a and b are fitted for different energy ranges [9] (see the discussion in [30, sect. IV]).

We set $\langle g \rangle = \int_B g(p)dp$ for a function $g(p)$, and we call the expression $\langle \kappa_i f \rangle$ the *ith moment* of f. Then we consider the constrained maximization problem

$$H(f^*) = \max\Big\{H(f) : \langle \kappa f(x, \cdot, t) \rangle = m(x, t) \text{ for } x \in \Omega,\ t > 0\Big\}. \tag{2.4}$$

The solution of this problem, if it exists, is given by

$$f^*(x, p, t) = \exp\big(\widetilde{\lambda}(x, t) \cdot \kappa(p) - \varepsilon(p)\big),$$

where $\widetilde{\lambda} = (\widetilde{\lambda}_0, \dots, \widetilde{\lambda}_N)$ are the Lagrange multipliers. Defining $\lambda_1 = \widetilde{\lambda}_1 - 1$ and $\lambda_i = \widetilde{\lambda}_i$ for all $i \neq 1$, we have the more compact formulation

$$f^*(x, p, t) = e^{\lambda(x,t) \cdot \kappa(p)}.$$

Remark 2.2. We notice that the mathematical solution of (2.4) is quite delicate. In [33], it has been shown that (2.4) can be uniquely solved whenever the multipliers $\widetilde{\lambda} = \widetilde{\lambda}(m)$ can be found. However, there are situations for which problem (2.4) has no solution. This is the case if the momentum space is unbounded and the polynomial weight functions have superquadratic growth at infinity [21, 35]. When the constraint of the highest degree is relaxed (as an inequality instead of an equality), the constrained maximization problem is always uniquely solvable [50]. In particular, the maximization problem can be uniquely solved if one of the following conditions holds:

1. General band structure: B is a bounded set and $\kappa = (1, \varepsilon, \varepsilon^2, \dots)$.
2. Kane's nonparabolic band approximation: $B = \mathbb{R}^3$ and $\kappa = (1, \varepsilon, \varepsilon^2)$, where ε is given by (2.3). Notice that $\varepsilon(p)$ grows linearly with p at infinity such that $\kappa_i(p)$ is at most quadratic.
3. Kane's nonparabolic band approximation: $B = \mathbb{R}^3$ and $\kappa = (1, \varepsilon, |u|^2, \varepsilon|u|^2, |u|^4, \varepsilon|u|^4, \dots)$, where ε is given by (2.3) [38]. Notice that the velocity $u = \nabla_p \varepsilon$ is bounded, and, therefore, $\kappa_i(p)$ is at most quadratic.
4. Parabolic band approximation: $B = \mathbb{R}^3$ and $\kappa = (1, |p|^2/2)$.

Given a function $f(x,p,t)$ with moments $m_i = \langle \kappa_i f \rangle = \int_B \kappa_i f dp$, we call the maximizer of (2.4) the *generalized Maxwellian* with respect to f, $f^* = M_f$. In view of the above comments, there are Lagrange multipliers λ_i such that

$$M_f = e^{\lambda \cdot \kappa}. \tag{2.5}$$

By definition, M_f and f have the same moments; i.e., $\langle \kappa_i M_f \rangle = \langle \kappa_i f \rangle = m_i$.

Below, we employ M_f to close the moment equations. This closure implicitly assumes nondegenerate Boltzmann statistics. For degenerate Fermi–Dirac statistics in the context of the energy-transport model, we refer to [6, 7]. Furthermore, it has been found that in certain semiconductor devices a mixture of hot and cold electrons exists, and a superposition of two (Maxwellian-type) distribution functions has been proposed as a closure [28].

Notice that generally the integrals relating the Lagrange multipliers and the moments cannot be solved analytically, so a numerical approach becomes necessary. However, in the case of the parabolic band energy-transport model (see Example 3.4), the integrals can be computed analytically. Moreover, for the fourth-order moment model (see Example 4.4), we show below that the function $\lambda \mapsto m$ is invertible. We also mention the approach of [4] where the exponentials in the integrals are expanded around the thermal equilibrium.

2.2. Assumptions on the collision operators. With the above definition of the generalized Maxwellian, we can state the following hypotheses on the collision operators.

(H2) For all functions $f(p)$ and all $i = 0, \ldots, N$, $\langle \kappa_i Q_1(f) \rangle = 0$. Furthermore, the null space $N(Q_1)$ of Q_1 consists of generalized Maxwellians, $N(Q_1) = \{f : f = M_f\}$.

(H3) For all functions $f(p)$, it holds that $\langle Q_2(f) \rangle = 0$.

These hypotheses express the collisional invariants. For instance, for elastic collisions, since $\kappa_0 = 1$ and $\kappa_1 = \varepsilon$ by (H1), we have mass and energy conservation:

$$\langle Q_1(f) \rangle = 0, \quad \langle \varepsilon Q_1(f) \rangle = 0.$$

Additionally, we suppose for Q_1 conservation properties for all moments with respect to the chosen weight functions. This assumption is rather strong; however, it is satisfied, for instance, for relaxation-time operators (see Example 2.3). Hypothesis (H3) simply expresses mass conservation for the collision operator Q_2, which is physically reasonable.

In [6], based on [20], the energy-transport model is derived by assuming that Q_1 represents elastic scattering and Q_2 includes inelastic and electron-electron scattering terms. There may be two criticisms: First, the elastic scattering is assumed to be of order one, but it can be seen from physics that elastic scattering gives only a small contribution to the total scattering rate. Second, there are no first-order scattering terms. For a derivation of the energy-transport model including first-order collision terms, we refer to [16]. In this paper, we follow a more formal approach: We derive diffusive models under the above assumptions on the collision operators which may not need to be specified; only some properties are assumed. A refinement of this argument may be the subject of future work. In the following example, we present some simple collision operators satisfying the above hypotheses.

Example 2.3. (i) Consider the relaxation-time operator

$$Q_1(f) = \frac{1}{\tau}(M_f - f), \tag{2.6}$$

where $\tau > 0$ is the (possibly space- and time-dependent) relaxation time. This collision operator satisfies $\langle \kappa_i Q_1(f) \rangle = 0$ for all f (since f and M_f have the same moments), and its null space consists of the functions $f = M_f$. Thus, Q_1 satisfies (H2).

(ii) Let $N = 1$ and $\kappa = (1, \varepsilon)$, and define the collision operator

$$Q_1(f) = Q_{\mathrm{imp}}(f) + Q_{\mathrm{ee}}(f)$$

as the sum of the impurity scattering operator Q_{imp} and the electron-electron binary collision operator Q_{ee},

$$\begin{aligned} Q_{\mathrm{imp}}(f)(p) &= \int_B \phi_{\mathrm{imp}}(p,p')\delta(\varepsilon' - \varepsilon)(f' - f)dp', \\ Q_{\mathrm{ee}}(f)(p) &= \int_B \phi_{\mathrm{ee}}(p,p',p_1,p_1')\delta(\varepsilon' + \varepsilon_1' - \varepsilon - \varepsilon_1)\delta_p(p' + p_1' - p - p_1) \\ &\quad \times (f'f_1' - ff_1)dp_1 dp' dp_1', \end{aligned}$$

where ϕ_{imp}, $\phi_{\mathrm{ee}} > 0$ are transition rates, δ_p is the periodized delta distribution, and $f' = f(p')$, $f_1 = f(p_1)$, $f_1' = f(p_1')$ (see [7]). It has been shown in [7] that $\langle \kappa_i Q_1(f) \rangle = 0$ and that the kernel of Q_1 consists of the functions $M_f = e^{\lambda_0 + \lambda_1 \varepsilon}$; i.e., Q_1 satisfies (H2) for $N = 1$.

(iii) Inelastic scattering may come from phonon collisions modeled by, for instance,

$$Q_{\mathrm{ph}}(f)(p) = \int_B \big(s_{\mathrm{ph}}(p,p')f' - s_{\mathrm{ph}}(p',p)f\big)dp',$$

where $s_{\mathrm{ph}}(p,p') = \phi_{\mathrm{ph}}(p,p')[(N_{\mathrm{ph}}+1)\delta(\varepsilon - \varepsilon' + \varepsilon_{\mathrm{ph}}) + N_{\mathrm{ph}}\delta(\varepsilon - \varepsilon' - \varepsilon_{\mathrm{ph}})]$ and $\varepsilon' = \varepsilon(p')$ [6]. The number N_{ph} is the phonon occupation number, and $\varepsilon_{\mathrm{ph}}$ is the phonon energy. An elementary computation shows that $\langle Q_{\mathrm{ph}}(f) \rangle = 0$; i.e., Q_{ph} satisfies (H3).

2.3. Chapman–Enskog expansion.

First we derive the balance equations.

PROPOSITION 2.4. *Let* (H1)–(H3) *hold, and let f_α be a solution to the Boltzmann equation* (2.1). *We assume that the formal limits $F = \lim_{\alpha \to 0} f_\alpha$ and $G = \lim_{\alpha \to 0}(f_\alpha - M_{f_\alpha})/\alpha$ exist. Then the moments $m_i = \langle \kappa_i M_F \rangle$ and the fluxes $J_i = \langle u \kappa_i G \rangle$ and $I_i = \langle \nabla_p \kappa_i G \rangle$ are solutions of*

$$\partial_t m_i + \mathrm{div} J_i - \nabla V \cdot I_i = W_i, \quad i = 0, \ldots, N, \tag{2.7}$$

where $W_i = \langle \kappa_i Q_2(F) \rangle$ are the averaged inelastic collision terms, $W_0 = 0$, and the divergence and gradient are to be taken with respect to x.

We notice that the definition of the moments is consistent with the notations in section 2.1 since $\langle \kappa_i M_F \rangle = \langle \kappa_i F \rangle$.

Proof. We multiply the Boltzmann equation (2.1) by the weight functions κ_i, integrate over the Brillouin zone B, and integrate by parts in the term involving the electric potential:

$$\alpha^2 \partial_t \langle \kappa_i f_\alpha \rangle + \alpha\big(\mathrm{div}_x \langle u \kappa_i f_\alpha \rangle - \nabla_x V \cdot \langle \nabla_p \kappa_i f_\alpha \rangle\big) = \langle \kappa_i Q_1(f_\alpha) \rangle + \alpha^2 \langle \kappa_i Q_2(f_\alpha) \rangle \tag{2.8}$$

for $i = 0, \ldots, N$. Next, we perform the following Chapman–Enskog expansion (see, e.g., [10]):

$$f_\alpha = M_{f_\alpha} + \alpha g_\alpha. \tag{2.9}$$

This equation in fact defines g_α, and, by assumption, $G = \lim_{\alpha \to 0} g_\alpha$. The generalized Maxwellian M_{f_α} is an even function in p, by hypothesis (H1), whereas $p \mapsto u(p)\kappa_i(p)$

and $p \mapsto \nabla_p \kappa_i(p)$ are odd functions in p. Therefore, $\langle u\kappa_i M_{f_\alpha}\rangle = 0$, $\langle \nabla_p \kappa_i M_{f_\alpha}\rangle = 0$. Then, substituting (2.9) into the moment equations (2.8), observing that the moments of $Q_1(f_\alpha)$ vanish by (H2), and dividing the resulting equation by α^2, we obtain

$$\partial_t \langle \kappa_i M_{f_\alpha}\rangle + \alpha \partial_t \langle \kappa_i g_\alpha\rangle + \mathrm{div}_x \langle u\kappa_i g_\alpha\rangle - \nabla_x V \cdot \langle \nabla_p \kappa_i g_\alpha\rangle = \langle \kappa_i Q_2(f_\alpha)\rangle.$$

Performing the formal limit $\alpha \to 0$ in this equation leads to

$$\partial_t \langle \kappa_i M_F\rangle + \mathrm{div}_x \langle u\kappa_i G\rangle - \nabla_x V \cdot \langle \nabla_p \kappa_i G\rangle = \langle \kappa_i Q_2(F)\rangle. \tag{2.10}$$

These are the balance equations (2.7). □

Remark 2.5. For $i = 0$, we have $I_0 = 0$ and $W_0 = 0$ such that the first balance equation just expresses mass conservation:

$$\partial_t m_0 + \mathrm{div} J_0 = 0. \tag{2.11}$$

Example 2.6. The integrals I_i can be expressed in terms of the fluxes J_i for special choices of the weight functions. For instance, if we choose $\kappa = (1, \varepsilon, \varepsilon^2, \dots)$ (see (2.2)), we obtain $\nabla_p \kappa_i = iu\varepsilon^{i-1}$ for $i \geq 1$ and $\nabla_p \kappa_0 = 0$, and thus $I_i = iJ_{i-1}$ for all $i \geq 0$ (for $i = 0$, we have $I_0 = 0$). In this situation the balance equations become

$$\partial_t m_i + \mathrm{div} J_i - i\nabla V \cdot J_{i-1} = W_i. \tag{2.12}$$

If we choose $\kappa = \kappa^{(2)}$ in (2.2), we cannot express I_i in terms of the integrals $J_0, \dots, J_N$ since, for instance, $\nabla_p \kappa_2 = \nabla_p |u|^2 = \varepsilon'' u$, where ε'' is the Hessian of $\varepsilon(p)$, and this cannot be written in general as a function of $|u|^{2j}$ and $\varepsilon |u|^{2j}$.

Next, we specify the flux equations J_i. For this, we need to determine G. We will see that this is equivalent to solving the operator equation $LG = H$, where $L = DQ_1(M_F)$ is the Fréchet derivative of Q_1 at $M_F = e^{\lambda \cdot \kappa} > 0$ and $H = u \cdot \nabla_x M_F + \nabla_x V \cdot \nabla_p M_F$. We introduce the Hilbert space $L^2(B)$ with the scalar product

$$(g_1, g_2)_F = \int_B g_1 g_2 M_F^{-1} dp$$

and the corresponding norm $\|\cdot\|_F$. In order to solve the equation $LG = H$, we impose the following hypothesis on the operator L.

(H4) The linear operator $L = DQ_1(M_F)$ is continuous, closed, and symmetric on $L^2(B)$, and its null space is spanned by M_F.

An example of an operator Q_1 satisfying (H4) is presented in [6, sect. 3.2].

By the Fredholm alternative, the linear, continuous, and closed operator L on the Hilbert space $L^2(B)$ satisfies the following property: The equation $LG = H$ is solvable if and only if $H \in N(L^*)^\perp$ and its solution is unique in $N(L^*)^\perp$. As L is assumed to be symmetric, $LG = H$ is solvable if and only if $H \in N(L)^\perp$ and the solution is unique in $N(L)^\perp$. Since the null space of L consists of the generalized Maxwellians, $LG = H$ is solvable if and only if $0 = (H, M_F)_F = \int_B H dp$.

Proposition 2.7. *Let* (H1)–(H4) *hold. Then the fluxes of Proposition* 2.4 *can be written as*

$$J_i = -\sum_{j=0}^{N} \big(D_{ij} \nabla \lambda_j + E_{ij} \nabla V \lambda_j \big), \quad i = 0, \dots, N, \tag{2.13}$$

where the diffusion matrices $D_{ij} \in \mathbb{R}^{3\times 3}$ and the matrices $E_{ij} \in \mathbb{R}^{3\times 3}$ are defined by

$$D_{ij} = -\langle \kappa_i u \otimes \phi_j \rangle, \quad E_{ij} = -\langle \kappa_i u \otimes \psi_j \rangle, \tag{2.14}$$

respectively, and $\phi_j = (\phi_{j1}, \phi_{j2}, \phi_{j3})$ *and* $\psi_j = (\psi_{j1}, \psi_{j2}, \psi_{j3})$ *are the (unique) solutions in* $N(L)^\perp$ *of the operator equations*

$$L\phi_{jk} = u_k \kappa_j M_F, \quad L\psi_{j\ell} = \frac{\partial \kappa_j}{\partial p_\ell} M_F, \quad j = 0, \ldots, N, \; k, \ell = 1, 2, 3. \tag{2.15}$$

Proof. Inserting the Chapman–Enskog expansion (2.9) into the Boltzmann equation (2.1), expanding formally the collision operator

$$Q_1(f_\alpha) = Q_1(M_{f_\alpha}) + \alpha D Q_1(M_{f_\alpha}) g_\alpha + O(\alpha^2),$$

and dividing the resulting equation by α, we obtain

$$\begin{aligned} &\alpha \partial_t (M_{f_\alpha} + \alpha g_\alpha) + u \cdot \nabla_x (M_{f_\alpha} + \alpha g_\alpha) + \nabla_x V \cdot \nabla_p (M_{f_\alpha} + \alpha g_\alpha) \\ &\quad = \alpha^{-1} Q_1(M_{f_\alpha}) + D Q_1(M_{f_\alpha}) g_\alpha + O(\alpha). \end{aligned}$$

By (H2), we have $Q_1(M_{f_\alpha}) = 0$. Hence, the formal limit $\alpha \to 0$ gives

$$u \cdot \nabla_x M_F + \nabla_x V \cdot \nabla_p M_F = D Q_1(M_F) G = LG. \tag{2.16}$$

Now let $j \in \{0, \ldots, N\}$ be fixed. The operator equations (2.15) are solvable in $L^2(B)$ since $u_k \kappa_j M_F$ and $(\partial \kappa_j / \partial p_\ell) M_F$ are odd functions in p, and hence, their integrals over B vanish. The unique solution G in $N(L)^\perp$ is given by

$$G = \sum_{j=0}^N \big(\phi_j \cdot \nabla_x \lambda_j + \nabla_x V \cdot \psi_j \lambda_j \big),$$

since, observing $\nabla_x M_F = \sum_j \nabla_x \lambda_j \kappa_j M_F$ and $\nabla_p M_F = \sum_j \lambda_j \nabla_p \kappa_j M_F$, we have

$$\begin{aligned} LG &= \sum_{j=0}^N \big(L\phi_j \cdot \nabla_x \lambda_j + \nabla_x V \cdot L\psi_j \lambda_j \big) = \sum_{j=0}^N \big(\kappa_j u \cdot \nabla_x \lambda_j + \nabla_x V \cdot \nabla_p \kappa_j \lambda_j \big) M_F \\ &= u \cdot \nabla_x M_F + \nabla_x V \cdot \nabla_p M_F. \end{aligned}$$

Hence, since $J_i = \langle u \kappa_i G \rangle$, we obtain (2.13). □

Example 2.8. In the case of the relaxation-time operator of Example 2.3(i), the function G can be found explicitly. Indeed, from Chapman–Enskog expansion (2.9) and Boltzmann equation (2.1), we derive

$$\begin{aligned} g_\alpha &= \frac{1}{\alpha}(f_\alpha - M_{f_\alpha}) = -\frac{\tau}{\alpha} Q_1(f_\alpha) \\ &= -\tau \alpha (\partial_t f_\alpha - Q_2(f_\alpha)) - \tau (u \cdot \nabla_x f_\alpha + \nabla_x V \cdot \nabla_p f_\alpha), \end{aligned}$$

and the formal limit $\alpha \to 0$ gives

$$G = -\tau \big(u \cdot \nabla_x M_F + \nabla_x V \cdot \nabla_p M_F \big) = -\tau \sum_{j=0}^N \big(\kappa_j u \cdot \nabla_x \lambda_j + \nabla_x V \cdot \nabla_p \kappa_j \lambda_j \big) M_F.$$

Thus, the solutions ϕ_j and ψ_j of (2.15) are

$$\phi_j = -\tau u \kappa_j M_F, \quad \psi_j = -\tau \nabla_p \kappa_j M_F. \tag{2.17}$$

Lemma 2.9. *Let $\kappa_i = \varepsilon^i$, $i = 0, \ldots, N$. Then the coefficients E_{ij} in* (2.14) *can be expressed in terms of D_{ij}:*

$$E_{ij} = jD_{i,j-1}, \quad E_{i0} = 0, \quad j = 1, \ldots, N. \tag{2.18}$$

Proof. The assumption $\kappa_i = \varepsilon^i$ gives $\nabla_p \kappa_{i+1} = (i+1)\varepsilon^i \nabla_p \varepsilon = (i+1)u\varepsilon^i$ and hence $L\psi_{i+1} = \nabla_p \kappa_{i+1} M_F = (i+1)L\phi_i$. By the unique solvability in $N(L)^\perp$, $\psi_{i+1} = (i+1)\phi_i + cM_F$ for all $i \geq 0$ and $\psi_0 = cM_F$, where c is a constant vector. Therefore,

$$E_{ij} = -\int_B \kappa_i u \otimes \psi_j dp = -j\int_B \varepsilon^i u \otimes \phi_{j-1} dp = jD_{i,j-1},$$

proving the lemma. □

2.4. Properties of the diffusion matrix. The diffusion matrix $D = (D_{ij})$ defined in (2.14) is symmetric; this expresses the Onsager principle [40].

Lemma 2.10. *The matrices $D = (D_{ij})$, $E = (E_{ij}) \in \mathbb{R}^{3(N+1)\times 3(N+1)}$ are symmetric in the sense that*

$$D_{ij}^\top = D_{ji}, \quad E_{ij}^\top = E_{ji} \quad \textit{for all } i, j = 0, \ldots, N.$$

Proof. We write $D_{ij} = (D_{ij}^{k\ell}) \in \mathbb{R}^{3\times 3}$. Since L is symmetric on $L^2(B)$, we have

$$\begin{aligned} D_{ij}^{k\ell} &= -(u_k \kappa_i M_F, \phi_{j\ell})_F = -(L\phi_{ik}, \phi_{j\ell})_F = -(\phi_{ik}, L\phi_{j\ell})_F \\ &= -(\phi_{ik}, u_\ell \kappa_j M_F)_F = D_{ji}^{\ell k}. \end{aligned}$$

The symmetry of E is proven in a similar way. □

Under additional assumptions on the derivative of the dominant collision operator and on the band structure, we can show that the diffusion matrix is positive definite.

(H5) Let the operator $-L = -DQ_1(M_F)$ be coercive on $N(L)^\perp$; i.e., there exists a constant $\mu > 0$ such that, for all $g \in N(L)^\perp$,

$$(-Lg, g)_F \geq \mu \|g\|_F^2.$$

Example 2.11. We claim that the relaxation-time operator (2.6) satisfies (H5) if the weight functions $\kappa_0, \ldots, \kappa_N$ are linearly independent. Let $g \in N(L)^\perp$. We show first that $M_g \in N(L)$. It is sufficient to prove that $M_{M_g} = M_g$. For this, let $M_g = e^{\lambda\cdot\kappa}$ and $M_{M_g} = e^{\widetilde{\lambda}\cdot\kappa}$. Since the moments of M_g and M_{M_g} coincide by construction, we have

$$\int_B \kappa(e^{\lambda\cdot\kappa} - e^{\widetilde{\lambda}\cdot\kappa})dp = 0 \quad \text{and} \quad \int_B (\lambda\cdot\kappa - \widetilde{\lambda}\cdot\kappa)(e^{\lambda\cdot\kappa} - e^{\widetilde{\lambda}\cdot\kappa})dp = 0.$$

By the strict monotonicity of $x \mapsto e^x$, the integrand vanishes, and, therefore, $(\lambda - \widetilde{\lambda})\cdot \kappa = 0$. Since $\kappa_0, \ldots, \kappa_N$ are linearly independent, $\lambda = \widetilde{\lambda}$. Hence, $M_{M_g} = M_g$, which proves that $M_g \in N(L)$. This property gives

$$(-Lg, g)_F = -\frac{1}{\tau}(M_g - g, g)_F = -\frac{1}{\tau}(M_g, g)_F + \frac{1}{\tau}\|g\|_F^2 = \frac{1}{\tau}\|g\|_F^2.$$

Lemma 2.12. *Let* (H5) *hold, and let $\{u_k\kappa_i : k = 1, 2, 3, i = 0, \ldots, N\}$ be linearly independent functions in p. Then the diffusion matrix $D = (D_{ij})$ is positive definite; i.e., for all $\xi_0, \ldots, \xi_N \in \mathbb{R}^{N+1}$, $(\xi_0, \ldots, \xi_N) \neq 0$,*

$$\sum_{i,j=0}^N \xi_i^\top D_{ij} \xi_j > 0.$$

The proof of the lemma can be found in the appendix. The diffusion matrices D_{ij} can be simplified under additional assumptions.

PROPOSITION 2.13. *Let $\kappa_i = \varepsilon^i$, $i = 0, \dots, N$ and $Q_1(f) = (M_f - f)/\tau$. Then the diffusion coefficients can be written as*

$$D_{ij} = \frac{\tau}{3} \int_B e(\tfrac{1}{2}|p|^2)^{i+j} e'(\tfrac{1}{2}|p|^2)^2 |p|^2 \exp\left(\sum_{k=0}^N \lambda_k e(\tfrac{1}{2}|p|^2)^k \right) dp\, I,$$

where $\varepsilon(p) = e(\frac{1}{2}|p|^2)$ and I is the unit matrix in $\mathbb{R}^{3\times 3}$.

Clearly, we may identify the matrix D_{ij} with its diagonal elements and obtain the $(N \times N)$ matrix $D = (D_{ij})$.

Proof. Since the collision operator Q_1 is assumed to be a relaxation-time operator, the solution of the operator equation (2.15) is equal to $\phi_j = -\tau u \kappa_j M_F = -\tau \varepsilon^j \nabla_p \varepsilon M_F$ (see (2.17)). Thus, by definition (2.14),

$$D_{ij} = -\int_B \varepsilon^i \nabla_p \varepsilon \otimes \phi_j dp = \tau \int_B \varepsilon^{i+j} \nabla_p \varepsilon \otimes \nabla_p \varepsilon M_F dp.$$

Since $\nabla_p \varepsilon(p) = p e'(\frac{1}{2}|p|^2)$, we obtain

$$D_{ij} = \tau \int_B e(\tfrac{1}{2}|p|^2)^{i+j} e'(\tfrac{1}{2}|p|^2)^2 p \otimes p M_F dp.$$

The function $p \mapsto p \otimes p$ is odd in every off-diagonal element such that the above integral vanishes except for the diagonal elements. Since each diagonal element has the same value and $M_F = e^{\lambda \cdot \kappa}$, the expression for D_{ij} is proven. ◻

The diffusion coefficients can be further simplified under additional assumptions on the energy band structure. We consider three examples.

Example 2.14 (monotone energy band). Let the assumption of Proposition 2.13 hold. We suppose additionally that $e(\frac{1}{2}|p|^2)$ is strictly monotone in $|p|$ and that $e(0) = 0$ and $e(\infty) = \infty$. This allows us to choose $B = \mathbb{R}^3$. Then, with spherical coordinates (ρ, θ, ϕ), for $i, j = 0, \dots, N$,

$$D_{ij} = \frac{\tau}{3} \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} e(\tfrac{1}{2}\rho^2)^{i+j} e'(\tfrac{1}{2}\rho^2)^2 \rho^4 \exp\left(\sum_{k=0}^N \lambda_k e(\tfrac{1}{2}\rho^2)^k \right) \sin\theta d\rho d\theta d\phi.$$

Now we perform the change of variables $\varepsilon = e(\frac{1}{2}\rho^2)$, setting $\gamma(\varepsilon) = \rho^2$. Then $d\rho = (\gamma'(\varepsilon)/2\sqrt{\gamma(\varepsilon)}) d\varepsilon$ such that

$$D_{ij} = \frac{8\pi\tau}{3} \int_0^{\infty} \varepsilon^{i+j} \frac{\gamma(\varepsilon)^{3/2}}{\gamma'(\varepsilon)} \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k \right) d\varepsilon. \tag{2.19}$$

In the special case $N = 1$, the same diffusion coefficients have been derived in [6, equations (3.36), (4.17)]. Notice that the above transformation allows us to simplify the expression for the moments:

$$\begin{aligned} m_i &= \int_B e(\tfrac{1}{2}|p|^2)^i \exp\left(\sum_{k=0}^N \lambda_k e(\tfrac{1}{2}|p|^2)^k\right) dp \\ &= 4\pi \int_0^\infty e(\tfrac{1}{2}\rho^2)^i \exp\left(\sum_{k=0}^N \lambda_k e(\tfrac{1}{2}|p|^2)^k\right) \rho^2 d\rho \\ &= 2\pi \int_0^\infty \varepsilon^i \sqrt{\gamma(\varepsilon)}\gamma'(\varepsilon) \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k\right) d\varepsilon, \end{aligned} \tag{2.20}$$

where $i = 0, \ldots, N$.

Example 2.15 (nonparabolic band approximation). In the case of Kane's nonparabolic band approximation (2.3), we can further simplify the integrals (2.19) and (2.20). Since $\gamma(\varepsilon) = |p|^2 = 2\varepsilon(1+\delta\varepsilon)$ and $\gamma'(\varepsilon) = 2(1+2\delta\varepsilon)$, we compute

$$D_{ij} = \frac{8\sqrt{2}\pi}{3}\tau \int_0^\infty \varepsilon^{i+j+3/2} \frac{(1+\delta\varepsilon)^{3/2}}{1+2\delta\varepsilon} \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k\right) d\varepsilon,$$

$$m_i = 4\sqrt{2}\pi \int_0^\infty \varepsilon^{i+1/2}(1+\delta\varepsilon)^{1/2}(1+2\delta\varepsilon) \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k\right) d\varepsilon, \quad i,j = 0,\ldots,N.$$

Example 2.16 (parabolic band approximation). Setting $\delta = 0$ in the formulas of Example 2.15, we obtain

$$D_{ij} = \frac{8\sqrt{2}\pi}{3}\tau \int_0^\infty \varepsilon^{i+j+3/2} \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k\right) d\varepsilon,$$

$$m_i = 4\sqrt{2}\pi \int_0^\infty \varepsilon^{i+1/2} \exp\left(\sum_{k=0}^N \lambda_k \varepsilon^k\right) d\varepsilon, \quad i,j = 0,\ldots,N.$$

3. Examples.

In this section we derive the diffusive models for $N = 0$, leading to the drift-diffusion equations, the case $N = 1$, leading to the energy-transport model, and $N = 2$, leading to a higher-order model.

3.1. Drift-diffusion equations.

We consider the case $N = 0$. Then $\kappa_0(p) = 1$, and the generalized Maxwellian reads $M_F = e^{\lambda_0 - \varepsilon(p)}$. The balance equation is given by (2.11). We need to compute the flux J_0 since, in section 2.3, the case $N = 0$ was excluded. For this, we have to solve $LG = u \cdot \nabla_x \lambda_0 M_F + \nabla_x V \cdot \nabla_p M_F = u \cdot \nabla_x(\lambda_0 - V)M_F$. Let ϕ_0 be the unique solution in $N(L)^\perp$ of $L\phi_0 = uM_F$. It is not difficult to check that $G = \nabla_x(\lambda_0 - V)\cdot\phi_0$ solves the above operator equation. This shows that

$$J_0 = \langle uG\rangle = \langle u \otimes \phi_0\rangle \nabla_x(\lambda_0 - V).$$

The flux can be written in terms of the particle density m_0. Indeed, since

$$m_0 = \int_B M_F dp = Ae^{\lambda_0}, \quad \text{where } A = \int_B e^{-\varepsilon(p)} dp > 0,$$

we obtain $\nabla_x \lambda_0 = (\nabla_x m_0)/m_0$ and, hence,

$$J_0 = -D_0(\nabla_x m_0 - m_0 \nabla_x V), \quad \text{where } D_0 = -\frac{1}{m_0}\int_B u \otimes \phi_0 dp.$$

This gives the well-known drift-diffusion equations for the particle density $n = m_0$ and the current density $J = J_0$:

$$\partial_t n + \mathrm{div} J = 0, \quad J = D_0(\nabla n - n\nabla V).$$

We specify the diffusion matrix D_0 and the relation between m_0 and λ_0 in the following example.

Example 3.1. Under the assumptions of Example 2.14, we obtain for the expressions for $D_0 = D_{00}/m_0$ and m_0:

$$D_0 = \frac{8\pi}{3}\,\frac{\tau}{m_0}\,e^{\lambda_0}\int_0^\infty \varepsilon^{3/2}\frac{\gamma(\varepsilon)^{3/2}}{\gamma'(\varepsilon)}e^{-\varepsilon}d\varepsilon,$$
$$m_0 = 2\pi\, e^{\lambda_0}\int_0^\infty \sqrt{\gamma(\varepsilon)}\gamma'(\varepsilon)e^{-\varepsilon}d\varepsilon.$$

For nonparabolic bands $\gamma(\varepsilon) = 2\varepsilon(1+\delta\varepsilon)$, this becomes

$$D_0 = \frac{8\sqrt{2}\pi}{3}\,\frac{\tau}{m_0}\,e^{\lambda_0}\int_0^\infty \varepsilon^{3/2}\frac{(1+\delta\varepsilon)^{3/2}}{1+2\delta\varepsilon}e^{-\varepsilon}d\varepsilon, \tag{3.1}$$

$$m_0 = 4\sqrt{2}\pi\, e^{\lambda_0}\int_0^\infty \varepsilon^{1/2}(1+\delta\varepsilon)^{1/2}(1+2\delta\varepsilon)e^{-\varepsilon}d\varepsilon, \tag{3.2}$$

and, for parabolic bands, the formulas simplify to

$$m_0 = 4\sqrt{2}\pi\, e^{\lambda_0}\int_0^\infty \varepsilon^{1/2}e^{-\varepsilon}d\varepsilon = 4\sqrt{2}\pi\, e^{\lambda_0}\Gamma(\tfrac{3}{2}) = (2\pi)^{3/2}e^{\lambda_0}, \tag{3.3}$$

$$D_0 = \frac{8\sqrt{2}\pi}{3}\,\frac{\tau}{m_0}\,e^{\lambda_0}\int_0^\infty \varepsilon^{3/2}e^{-\varepsilon}d\varepsilon = \frac{4\tau}{3\sqrt{\pi}}\Gamma(\tfrac{5}{2}) = \tau, \tag{3.4}$$

where Γ is the Gamma function satisfying $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(x+1) = x\Gamma(x)$. The expressions (3.3) and (3.4) coincide with the standard drift-diffusion model; see, for instance, [36, 44].

3.2. Energy-transport equations. We take $N = 1$ and $\kappa = (1, \varepsilon)$. Then $M_F = e^{\lambda_0+\lambda_1\varepsilon}$. The balance equations are, according to Proposition 2.4 and Example 2.6,

$$\partial_t m_0 + \mathrm{div} J_0 = 0, \quad \partial_t m_1 + \mathrm{div} J_1 - \nabla V\cdot J_0 = W_1. \tag{3.5}$$

The diffusion coefficients D_{ij} are, by (2.14),

$$D_{00} = -\langle u\otimes\phi_0\rangle,\ D_{01} = -\langle u\otimes\phi_1\rangle,\ D_{10} = -\langle \varepsilon u\otimes\phi_0\rangle,\ D_{11} = -\langle \varepsilon u\otimes\phi_1\rangle,$$

and the coefficients E_{ij} can be expressed in terms of D_{ij}, according to (2.18),

$$E_{00} = E_{10} = 0, \quad E_{01} = D_{00}, \quad E_{11} = D_{01}.$$

Notice that $D_{01} = D_{10}$ since $\langle u\otimes\phi_1\rangle = (L\phi_0, \phi_1)_F = (\phi_0, L\phi_1)_F = \langle \varepsilon u\otimes\phi_0\rangle$. Then the particle and energy current densities (2.13) can be written as follows:

$$J_0 = -D_{00}(\nabla\lambda_0 + \nabla V\lambda_1) - D_{01}\nabla\lambda_1, \tag{3.6}$$

$$J_1 = -D_{10}(\nabla\lambda_0 + \nabla V\lambda_1) - D_{11}\nabla\lambda_1, \tag{3.7}$$

and the moments are given by

$$m_0 = e^{\lambda_0} \int_B e^{\lambda_1 \varepsilon(p)} dp, \quad m_1 = e^{\lambda_0} \int_B \varepsilon(p) e^{\lambda_1 \varepsilon(p)} dp. \tag{3.8}$$

Equations (3.5)–(3.8) are called the energy-transport model.

Notice that, in [47], a related energy-transport model based on entropy maximization has been derived. More precisely, the model is derived through the relaxation-time limit from the hydrodynamic equations which have been found by a moment method employing the entropy maximization principle. Moreover, an assumption of small anisotropy for the Maxwellian has been used. Depending on the concrete scattering terms, the diffusion coefficients seem to be different compared to our model, but there are similar properties (such as positive definiteness of the diffusion matrix).

Example 3.2 (monotone energy band). In the situation of Example 2.14, we can make the above expressions more explicit. As we have assumed that the constrained maximization problem (2.4) is solvable, the integral expressions defining the moments have to exist. Consequently, we must have $\lambda_1 < 0$ in order to guarantee integrability of $M_F = e^{\lambda_0 + \lambda_1 \varepsilon(p)}$ in $B = \mathbb{R}^3$. Thus, we can define $T = -1/\lambda_1$, and we call $T > 0$ the particle temperature. Formulas (2.19) and (2.20) give

$$\begin{aligned} D_{ij} &= \frac{8\pi}{3} \tau e^{\lambda_0} \int_0^\infty \varepsilon^{i+j} \frac{\gamma(\varepsilon)^{3/2}}{\gamma'(\varepsilon)} e^{-\varepsilon/T} d\varepsilon, \\ m_i &= 2\pi e^{\lambda_0} \int_0^\infty \varepsilon^i \sqrt{\gamma(\varepsilon)} \gamma'(\varepsilon) e^{-\varepsilon/T} d\varepsilon, \quad i, j = 0, 1. \end{aligned}$$

Example 3.3 (nonparabolic band approximation). For nonparabolic bands according to (2.3), i.e., $\gamma(\varepsilon) = 2\varepsilon(1 + \delta\varepsilon)$, we can specify the above formulas, as in Example 2.15:

$$\begin{aligned} D_{ij} &= \frac{8\sqrt{2}\pi}{3} \tau e^{\lambda_0} \int_0^\infty \varepsilon^{i+j+3/2} \frac{(1+\delta\varepsilon)^{3/2}}{1 + 2\delta\varepsilon} e^{-\varepsilon/T} d\varepsilon, \\ m_i &= 4\sqrt{2}\pi e^{\lambda_0} \int_0^\infty \varepsilon^{i+1/2} (1+\delta\varepsilon)^{1/2} (1 + 2\delta\varepsilon) e^{-\varepsilon/T} d\varepsilon, \quad i = 0, 1. \end{aligned}$$

These expressions coincide with those in [15].

Example 3.4 (parabolic band approximation). For $\delta = 0$, the integrals of the previous example can be computed explicitly. Since

$$m_i = 4\sqrt{2}\pi e^{\lambda_0} \int_0^\infty \varepsilon^{i+1/2} e^{-\varepsilon/T} d\varepsilon = 4\sqrt{2}\pi e^{\lambda_0} T^{i+3/2} \Gamma(i + \tfrac{3}{2}), \tag{3.9}$$

we compute the moments

$$m_0 = (2\pi)^{3/2} T^{3/2} e^{\lambda_0}, \quad m_1 = \tfrac{3}{2} (2\pi)^{3/2} T^{5/2} e^{\lambda_0} = \tfrac{3}{2} m_0 T.$$

Calling $n = m_0$ the particle density, $m_1 = \frac{3}{2} nT$ can be interpreted as the electron energy with the temperature T. The diffusion coefficients become

$$D_{ij} = \frac{8\sqrt{2}\pi}{3} \tau e^{\lambda_0} \int_0^\infty \varepsilon^{i+j+3/2} e^{-\varepsilon/T} d\varepsilon = \frac{8\sqrt{2}\pi}{3} \tau e^{\lambda_0} T^{i+j+5/2} \Gamma(i + j + \tfrac{5}{2}),$$

and, computing the Gamma functions, we derive for $D = (D_{ij})$

$$D = \tau n T \begin{pmatrix} 1 & \frac{5}{2}T \\ \frac{5}{2}T & \frac{35}{4}T^2 \end{pmatrix}.$$

The relaxation time τ may be defined as the inverse of the (averaged) collision rate which generally depends on the energy. For instance, we may take

$$\tau = \tau_0 \left(\frac{\langle M_F \rangle}{\langle \varepsilon M_F \rangle} \right)^{\beta},$$

where $\tau_0 > 0$ and $\beta \in \mathbb{R}$ [52]. Then $\tau = \tau_0(m_0/m_1)^\beta = (\frac{2}{3})^\beta \tau_0 T^{-\beta}$, and the diffusion matrix can be written as

$$D = \left(\frac{2}{3}\right)^{\beta} \tau_0 m_0 T^{1-\beta} \begin{pmatrix} 1 & \frac{5}{2}T \\ \frac{5}{2}T & \frac{35}{4}T^2 \end{pmatrix}.$$

We observe that D is very similar to the matrix derived in [15] for $\beta = 1$, but the coefficients are different. The matrix of [15] can be obtained if the relaxation time depends on the microscopic kinetic energy, $\tau = \tau(\varepsilon) = \varepsilon_0/\varepsilon$ for some $\varepsilon_0 > 0$, such that

$$D_{ij} = \frac{8\sqrt{2}\pi}{3} e^{\lambda_0} \int_0^\infty \tau(\varepsilon) \varepsilon^{i+j+3/2} e^{-\varepsilon/T} d\varepsilon = \frac{8\sqrt{2}\pi\varepsilon_0}{3} e^{\lambda_0} T^{i+j+3/2} \Gamma(i+j+\tfrac{3}{2}),$$

which gives the matrix

$$D = \frac{2}{3}\varepsilon_0 n \begin{pmatrix} 1 & \frac{3}{2}T \\ \frac{3}{2}T & \frac{15}{4}T^2 \end{pmatrix}.$$

3.3. Fourth-order moment equations. Finally, we consider the case $N = 2$ and $\kappa = (1, \varepsilon, \varepsilon^2)$. The coefficients are taken from Example 2.15, which uses the hypotheses of Proposition 2.13. The balance equations are given by (2.7), which, taking into account Example 2.6, read as

$$\partial_t m_0 + \mathrm{div} J_0 = 0, \tag{3.10}$$

$$\partial_t m_1 + \mathrm{div} J_1 - \nabla V \cdot J_0 = W_1, \tag{3.11}$$

$$\partial_t m_2 + \mathrm{div} J_2 - 2\nabla V \cdot J_1 = W_2, \tag{3.12}$$

where W_i are the averaged inelastic collision terms (see Proposition 2.4), and the fluxes are given by (2.13):

$$J_i = -D_{i0}(\nabla\lambda_0 + \nabla V \lambda_1) - D_{i1}(\nabla\lambda_1 + 2\nabla V \lambda_2) - D_{i2}\nabla\lambda_2, \quad i = 0, 1, 2.$$

The diffusion coefficients are expressed as in Example 2.15 with $N = 2$. In the limiting case $\delta \to 0$ we obtain the parabolic band approximation, which allows for a more explicit formulation of the fourth-order model. Since the parabolic band approximation cannot be taken directly in the case $N = 2$ (the entropy maximization problem may be unsolvable; see Remark 2.2), we derive the model for $\delta = 0$ by taking formally the limit $\delta \to 0$ in the expressions for D_{ij} and m_i in Example 2.15. This leads to

(3.13)

$$m_i = 4\sqrt{2}\pi e^{\lambda_0} \int_0^\infty \varepsilon^{i+1/2} e^{\lambda_1\varepsilon+\lambda_2\varepsilon^2} d\varepsilon, \quad D_{ij} = \frac{8\sqrt{2}\pi}{3} \tau e^{\lambda_0} \int_0^\infty \varepsilon^{i+j+3/2} e^{\lambda_1\varepsilon+\lambda_2\varepsilon^2} d\varepsilon,$$

where $i, j = 0, 1, 2$. We argue as in Example 3.2 to conclude that $\lambda_2 < 0$ must hold. Notice that we can express the diffusion coefficients in terms of the moments:

$$D_{ij} = \frac{2\tau}{3} m_{i+j+1}. \tag{3.14}$$

The moments m_j for $j \geq 3$ are defined as above. In section 4 we discuss several reformulations of this model and compare it with higher-order models in the literature.

4. Properties of the model equations. We suppose that (H1)–(H5) hold and that the weight functions are given by $\kappa_i = \varepsilon^i$, $i = 0, \ldots, N$. Then, by (2.12), (2.13), and (2.18), the higher-order moment model can be written as

$$\partial_t m_i + \mathrm{div} J_i - iJ_{i-1} \cdot \nabla V = W_i, \quad J_i = -\sum_{j=0}^N \left(D_{ij}\nabla\lambda_j + jD_{i,j-1}\nabla V \lambda_j\right), \tag{4.1}$$

where $i = 0, \ldots, N$, $D_{i,-1} = 0$, and the moments m_i and the Lagrange multipliers λ_j are related by the formula

$$m_i = \int_B \varepsilon(p)^i \exp\left(\sum_{j=0}^N \varepsilon(p)^j \lambda_j\right) dp. \tag{4.2}$$

In this section we show that these equations can be written in two different ways, which allows us to recover some important properties of the model.

4.1. Drift-diffusion formulation. We can write the fluxes in a drift-diffusion formulation which allows a numerical decoupling of the stationary higher-order moment model.

Proposition 4.1. *Let* (H1)–(H5) *and the assumptions of Lemma* 2.12 *hold, and let $\kappa_i = \varepsilon^i$ for $i = 0, \ldots, N$. Then we can write*

$$J_i = -\nabla d_i - F_i(d) d_i \nabla V,$$

where $d_i = D_{i0}$, $d = (d_0, \ldots, d_N)^\top$, and

$$F_i(d) = \sum_{j=1}^N j \frac{D_{i,j-1}}{D_{i0}} \lambda_j, \quad i = 0, \ldots, N.$$

The Lagrange multipliers λ_j are implicitly given by the values of d_i:

$$d_i = -\langle \varepsilon^i u \otimes \phi_0 \rangle, \quad L\phi_0 = ue^{\lambda\cdot\kappa}.$$

The operator L is the linearization of the dominant collision operator; see (H4). *The mapping $d = d(\lambda)$ can be inverted since* $\det d'(\lambda) = \det D > 0$.

Proof. We claim that the first sum in the second equation in (4.1) equals ∇D_{0i}. Indeed, from

$$L(\nabla\phi_{jk}) = u_k\varepsilon^j \sum_{\ell=0}^{N} \nabla\lambda_\ell \varepsilon^\ell M_F = \sum_{\ell=0}^{N} \nabla\lambda_\ell u_k \varepsilon^{j+\ell} M_F = L\left(\sum_{\ell=0}^{N} \nabla\lambda_\ell \phi_{j+\ell,k}\right)$$

and the unique solvability in $N(L)^\perp$, we obtain the relation

$$\nabla\phi_j = \sum_{\ell=0}^{N} \nabla\lambda_\ell \phi_{j+\ell} + cM_F,$$

where c is a constant vector. Hence, by (2.14), setting $j = 0$,

$$\nabla D_{i0} = -\langle \varepsilon^i u \otimes \nabla\phi_0 \rangle = -\sum_{\ell=0}^{N} \nabla\lambda_\ell \langle \varepsilon^i u \otimes \phi_\ell \rangle = -\sum_{\ell=0}^{N} \nabla\lambda_\ell D_{i\ell}.$$

Then (4.1) becomes

$$J_i = -\nabla D_{i0} - D_{i0}\nabla V \sum_{j=0}^{N} j\, \frac{D_{i,j-1}}{D_{i0}} \lambda_j,$$

showing the first assertion.

It remains to show that the determinant of the matrix $d'(\lambda)$ is positive. Since

$$L\Big(\frac{\partial\phi_{jk}}{\partial\lambda_\ell}\Big) = u_k\varepsilon^j \frac{\partial M_F}{\partial\lambda_\ell} = u_k\varepsilon^{j+\ell} M_F = L\phi_{j+\ell,k},$$

which gives $\partial\phi_0/\partial\lambda_\ell = \phi_\ell + cM_F$ and thus

$$\frac{\partial D_{i0}}{\partial\lambda_\ell} = -\Big\langle \varepsilon^i u \otimes \frac{\partial\phi_0}{\partial\lambda_\ell} \Big\rangle = -\langle \varepsilon^i u \otimes \phi_\ell \rangle = D_{i\ell}, \tag{4.3}$$

the Jacobian of $d(\lambda)$ consists of the elements $\partial d_i/\partial\lambda_j = \partial D_{i0}/\partial\lambda_j = D_{ij}$. The matrix $D = (D_{ij})$ is positive definite (see Lemma 2.12), and we have $\det d'(\lambda) = \det D > 0$. □

Remark 4.2. The decoupling of the higher-order moment model can be done as follows. Under the assumptions of the above proposition, the stationary model reads

$$\mathrm{div} J_i = i\nabla V \cdot J_{i-1} + W_i, \quad J_i = -\nabla d_i - F_i(d) d_i \nabla V, \quad i = 0, \ldots, N.$$

We assume that V is given, and $W_i = W_i(d, V)$ may depend on d and V. We also write $J_i = J_i(d, V)$. During the iteration procedure, we may "freeze" the nonlinearities: Let $\widetilde{d}$ be given (e.g., from the previous iteration step), and consider the system

$$\mathrm{div} J_i(d, V) = i\nabla V \cdot J_{i-1}(d, V) + W_i(\widetilde{d}, V), \quad J_i(d, V) = -\nabla d_i - F_i(\widetilde{d}) d_i \nabla V.$$

This system is decoupled since each equation is a scalar elliptic differential equation for d_i. Furthermore, the linear equations can by "symmetrized" by local Slotboom variables as described, for instance, in [15] to treat the convective part $F_i(\widetilde{d}) d_i \nabla V$. Finally, the "symmetrized" equations can be numerically discretized by mixed finite elements [15, 32]. We will numerically explore this idea for a higher-order moment model in a future paper.

Example 4.3 (energy-transport model). In the case of the energy-transport equations ($N = 1$), the functions $F_i(\lambda)$ in Proposition 4.1 simplify. Introducing the particle temperature $T = -1/\lambda_1$ as in Example 3.2, we obtain $F_0(d) = F_1(d) = \lambda_1 = -1/T$ and hence

$$J_i = -\nabla d_i + \frac{d_i}{T}\nabla V, \quad i = 0, 1.$$

The temperature is implicitly defined through the relation

$$f(T) = \frac{d_1}{d_0} = \frac{D_{10}}{D_{00}} = \frac{\langle \varepsilon u \otimes \phi_0\rangle}{\langle u \otimes \phi_0\rangle},$$

where ϕ_0 solves $L\phi_0 = uM_F$. A similar expression has been given in [15] but only in the case of monotone energy bands. For given d_0 and d_1, this defines T uniquely since $f'(T) = \det D/(Td_0)^2 > 0$. In order to check this derivative, we first compute

$$L\Big(\frac{\partial \phi_0}{\partial T}\Big) = \frac{\partial}{\partial T}(ue^{\lambda_0 - \varepsilon/T}) = \frac{1}{T^2}\varepsilon u M_F = \frac{1}{T^2}L\phi_1.$$

Hence, $\partial\phi_0/\partial T = \phi_1/T^2 + cM_F$, where c is a constant. Thus, since $\langle \varepsilon u \otimes \phi_0\rangle = D_{10} = D_{01} = \langle u \otimes \phi_1\rangle$ and $D_{11} = \langle \varepsilon u \otimes \phi_1\rangle$,

$$\begin{aligned} f'(T) &= \frac{1}{T^2 d_0^2}\big(\langle \varepsilon u \otimes \phi_1\rangle\langle u \otimes \phi_0\rangle - \langle \varepsilon u \otimes \phi_0\rangle\langle u \otimes \phi_1\rangle\big) \\ &= \frac{1}{T^2 d_0^2}(D_{11}D_{00} - D_{10}D_{01}) = \frac{\det D}{T^2 d_0^2} > 0. \end{aligned}$$

Example 4.4 (fourth-order model). We take $N = 2$ and assume the parabolic band approximation. The functions $F_i(d)$ read as follows:

$$F_i(d) = \lambda_1 + 2\frac{d_{i+1}}{d_i}\lambda_2, \quad i = 0, 1, 2.$$

Notice that, by (3.14), $d_i = (2\tau/3)m_{i+1}$. Moreover, integration by parts gives, using (3.13),

$$\begin{aligned} m_i &= -4\sqrt{2}\pi e^{\lambda_0}\int_0^\infty \frac{2}{2i+3}\varepsilon^{i+3/2}(\lambda_1 + 2\lambda_2\varepsilon)e^{\lambda_1\varepsilon + \lambda_2\varepsilon^2}\,d\varepsilon \\ &= -\frac{2}{2i+3}(\lambda_1 m_{i+1} + 2\lambda_2 m_{i+2}) = -\frac{3}{(2i+3)\tau}(\lambda_1 d_i + 2\lambda_2 d_{i+1}). \end{aligned} \tag{4.4}$$

Hence,

$$F_i(d) = \frac{1}{d_i}(\lambda_1 d_i + 2\lambda_2 d_{i+1}) = -\frac{(2i+3)\tau}{3}\,\frac{m_i}{d_i},$$

and the fluxes become, for constant relaxation time,

$$J_i = -\nabla d_i - F_i(d)d_i\nabla V = -\frac{2}{3}\tau\Big(\nabla m_{i+1} - \frac{2i+3}{2}m_i\nabla V\Big), \quad i = 0, 1, 2. \tag{4.5}$$

Together with the balance equations (2.12), we obtain a system of three equations for the unknowns m_0, m_1, and m_2. If τ depends on x or t, the variables are τm_0, τm_1, and τm_2. In the expression for J_2, the moment m_3 is needed. However, it can be computed from m_0, m_1, and m_2 using the relation

$$m_3 = -\frac{1}{2\lambda_2}\Big(\frac{5}{2}m_1 + \lambda_1 m_2\Big), \tag{4.6}$$

which comes from (4.4), where λ_1, λ_2 are functions of $m = (m_0, m_1, m_2)$. The fourth-order model with the above current relations can be also seen as a system of parabolic equations in the variables m_1, m_2, and m_3; the particle density m_0 is then a function of m_1, m_2, and m_3.

It remains to show that the function $m(\lambda)$, with $\lambda = (\lambda_0, \lambda_1, \lambda_2)$, can be inverted. This comes from the fact that the matrix $m'(\lambda) = (m_{i+j})_{i,j} \in \mathbb{R}^{3\times 3}$ is positive definite (and hence, its determinant is positive) since it is equal to the Hessian of the strictly convex function

$$\lambda \mapsto m_0 = 4\sqrt{2}\pi\tau \int_0^\infty \varepsilon^{1/2} e^{\lambda_0 + \lambda_1\varepsilon + \lambda_2\varepsilon^2} d\varepsilon.$$

The final fourth-order model consists of the balance equations (3.10)–(3.12) and the current relations (4.5) in the variables m_1, m_2, and m_3.

Remark 4.5. Grasser et al. have derived a related fourth-order model, called the six-moments transport equations (see (124)–(129) in [29]). The model equations are given by (3.10)–(3.12) and (4.5), where

$$m_0 = n, \quad m_1 = \frac{3}{2}nT, \quad m_2 = \frac{5\cdot 3}{4}nT^2\beta_n. \tag{4.7}$$

Here the variables are the particle density n, the electron temperature T, and the kurtosis β_n. This notation is inspired from the energy-transport model in the parabolic band approximation (see Example 3.4), where $m_2 = \frac{15}{4}nT^2$ (see (3.9)). In this sense, β_n measures the deviation from the heated Maxwellian $M_F = e^{\lambda_0 - \varepsilon/T}$. More generally, the kurtosis is defined by

$$\beta_n = \frac{3}{5}\,\frac{m_0 m_2}{m_1^2}.$$

By the Cauchy–Schwarz inequality

$$\begin{aligned} m_1^2 &= 32\pi^2 e^{2\lambda_0}\left(\int_0^\infty \varepsilon^{1/4}\varepsilon^{5/4} e^{\lambda_1\varepsilon + \lambda_2\varepsilon^2} d\varepsilon\right)^2 \\ &\le 32\pi^2 e^{2\lambda_0}\int_0^\infty \varepsilon^{1/2} e^{\lambda_1\varepsilon + \lambda_2\varepsilon^2} d\varepsilon \int_0^\infty \varepsilon^{5/2} e^{\lambda_1\varepsilon + \lambda_2\varepsilon^2} d\varepsilon = m_0 m_2, \end{aligned}$$

we obtain the restriction $\beta_n \ge 3/5$.

Grasser et al. [29] define heuristically m_3 in terms of the lower-order moments by setting

$$m_3 = \frac{7\cdot 5\cdot 3}{8}nT^2\beta_n^c, \tag{4.8}$$

where the constant exponent c is fitted from Monte Carlo simulations of the Boltzmann equation, computing the numerical moment m_3^{MC}. It has been found that the choice

$c = 3$ gives the smallest deviation of the ratio m_3^{MC}/m_3 from the desired value of one [29].

In the model derived in Example 4.4, m_3 is implicitly defined in terms of the lower-order moments; see (4.6). Using notation (4.7) and setting $\lambda_1 = -1/T$ as in the energy-transport equations, we obtain from (4.6)

$$m_3 = -\frac{15}{8}\frac{(1-\beta_n)nT}{\lambda_2}.$$

The expression (4.8) is obtained by setting $\lambda_2 = -(1-\beta_n)/7T^2\beta_n^c$. Since it should hold that $\lambda_2 < 0$, we conclude the restriction $\beta_n \leq 1$. Together with the above condition, the kurtosis has to satisfy the inequality $3/5 \leq \beta_n \leq 1$ [26]. Clearly, $\beta_n = 1$ corresponds to the energy-transport case for which $\lambda_2 = 0$.

Thus, the model of Grasser et al. is contained in our model hierarchy with the heuristic choice $\lambda_2 = -(1-\beta_n)/7T^2\beta_n^c$.

4.2. Dual-entropy variable formulation. It is well known from nonequilibrium thermodynamics that the electric force terms in (4.1) can be removed by employing so-called dual-entropy variables [19, 40]. Here we extend this methodology to higher-order moment models by defining the (generalized) dual-entropy variables $\nu = (\nu_0, \ldots, \nu_N)^\top$ by $\lambda = P\nu$, where $\lambda = (\lambda_0, \ldots, \lambda_N)^\top$ are the Lagrange multipliers (or the primal entropy variables), and the transformation matrix $P = (P_{ij}) \in \mathbb{R}^{(N+1)\times(N+1)}$ is defined by

$$P_{ij} = (-1)^{i+j}\binom{j}{i}a_{ij}V^{j-i} \quad \text{with} \quad a_{ij} = \begin{cases} 1 & \text{if } i \leq j, \\ 0 & \text{if } i > j, \end{cases}$$

where $i, j = 0, \ldots, N$. The dual-entropy formulation "symmetrizes" the system of equations [13]. It is well known that the existence of such variables is equivalent to the existence of an entropy functional [14, 39].

PROPOSITION 4.6. *Define the dual-entropy variables $\nu = (\nu_0, \ldots, \nu_N)^\top$, the transformed moments $\rho = (\rho_0, \ldots, \rho_N)^\top$, and the thermodynamic fluxes $F = (F_0, \ldots, F_N)^\top$ by*

$$\lambda = P\nu, \quad \rho = P^\top m, \quad \textit{and} \quad F = P^\top J,$$

respectively. Then the model equations (4.1) can be equivalently written as

$$\partial_t\rho_i + \mathrm{div}F_i = (P^\top W + V^{-1}\partial_t VRm)_i, \quad F_i = -\sum_{j=0}^{N} C_{ij}\nabla\nu_i,$$

where $W = (0, W_1, \ldots, W_N)^\top$, $R = (R_{ij})$ is given by $R_{ij} = (i-j)P_{ji}$, and the new diffusion matrix $C = (C_{ij})$ is defined by $C = P^\top DP$.

The proposition is proved in the appendix. Notice that the new diffusion matrix C is symmetric and positive definite if and only if D is symmetric and positive definite (see Lemma 2.12).

Example 4.7 (energy-transport model). The transformation matrix P and its inverse Q read in the case $N = 1$ as

$$P = \begin{pmatrix} 1 & -V \\ 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & V \\ 0 & 1 \end{pmatrix}.$$

Defining the chemical potential μ by $\lambda_0 = \mu/T$, where $T = -1/\lambda_1 > 0$ is the particle temperature, the dual-entropy variable $\nu = Q\lambda$ becomes (see, e.g., [13, 40])

$$\nu_0 = \lambda_0 + V\lambda_1 = \frac{\mu - V}{T}, \quad \nu_1 = \lambda_1 = -\frac{1}{T}.$$

The quantity $\mu - V$ is known as the *electrochemical potential.*

Example 4.8 (fourth-order model). For $N = 2$, the transformation matrix is given by

$$P = \begin{pmatrix} 1 & -V & V^2 \\ 0 & 1 & -2V \\ 0 & 0 & 1 \end{pmatrix}.$$

Introducing the chemical potential and the temperature as in the previous example and the *second-order temperature* θ as in [27] by $\lambda_2 = -1/\theta T$, the dual-entropy variables are

$$\nu_0 = \frac{\mu - V}{T} - \frac{V^2}{\theta T}, \quad \nu_1 = -\frac{1}{T} - \frac{2V}{\theta T}, \quad \nu_2 = -\frac{1}{\theta T}.$$

The dual-entropy formulation allows us to prove entropy dissipation. We define the relative entropy H_0 by

$$H_0(t) = -\int_{\mathbb{R}^3} (m \cdot (\lambda - \bar\lambda) - m_0 + \bar m_0) dx \le 0,$$

where $\lambda = (\lambda_0, \dots, \lambda_N)^\top$, $m = (m_0, \dots, m_N)^\top$, $\bar\lambda = (V, -1, 0, \dots, 0)^\top$, and $\bar m_0 = m_0(\bar\lambda)$ are the equilibrium values (since $e^{\bar\lambda \cdot \kappa} = e^{V - \varepsilon}$ is the equilibrium distribution function in the presence of an electric field). Notice that, in the situation of Example 3.4 (i.e., $N = 1$), the relative entropy becomes

$$H = -\int_{\mathbb{R}^3} \Big(n\Big(\ln n - \frac{3}{2}\ln T - \frac{5}{2} - V\Big) + \frac{3}{2}nT + e^V \Big) dx.$$

PROPOSITION 4.9. *Assume that*

$$\int_{\mathbb{R}^3} W \cdot (\lambda - \bar\lambda) dx \le 0. \tag{4.9}$$

Then any (smooth) solution λ to the higher-order moment equations (4.1) *satisfies the entropy inequality*

$$-\frac{dH}{dt} + \int_{\mathbb{R}^3} \sum_{i,j=0}^{N} C_{ij} \nabla\nu_i \cdot \nabla\nu_j dx \le 0.$$

The second integral on the left-hand side is called *entropy dissipation.* Clearly, it is nonnegative if the diffusion matrix D is positive (semi)definite. Thus, the entropy is nondecreasing in time.

Proof. We introduce the relative entropy density $h(\lambda) = -m \cdot (\lambda - \bar\lambda) + m_0 - \bar m_0$. The moments are given by (4.2) such that $\partial m_0/\partial\lambda_i = m_i$, from which we obtain

$$\frac{\partial h}{\partial \lambda_i} = -\frac{\partial m}{\partial \lambda_i} \cdot (\lambda - \bar\lambda) - m_i + \frac{\partial m_0}{\partial \lambda_i} = -\frac{\partial m}{\partial \lambda_i} \cdot (\lambda - \bar\lambda)$$

and

$$\partial_t m \cdot (\lambda - \bar{\lambda}) = \sum_{i=0}^{N} \frac{\partial m}{\partial \lambda_i} \cdot (\lambda - \bar{\lambda}) \partial_t \lambda_i = -\sum_{i=0}^{N} \frac{\partial h}{\partial \lambda_i} \partial_t \lambda_i = -\partial_t h(\lambda). \tag{4.10}$$

The balance equations (4.1) are formally equivalent to (A.2); multiplying the latter equations by $\nu_i - \bar{\nu}_i$, where $\bar{\nu} = Q\bar{\lambda}$, and summing over $i = 0, \ldots, N$, it follows that

$$(P^\top \partial_t m)^\top (\nu - \bar{\nu}) + (\mathrm{div} F)^\top (\nu - \bar{\nu}) = (P^\top W)^\top (\nu - \bar{\nu}).$$

Integrating over x and employing the definition $\nu = Q\lambda$ gives

$$\int_{\mathbb{R}^3} \partial_t m^\top PQ(\lambda - \bar{\lambda}) dx + \int_{\mathbb{R}^3} \sum_{i,j=0}^{N} \mathrm{div}(C_{ij} \nabla \nu_j)(\nu_i - \bar{\nu}_i) dx = \int_{\mathbb{R}^3} W^\top PQ(\lambda - \bar{\lambda}) dx.$$

Finally, integrating by parts in the second integral, taking into account that $\nabla \bar{\nu} = 0$, and using (4.10) yields

$$-\int_{\mathbb{R}^3} \partial_t h(\lambda) dx + \int_{\mathbb{R}^3} \sum_{i,j=0}^{N} C_{ij} \nabla \nu_i \cdot \nabla \nu_j dx = \int_{\mathbb{R}^3} W^\top (\lambda - \bar{\lambda}) dx \le 0,$$

which proves the lemma. □

In [7, Lem. 4.11], it has been shown that assumption (4.9) on W holds for an inelastic phonon collision operator, in the case of the energy-transport model. This hypothesis also holds if

$$W_i = -\frac{1}{\tau_1}(m_i - \bar{m}_i), \quad \text{where } \bar{m}_i = m_i(\bar{\lambda}),$$

since

$$W \cdot (\lambda - \bar{\lambda}) = -\frac{1}{\tau_1} \int_B (e^{\kappa \cdot \lambda} - e^{\kappa \cdot \bar{\lambda}})(\kappa \cdot \lambda - \kappa \cdot \bar{\lambda}) dp \le 0.$$

5. Conclusions. In this paper, we have derived a new hierarchy of diffusive models from the semiconductor Boltzmann equation by using a moment method and a Chapman–Enskog expansion, based on the entropy maximization principle of Levermore. The hierarchy contains well-known transport models, such as the drift-diffusion equations, the energy-transport equations, and a variant of the six-moments model of Grasser et al. Some features of the new models are (formally) shown: The diffusion matrix is positive definite, the flux equations can be written in a drift-diffusion form suitable for numerical discretizations, and the convective parts due to the electric field can be eliminated by employing generalized dual-entropy variables.

We mention some limitations of this model hierarchy. First, one may criticize hypothesis (H2) in which we require that all moments of the dominant part of the collision operator vanish. The hypothesis is clearly satisfied by a relaxation-type operator as shown in section 2.2, but it is not clear whether more realistic scattering operators satisfy this hypothesis. Second, numerical experiments have to show whether realistic simulation results for the higher-order models applied to small-channel devices can be obtained and whether the numerical effort is moderate compared to other (hydrodynamic or diffusive) models.

An important question is how many moments are actually needed in order to obtain accurate numerical results. In this direction, we mention the works of Schmeiser and Zwirchmayr [49] and of Struchtrup [54] for hyperbolic transport equations. In [54] the number of moments could be significantly reduced by a proper construction of moments.

In a future work, we intend to implement the fourth-order moment model using a mixed finite-element method and to compare the numerical results with those from the (similar) six-moments model of Grasser et al. [28]. Moreover, we intend to extend the hierarchy of diffusive models to Fermi–Dirac statistics (see, for instance, [6]). We expect that the decoupled drift-diffusion formulation has the potential to keep the computational cost down.

Appendix. We present the technical proofs of some results.

Proof of Lemma 2.12. The proof is inspired from the proof of Proposition IV.6 in [6]. We write as above $D_{ij} = (D_{ij}^{k\ell})$ and $\xi_i = (\xi_{ik})$. Let $(\xi_0, \dots, \xi_N) \neq 0$. Then, by the definition of the matrices D_{ij},

$$\sum_{i,j=0}^{N} \xi_i^\top D_{ij} \xi_j = \sum_{i,j=0}^{N} \sum_{k,\ell=1}^{3} \xi_{ik} D_{ij}^{k\ell} \xi_{j\ell} = -\sum_{i,j=0}^{N} \sum_{k,\ell=1}^{3} \int_B \xi_{ik} \kappa_i u_k \phi_{j\ell} \xi_{j\ell} dp.$$

Since $\kappa_i u_k M_F = L\phi_{ik}$, we obtain

$$\begin{aligned}\sum_{i,j=0}^{N} \xi_i^\top D_{ij} \xi_j &= -\sum_{i,j=0}^{N} \sum_{k,\ell=1}^{3} \int_B \xi_{ik} L\phi_{ik} \phi_{j\ell} \xi_{j\ell} M_F^{-1} dp \\ &= \sum_{i,j=0}^{N} \sum_{k,\ell=1}^{3} \big(- L(\xi_{ik}\phi_{ik}), \xi_{j\ell}\phi_{j\ell} \big)_F \\ &= \left(-L \left(\sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik}\phi_{ik} \right), \sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik}\phi_{ik} \right)_F .\end{aligned}$$

As $\phi_{ik} \in N(L)^\perp$, assumption (H5) and the boundedness of L (with constant $c_L > 0$) give

$$\begin{aligned}\sum_{i,j=0}^{N} \xi_i^\top D_{ij} \xi_j &\geq \mu \left\| \sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik}\phi_{ik} \right\|_F^2 \geq \frac{\mu}{c_L^2} \left\| L \left(\sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik}\phi_{ik} \right) \right\|_F^2 \\ &= \frac{\mu}{c_L^2} \left\| \sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik} u_k \kappa_i M_F \right\|_F^2 = \frac{\mu}{c_L^2} \int_B \left| \sum_{i=0}^{N} \sum_{k=1}^{3} \xi_{ik} u_k \kappa_i \right|^2 M_F dp > 0,\end{aligned}$$

since the functions $u_k\kappa_i$ are linearly independent.

Proof of Proposition 4.6. First, we prove some properties of the transformation matrix P which is needed in the proof of the proposition.

LEMMA A.1. (i) *The matrix $Q = (Q_{ij})$ given by $Q_{ij} = \binom{j}{i} a_{ij} V^{j-i}$ is the inverse of P.*

(ii) *For all i, $j = 0, \dots, N$,*

$$\sum_{k=0}^{N} (j-k) P_{ik} Q_{kj} = -\sum_{k=0}^{N} (j-k) Q_{ik} P_{kj} = j\delta_{i,j-1} V,$$

where $j\delta_{i,j-1} = 0$ for $j = 0$.

(iii) *For all* $i = 0, \dots, N-1$, $j = 1, \dots, N$,

$$-jP_{i,j-1} + (i+1)P_{i+1,j} = 0.$$

Proof. (i) By the definition of the coefficients a_{ij}, we have $\sum_k P_{ik}Q_{kj} = 0$ for all $i > j$. Let $i < j$. Then

$$\begin{aligned}\sum_{k=0}^{N} P_{ik}Q_{kj} &= \sum_{k=i}^{j}(-1)^{i+k}\binom{k}{i}\binom{j}{k}V^{j-i} = V^{j-i}\sum_{k=i}^{j}(-1)^{i+k}\binom{j}{i}\binom{j-i}{k-i} \\ &= V^{j-i}\binom{j}{i}\sum_{\ell=0}^{j-i}(-1)^{\ell}\binom{j-i}{\ell} = 0.\end{aligned}$$

Furthermore, for $i = j$, we obtain

$$\sum_{k=0}^{N} P_{ik}Q_{ki} = \sum_{k=i}^{i}(-1)^{i+k}\binom{k}{i}\binom{i}{k} = 1.$$

(ii) The definition of a_{ij} yields $\sum_k (j-k)P_{ik}Q_{kj} = 0$ for $i \geq j$. Next, let $i < j-1$. Then

$$\begin{aligned}\sum_{k=0}^{N}(j-k)P_{ik}Q_{kj} &= V^{j-i}\sum_{k=i}^{j-1}(j-k)(-1)^{i+k}\binom{k}{i}\binom{j}{k} \\ &= V^{j-i}\sum_{k=i}^{j-1}(-1)^{i+k}j\binom{j-1}{i}\binom{j-1-i}{k-i} \\ &= jV^{j-i}\binom{j-1}{i}\sum_{\ell=0}^{j-1-i}(-1)^{\ell}\binom{j-1-i}{\ell} = 0.\end{aligned}$$

If $i = j-1$, then

$$\sum_{k=0}^{N}(j-k)P_{ik}Q_{kj} = V\sum_{k=j-1}^{j-1}(j-k)(-1)^{j-1+k}\binom{k}{j-1}\binom{j}{k} = V\binom{j-1}{j-1}\binom{j}{j-1} = jV.$$

The second equality is shown in a similar way.

(iii) For $i \geq j$ we have $P_{i,j-1} = 0$ and $P_{i+1,j} = 0$. If $i < j$, then

$$-jP_{i,j-1} + (i+1)P_{i+1,j} = (-1)^{i+j+1}V^{j-1-i}\Big(-j\binom{j-1}{i} + (i+1)\binom{j}{i+1}\Big) = 0.$$

This proves the lemma. □

Now we proceed to the proof of Proposition 4.6. First we show the relation for the new fluxes. Employing the definitions $C = P^{\top}DP$ and $\nu = Q\lambda$ and the property

$QP = I$ (I being the identity matrix), we obtain

$$\begin{aligned}\sum_{j=0}^{N} C_{ij}\nabla\nu_j &= \sum_{j,k,\ell,n=0}^{N} P_{ki}D_{k\ell}P_{\ell j}\nabla(Q_{jn}\lambda_n)\\ &= \sum_{j,k,\ell,n=0}^{N} P_{ki}D_{k\ell}(P_{\ell j}Q_{jn}\nabla\lambda_n + P_{\ell j}\nabla Q_{jn}\lambda_n)\\ &= \sum_{k,\ell=0}^{N} P_{ki}D_{k\ell}\nabla\lambda_\ell + \sum_{k,\ell,n=0}^{N} P_{ki}D_{k\ell}\left(\sum_{j=0}^{N}(n-j)P_{\ell j}Q_{jn}\right)V^{-1}\nabla V\lambda_n,\end{aligned}$$

since $\nabla Q_{jn} = (n-j)V^{-1}\nabla V Q_{jn}$. Now, using Lemma A.1(ii),

$$\begin{aligned}\sum_{j=0}^{N} C_{ij}\nabla\nu_j &= \sum_{k,\ell=0}^{N} P_{ki}D_{k\ell}\nabla\lambda_\ell + \sum_{k,\ell,n=0}^{N} P_{ki}D_{k\ell}n\delta_{\ell,n-1}\nabla V\lambda_n\\ &= \sum_{k,n=0}^{N} P_{ki}(D_{kn}\nabla\lambda_n + nD_{k,n-1}\nabla V\lambda_n) = -\sum_{k=0}^{N} P_{ki}J_k = -F_i.\end{aligned}$$

Next we compute the transformed balance equations. By the definition of F_i,

$$\begin{aligned}\mathrm{div}F_i &= \sum_{j=0}^{N}\mathrm{div}(P_{ji}J_j) = \sum_{j=0}^{N}(P_{ji}\mathrm{div}J_j + \nabla P_{ji}\cdot J_j)\\ &= \sum_{j=0}^{N} P_{ji}(\mathrm{div}J_j - jJ_{j-1}\cdot\nabla V) + \sum_{j=0}^{N}(\nabla P_{ji}\cdot J_j + jP_{ji}J_{j-1}\cdot\nabla V).\end{aligned}\tag{A.1}$$

We show that the second sum vanishes. Observing that $\nabla P_{ji} = (i-j)V^{-1}\nabla V P_{ji}$, we find that

$$A := \sum_{j=0}^{N}(\nabla P_{ji}\cdot J_j + jP_{ji}J_{j-1}\cdot\nabla V) = \sum_{j=0}^{N}\big((i-j)P_{ji}V^{-1}\nabla V\cdot J_j + jP_{ji}J_{j-1}\cdot\nabla V\big).$$

Since the first sum can be rewritten, by Lemma A.1(ii), as

$$\begin{aligned}&\sum_{j=0}^{N}(i-j)P_{ji}V^{-1}\nabla V\cdot J_j = \sum_{j,k=0}^{N}(i-k)\delta_{jk}P_{ki}V^{-1}J_j\cdot\nabla V\\ &\quad= \sum_{j,k,\ell=0}^{N}(i-k)P_{j\ell}Q_{\ell k}P_{ki}V^{-1}J_j\cdot\nabla V = \sum_{j,\ell=0}^{N}\left(\sum_{k=0}^{N}(i-k)Q_{\ell k}P_{ki}\right)P_{j\ell}V^{-1}J_j\cdot\nabla V\\ &\quad= -\sum_{j,\ell=0}^{N} i\delta_{\ell,i-1}P_{j\ell}J_j\cdot\nabla V = -\sum_{j=0}^{N} iP_{j+1,i}J_j\cdot\nabla V,\end{aligned}$$

we obtain

$$A = \sum_{j=0}^{N-1}(-iP_{j,i-1} + (j+1)P_{j+1,i})J_j\cdot\nabla V = 0,$$

using Lemma A.1(iii). Hence, with the balance equations (4.1), (A.1) becomes

$$\mathrm{div} F_i = \sum_{j=0}^{N} P_{ji}(-\partial_t m_j + W_j). \tag{A.2}$$

We employ the definition $\rho = P^\top m$ to rewrite the first sum:

$$\begin{aligned}\sum_{j=0}^{N} P_{ji}\partial_t m_j &= \sum_{j=0}^{N}\big(\partial_t(P_{ji}m_j) - \partial_t P_{ji} m_j\big) \\ &= \partial_t \rho_i - V^{-1}\partial_t V \sum_{j=0}^{N}(i-j)P_{ji}m_j = \partial_t\rho_i - V^{-1}\partial_t V\sum_{j=0}^{N} R_{ij}m_j.\end{aligned}$$

This finishes the proof. □

REFERENCES

[1] A. M. Anile and O. Muscato, *Improved hydrodynamic model for carrier transport in semiconductors*, Phys. Rev. B, 51 (1995), pp. 16728–16740.

[2] A. M. Anile, N. Nikiforakis, V. Romano, and G. Russo, *Discretization of semiconductor device problems* (II), in Handbook of Numerical Analysis XIII, W. H. A. Schilders and E. J. W. ter Maten, eds., Numerical Methods in Electromagnetics, North-Holland, Amsterdam, 2005.

[3] A. M. Anile and S. Pennisi, *Thermodynamic derivation of the hydrodynamical model for charge transport in semiconductors*, Phys. Rev. B, 46 (1992), pp. 13186–13193.

[4] A. M. Anile, G. Russo, and V. Romano, *Extended hydrodynamical model of carrier transport in semiconductors*, SIAM J. Appl. Math., 61 (2000), pp. 74–101.

[5] G. Baccarani and M. R. Wordeman, *An investigation of steady-state velocity overshoot in silicon*, Solid-State Electr., 29 (1982), pp. 970–977.

[6] N. B. Abdallah and P. Degond, *On a hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.

[7] N. B. Abdallah, P. Degond, and S. Génieys, *An energy-transport model for semiconductors derived from the Boltzmann equation*, J. Stat. Phys., 84 (1996), pp. 205–231.

[8] K. Bløtekjær, *Transport equations for electrons in two-valley semiconductors*, IEEE Trans. Electron Devices, 17 (1970), pp. 38–47.

[9] D. Cassi and B. Riccò, *An analytical model of the energy distribution of hot electrons*, IEEE Trans. Electron Devices, 37 (1990), pp. 1514–1521.

[10] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non–Uniform Gases*, Cambridge University Press, Cambridge, 1958.

[11] D. Chen, E. Kan, U. Ravaioli, C. Shu, and R. Dutton, *An improved energy transport model including nonparabolicity and non-Maxwellian distribution effects*, IEEE Electron Device Lett., 13 (1992), pp. 26–28.

[12] J. F. Coulombel, F. Golse, and T. Goudon, *Diffusion approximation and entropy-based moment closure for kinetic equations*, Asymptot. Anal., 45 (2005), pp. 1–39.

[13] P. Degond, S. Génieys, and A. Jüngel, *A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects*, J. Math. Pures Appl., 76 (1997), pp. 991–1015.

[14] P. Degond, S. Génieys, and A. Jüngel, *Symmetrization and entropy inequality for general diffusion equations*, C. R. Acad. Sci. Paris, 325 (1997), pp. 963–968.

[15] P. Degond, A. Jüngel, and P. Pietra, *Numerical discretization of energy-transport models for semiconductors with nonparabolic band structure*, SIAM J. Sci. Comput., 22 (2000), pp. 986–1007.

[16] P. Degond, C. Levermore, and C. Schmeiser, *A note on the energy-transport limit of the semiconductor Boltzmann equation*, in Proceedings of Transport in Transition Regimes (Minneapolis, 2000), IMA Vol. Math. Appl. 135, Springer, New York, 2004.

[17] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Stat. Phys., 112 (2003), pp. 587–628.

[18] W. Dreyer, *Maximisation of the entropy in non-equilibrium*, J. Phys. A, 20 (1987), pp. 6505–6517.

[19] S. de Groot, *Thermodynamik irreversibler Prozesse*, Bibliographisches Institut, Mannheim, 1960.

[20] P. Dmitruk, A. Saul, and L. Reyna, *High electric field approximation to charge transport in semiconductor devices*, Appl. Math. Lett., 5 (1992), pp. 99–102.

[21] W. Dreyer, M. Junk, and M. Kunik, *On the approximation of kinetic equations by moment systems*, Nonlinearity, 14 (2001), pp. 881–906.

[22] W. Fang and K. Ito, *Existence of stationary solutions to an energy drift-diffusion model for semiconductor devices*, Math. Models Methods Appl. Sci., 11 (2001), pp. 827–840.

[23] S. Gadau and A. Jüngel, *Finite-Element Approximation of the 3D Energy-Transport Models for Semiconductors*, manuscript, 2007.

[24] B. Geurts, M. Nekovee, H. Boots, and M. Schuurmans, *Exact and moment equation modeling of electron transport in submicron structures*, Appl. Phys. Lett., 59 (1991), pp. 1743–1745.

[25] T. Grasser, *Non-parabolic macroscopic transport models for semiconductor device simulation*, Phys. A, 349 (2005), pp. 221–258.

[26] T. Grasser, R. Kosik, C. Jungemann, H. Kosina, and S. Selberherr, *Nonparabolic macroscopic transport models for device simulation based on bulk Monte Carlo data*, J. Appl. Phys., 97 (2005), 093710.

[27] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, *Using six moments of Boltzmann's equation for device simulation*, J. Appl. Phys., 90 (2001), pp. 2389–2396.

[28] T. Grasser, H. Kosina, C. Heitzinger and S. Selberherr, *Characterization of the hot electron distribution function using six moments*, J. Appl. Phys., 91 (2002), pp. 3869–3879.

[29] T. Grasser, H. Kosina, and S. Selberherr, *Hot carrier effects within macroscopic transport models*, Internat. J. High Speed Electr. Sys., 13 (2003), pp. 873–901.

[30] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, *A review of hydrodynamic and energy-transport models for semiconductor device simulation*, Proc. IEEE, 91 (2003), pp. 251–274.

[31] J. Griepentrog, *An application of the implicit function theorem to an energy model of the semiconductor theory*, Z. Angew. Math. Mech., 79 (1999), pp. 43–51.

[32] S. Holst, A. Jüngel, and P. Pietra, *A mixed finite-element discretization of the energy-transport model for semiconductors*, SIAM J. Sci. Comput., 24 (2003), pp. 2058–2075.

[33] S. Ihara, *Information Theory for Continuous Systems*, World Scientific, Singapore, 1993.

[34] J. Jerome, *Analysis of Charge Transport. A Mathematical Study of Semiconductor Devices*, Springer, Berlin, 1996.

[35] M. Junk, *Domain of definition of Levermore's five-moment system*, J. Stat. Phys., 93 (1998), pp. 1143–1167.

[36] A. Jüngel, *Quasi-hydrodynamic Semiconductor Equations*, Progr. Nonlinear Differential Equations, Birkhäuser, Basel, 2001.

[37] A. Jüngel, D. Matthes, and J. Milišić, *Derivation of new quantum hydrodynamic equations using entropy minimization*, SIAM J. Appl. Math., 67 (2006), pp. 46–68.

[38] M. Junk and V. Romano, *Maximum entropy moment systems of the semiconductor Boltzmann equation using Kane's dispersion relation*, Contin. Mech. Thermodyn., 17 (2004), pp. 247–267.

[39] S. Kawashima and Y. Shizuta, *On the normal form of the symmetric hyperbolic-parabolic systems associated with conservation laws*, Tohoku Math. J., 40 (1988), pp. 449–464.

[40] H. Kreuzer, *Nonequilibrium Thermodynamics and Its Statistical Foundation*, Clarendon Press, Oxford, 1981.

[41] C. Levermore, *Moment closure hierarchies for kinetic theories*, J. Stat. Phys., 83 (1996), pp. 1021–1065.

[42] S. Liotta and H. Struchtrup, *Moment equations for electrons in semiconductors: Comparison of spherical harmonics and full moments*, Solid-State Electr., 44 (2000), pp. 95–103.

[43] E. Lyumkis, B. Polsky, A. Shur, and P. Visocky, *Transient semiconductor device simulation including energy balance equations*, COMPEL, 11 (1992), pp. 311–325.

[44] P. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Vienna, 1990.

[45] I. Müller and T. Ruggeri, *Extended Thermodynamics*, Springer, New York, 1993.

[46] M. Nekovee, B. Geurts, H. Boots, and M. Schuurmans, *Failure of extended-moment-equation approaches to describe ballistic transport in submicrometer structures*, Phys. Rev. B, 45 (1992), pp. 6643–6651.

[47] V. Romano, *Non-parabolic band hydrodynamical model of silicon semiconductors and simulation of electron devices*, Math. Methods Appl. Sci., 24 (2001), pp. 439–471.

[48] M. Rudan and G. Baccarani, *On the structure and closure condition of the hydrodynamical model*, VLSI Design, 3 (1995), pp. 115–129.

[49] C. Schmeiser and A. Zwirchmayr, *Convergence of moment methods for linear kinetic equations*, SIAM J. Numer. Anal., 36 (1998), pp. 74–88.

[50] J. Schneider, *Entropic approximation in kinetic theory*, ESAIM: Math. Mod. Numer. Anal., 38 (2004), pp. 541–561.

[51] K.-I. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. Dunham, *Moment expansion approach to calculate impact ionization rate in submicron silicon devices*, J. Appl. Phys., 80 (1996), pp. 5444–5448.

[52] R. Stratton, *Diffusion of hot and cold electrons in semiconductor barriers*, Phys. Rev., 126 (1962), pp. 2002–2014.

[53] H. Struchtrup, *Extented moment method for electrons in semiconductors*, Phys. A, 275 (2000), pp. 229–255.

[54] H. Struchtrup, *Derivation of* 13 *moment equations for rarefied gas flow to second order accuracy for arbitrary interaction potentials*, Multiscale Model. Simul., 3 (2005), pp. 221–243.

[55] D. Woolard, H. Tian, R. Trew, M. Littlejohn, and K. Kim, *Hydrodynamic electron-transport model: Nonparabolic corrections to the streaming terms*, Phys. Rev. B, 44 (1991), pp. 11119–11132.

SIAM J. APPL. MATH.
Vol. 68, No. 1, pp. 199–221

A DIRICHLET-INTEGRAL–BASED DUAL-ACCESS COLLOCATION-KERNEL APPROACH TO POINT SOURCE GRAVITY-FIELD MODELING[*]

ALAN RUFTY[†]

Abstract. Problems in $\mathbb{R}^3$ are addressed where the scalar potential of an associated vector field satisfies Laplace's equation in some unbounded external region and is to be approximated by unknown (point) sources contained in the complimentary subregion. Two specific field geometries are considered: $\mathbb{R}^3$ half-space and the exterior of an $\mathbb{R}^3$ sphere, which are the two standard settings for geophysical and geoexploration gravitational problems. For these geometries it is shown that a new type of kernel space exists, which is labeled a Dirichlet-integral dual-access collocation-kernel space (DIDACKS) and which is well suited for many applications. The DIDACKS examples studied are related to reproducing kernel Hilbert spaces, and they have a replicating kernel (as opposed to a reproducing kernel) that has the ubiquitous form of the inverse of the distance between a field point and a corresponding source point. Underpinning this approach are three basic mathematical relationships of general interest. Two of these relationships—corresponding to the two geometries—yield exact closed-form inner products and thus exact linear equation sets for the corresponding point source strengths of various types (i.e., point mass, point dipole, and/or point quadrupole sets) at specified source locations. The given field is reconstructed not only in a point collocation sense, but also in a (weighted) field-energy error-minimization sense.

Key words. Laplace's equation, inverse problem, reproducing kernels, fundamental solutions, point collocation, point source

AMS subject classifications. 35J05, 31B10, 86A22, 65D05

DOI. 10.1137/060659090

1. Introduction.

The goal of this article is to set forth a mathematical framework for the approximation of $\mathbb{R}^3$ harmonic fields in unbounded domains by point sources contained inside the complimentary region. The proposed Dirichlet-integral dual-access collocation-kernel (DIDACK) approach has a mathematically and physically well-motivated underpinning. The associated space (DIDACKS) has certain similarities to reproducing kernel Hilbert space (RKHS) but is distinct from it. Two concrete $\mathbb{R}^3$ geometries are considered: (A) The harmonic field region consisting of a half-space (denoted Ω_1). (B) The harmonic field region consisting of the exterior of a sphere (denoted Ω_0). Within this geometric context, the developed formalism easily handles various combinations of diverse types of point sources (such as point masses, point mass dipoles, or point mass quadrupoles); moreover, for a set of specified source locations the formalism yields closed-form linear equation sets that simultaneously minimize the volume integrals of (weighted) field energy densities (i.e., (weighted) Dirichlet integrals).

Techniques introduced here can either be applied directly or adapted for use in many mathematical and physical areas. Examples on the mathematical side include

[*] Received by the editors May 5, 2006; accepted for publication (in revised form) April 18, 2007; published electronically October 19, 2007. This work was performed by an employee of the U.S. Government or under U.S. Government contact. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published from of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siap/68-1/65909.html

[†] Naval Surface Warfare Center Dahlgren (Attn: Alan Rufty), 5443 Bronson Road, Suite 112, Dahlgren, VA 22448–5159.

potential theory, point source elliptic boundary value modeling (i.e., method of fundamental solutions), fast multipole techniques, radial-basis function techniques, RKHS techniques, geophysical collocation (GC) techniques, and standard energy minimization–based techniques (such as Galerkin and Raleigh–Ritz-based approaches). Also see [22]. Examples on the physical side include geoexploration (where gravity is often used to locate oil or other minerals), geophysics, and magnetostatics (for a survey of these three areas, see [12, 25, 23, 4]), as well as general electromagnetic source analysis, electrostatics, and hydrostatics (for the physical significance of these three areas, and of field energy with regards to them, see [3]). Also somewhat similar mathematical structures arise in many other areas such as biophysical or biomedical engineering (where, for example, electric dipole models are used in electrocardiographical (ECG) modeling and current dipole models are used in electroencephalography (EEG) [2]).

While these diverse candidate application areas exist, DIDACKS theory was developed to handle gravitational problems, and all of the author's direct numerical experience with it is in this arena; hence, gravitation is the application setting considered here. Geophysical gravitation has an easily understood notation and a readily accessible mature literature [10, 19]. (Note that [10] is based on the earlier book [9], which, in turn, was largely based on [8].) Geophysical gravitation also has a direct historical association with potential theory (for example, Gauss played a very significant role in the history of geophysics [10, p. 1]).

A family of significant and challenging problems (including representative geophysical inverse source problems) is associated with geophysical gravitation. In addition to Laplacian inverse source theory, where the goal is to determine source strength, geophysical gravitation has two separate problem categories that point sources can be used in: field modeling and field estimation. For gravitational modeling problems the field is assumed to be known throughout the region of interest, and a more compact, but accurate, representation is desired. For estimation problems it is assumed that values are known accurately at a certain number of points in the field region and that one wants to predict gravity values over some part of this field region. Besides combinations of point sources, other techniques (such as GC) exist for treating gravitational field modeling and field estimation problems. As currently understood, each of these approaches has certain advantages on its own home ground, and the application areas where they all can be considered direct competitors is somewhat limited. In a wider mathematical context, since the fundamental solutions used here are reinterpreted as kernels and satisfy a generalized collocation property, the DIDACKS modeling and estimation approaches described here can also be considered harmonic interpolation and extrapolation techniques, respectively. GC and its extensions are thus the family of approaches that are mathematically closest to DIDACKS theory.

With minor differences these gravitation ($\vec{G}$) problems use the same standard notation and techniques employed in electrostatics, where $\vec{E}$ is the field [11], and so this arena should be readily accessible to all interested applied mathematicians, physicists, and engineers. (Within this electrostatic context the formalism developed here handles multiple types of point sources—such as point charges, electrostatic dipoles, or electrostatic quadrupoles.) Thus, although the notation used in the first four sections is specialized to the gravitational setting, the intrusion of this setting is otherwise minimal in these sections. In fact, the required mathematical background for this "core part of the article" is limited to classical $\mathbb{R}^3$ potential theory [13, 11] and a basic understanding of functional analysis, as utilized in approximation theory. Because no previous geophysical background is presupposed in these first four sections of the paper, the relevant overall geophysical and point mass modeling context is supplied in section 5. Since GC is the most commonly used approach for regional gravitational data processing and estimation it is briefly described at the end of section 5. While a knowledge of neither RKHS nor GC is a necessity in these first five sections, a better understanding of both these areas is presupposed in section 6, which describes the mathematical connections, noted above, of DIDACKS theory to other kernel-based approaches. GC, which is often called least-squares collocation, is an RKHS-based approach that differs from standard collocation techniques utilized by applied mathematicians in several relevant ways and, as such, may not be familiar to many readers. Moritz [19] provides a readily accessible treatment of both RKHS and GC theory; nevertheless, the thrust of section 6 should be directly accessible to those familiar with only standard RKHS theory ([17], [1], or [27] (which contains [1])).

On a first reading, the first four sections of the article should probably be considered only from a mathematical perspective. One implicit, but strong, motivation exists for limiting the level of this core material—there is often a given mathematical level of sophistication from which certain research results tend to emerge, and the same results may not be apparent at either a higher or lower level of mathematical sophistication. The first four sections and the geophysical applications setting given in section 5 are aimed at this default level of sophistication; however, there are open-ended physical and mathematical foundational issues associated with DIDACKS theory implicitly touched on in the last half of section 2 that may be of primary interest to only the type two and type four readers/researchers mentioned below. Other readers may wish to simply skip this material in the second half of section 2 on a first reading. From a general reader's perspective, it is still worth noting that a central theme that emerges from DIDACKS theory is that well-known kernels (such as the fundamental solutions studied here) can be regarded as part of a new kernel setting, and taking this perspective seriously seems to imply not only applied research but also new mathematical results, of which the relationships given below are only one example. To see that all of the various mathematical, applied, and physical facets of DIDACKS theory hang together as a whole and then to envision what the research possibilities are is surely not an easy task, but the material itself can be approached in a step-by-step way and, when this is done, even cross-fertilization and educational possibilities emerge. Given all of this, it is useful to set the stage for the rest of the paper by taking a step back and previewing the basic mathematical relationships that produce closed-form inner products. Closed-form (weighted) energy inner products are the cornerstone of both the theory and application of the whole approach.

Basic DIDACKS relationships in $\mathbb{R}^3$. The DIDACKS approach was conceived and is best understood on its own merits as a new self-consistent mathematical theory that is independent of the geophysical connections just indicated and, as noted above, it is this exposition that occupies the first four sections. Underpinning this mathematical side of the DIDACKS approach are three intrinsically interesting relationships that will now be briefly surveyed. First consider the notation employed. For overall accessibility, for consistency with the geophysical and electrostatic literature, and to avoid various possible notational conflicts, preference is given to a pedestrian but unambiguous notation: Cartesian coordinates are used and overset arrows employed to denote $\mathbb{R}^3$ vectors, $\vec{X} = (x, y, z)^T \in \mathbb{R}^3$ (for superscript $T :=$ transpose), while for $n \neq 3$, n-dimensional vectors and matrices are denoted by lower- and uppercase bold letters, respectively. R_0 is used to denote the radius of the sphere associated with Ω_0 that is centered over the origin: $\Omega_0 := \{\vec{X} \in \mathbb{R}^3 \mid |\vec{X}| \geq \mathrm{R}_0\}$. Likewise for the half-space case, the origin is chosen in the plane $\partial\Omega_1 := \{\vec{X} \in \mathbb{R}^3 \mid z = 0\}$ so that $\Omega_1 := \{\vec{X} \in \mathbb{R}^3 \mid z \geq 0\}$. (Observe that the overall shape of the subscripts here matches that of the associated boundary.) Frequently these two settings will be denoted by a subscript j (for example, Ω_j), where $j = 0$ or 1 is always implied.

Temporarily leaving aside the issue of admissible functions, these relationships can be compactly stated in terms of a Dirichlet integral over some connected but possibly unbounded region Ω, which is usually denoted by $\mathrm{D}[v,\, w] = \iiint_\Omega \vec{\nabla} v \cdot \vec{\nabla} w \; dV$ for admissible harmonic functions $v(\vec{X})$ and $w(\vec{X})$, or in terms of the more inclusive concept of a weighted Dirichlet integral for the region Ω denoted by $\mathrm{D}[v,\, w,\, \mu,\, \Omega]$, where $\mu = \mu(\vec{X})$ is the weighting function so that $\mathrm{D}[v,\, w,\, \mu,\, \Omega] := \iiint_\Omega \mu\, \vec{\nabla} v \cdot \vec{\nabla} w \; dV$. Clearly $\mathrm{D}[v,\, w,\, 1,\, \Omega] = \mathrm{D}[v,\, w]$. Let $\ell^{-1} := 1/|\vec{X} - \vec{X}'|$, where $\vec{X} \in \Omega_j$ and $\vec{X}'$ is in the corresponding closed source region $:= \Omega_{S_j} \subset \Omega'_j :=$ compliment of Ω_j. (By convention, generally primed variables occur in Ω'_j and unprimed ones in Ω_j.) Then the first two relations give the replication (or generalized collocation) properties of the DIDACKS kernel ℓ^{-1}:

$$\mathrm{D}[w,\, \ell^{-1},\, 1,\, \Omega_1] \,=\, 2\pi\, w(x',\, y', -z') \tag{1.1}$$

and

$$\mathrm{D}[w,\, \ell^{-1},\, \mu_0,\, \Omega_0] \,=\, 2\pi\, |\vec{P}|\; w(\vec{P})/\mathrm{R}_0^2, \tag{1.2a}$$

with $\mu_0 = 1/r$ $(r := |\vec{X}|)$ and where

$$\vec{P} = \left(\frac{\mathrm{R}_0^2}{|\vec{X}'|^2} \right) \vec{X}'. \tag{1.2b}$$

Finally the third relationship ties the unweighted Dirichlet integral over Ω_0 to the weighted integral given on the left-hand side (LHS) of (1.2a) and can be written as

$$\mathrm{D}[v,\, w,\, 1,\, \Omega_0] = \mathrm{R}_0 \cdot \mathrm{D}[v,\, w,\, \mu_0,\, \Omega_0] + (2\pi \mathrm{R}_0) \cdot (v,\, w)_\sigma, \tag{1.3a}$$

where the surface inner product on the right-hand side (RHS) here is defined as

$$(v,\, w)_\sigma := (1/4\pi) \iint_\sigma v(r, \theta,\, \phi)\, w(r, \theta,\, \phi)\, d\,\sigma \tag{1.3b}$$

and where, as in [9] and [10], σ and $d\,\sigma$ have the following meaning when associated with the integral of $f(\vec{X})$:

$$\iint_\sigma f(r, \theta,\, \phi)\, d\,\sigma \;:= \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} [f(r,\, \theta,\, \phi)]\Big|_{r=R_0} \sin\,\theta \;\, d\,\theta\, d\,\phi \tag{1.4}$$

for standard spherical coordinates r, θ, ϕ. (Occasionally the limits implicit in (1.4) will be stated explicitly for emphasis.) Due to the way r dependence enters in (1.4) it can be used to derive expressions that are otherwise not obvious, as will be apparent in what follows.

Clearly (1.1) and (1.2a) give a means of performing closed-form inner products based on the Dirichlet integral, while (1.3a) links the weighted Dirichlet inner product to the unweighted one over Ω_0. The general approximation and functional analysis framework for these relationships is given in section 2. The derivation of (1.1) is given in section 3, and that of (1.2a) and (1.3a) in section 4. In order to understand the connections of these relationships to the applications described in section 5 and to put them in proper historical context, it is useful to briefly consider the history of these relationships. Originally (1.2a) and (1.3a) were discovered in a different form by the author in the early 1980s—namely, (4.8) and (4.3), where the "integral norm" introduced in section 4 is used in place of the weighted Dirichlet integral. The history of these relationships in this form and of their application from inception to the mid-1990s can be found in [21]. The germane part of this internal history is summarized and updated in section 5. As an aside, although (1.1) is not explicitly mentioned in [21], the derivation here of (1.1) given in section 3 is the one originally found by the author in the early 1980s. The direct relationship of the integral norm to the weighted Dirichlet norm (4.13) is new. Also unless otherwise noted, the presentation itself (including various terms and concepts) is completely new here. Finally, while the general DIDACKS technique is a synthesis of several results that seem to have been overlooked by the broader scientific community, for any mathematical approach that directly touches on harmonic analysis, RKHS theory, and gravimetric inverse source theory either precedents or specialized parallel lines of development would seem to be a necessity. The known ones for $\mathbb{R}^3$ DIDACKS theory are addressed in section 6. In one way or another all of the discussions of section 6 pertain to the second DIDACKS relation in weighted Dirichlet form, (1.2a). As discussed there, (1.2a) should clearly be viewed as going back to Krarup in [15], since he derived it there in a directly equivalent form. Aside from the author's work, there are no known instances of (1.1) and (1.3a), or for that matter for the second DIDACKS relation expressed in the integral norm form (4.8).

Four types of readers. To motivate a further exploration of the paper's scope and limits it is useful to envision the possible reactions of four typical classes of readers to the above relations. First, consider an applied mathematician, physicist, engineer, or other scientist who may be familiar with the material in the first four chapters of [11], but is not yet a seasoned practitioner and may benefit by consulting [17], [19], or [13]. (The physical and historical importance of Dirichlet integrals and of the associated Dirichlet principle may be obtained from other sources [3, 18, 6].) This reader may find all three of the above field energy expressions somewhat surprising: (1.1) and (1.2a) because they allow for the closed-form evaluation of volume integrals, and (1.3a) since it allows for an unusual reexpression of the volume field energy. This reader may also observe that (1.1) and (1.2a) appear to have some connection to Green's functions and the method of images [11] and may recognize (1.2b) as the coordinate transformation part of a Kelvin transformation [13, 14]. These particular connections arise from the nature of Green's functions for Ω_j, but the most direct explanation requires some knowledge of RKHS theory. First, as noted in [3], the existence of a closed-form reproducing kernel occurs when closed-form expressions for both Neumann and Dirichlet Green's functions exist. Second, for the cases studied here a dual-action collocation kernel (DACK) arises from the result of a reflection, (1.1), or Kelvin transformation, (1.2b), applied to a reproducing kernel of the right form for the relevant geometry.

Second, while the possible reactions of any number of specialists might also be examined, consider a reader who has a particular interest in integral kernels or RKHS theory. For various reasons this reader may also find the above relationships somewhat surprising. This reader might, for example, observe that ℓ^{-1} plays the role of a kernel and then recall that a symmetric reproducing kernel (SRK) of the form $|\vec{X} - \vec{Y}|^{-1}$ for $\vec{X}$ and $\vec{Y}$ both in the same region cannot exist, since a SRK must be bounded and this kernel is not. Here $\vec{X}'$ is a fixed interior point and $\vec{X}$ is in the exterior region, so ℓ^{-1} is bounded, which is a very different situation and implies a change of perspective. This in itself could raise further questions since although DACKs are not reproducing kernels they have some properties analogous to them. In the previous paragraph the kernels studied here were linked to reproducing kernels, but it is unclear whether minimum norm DACKs can arise in other ways. Moreover, consideration of DACKs as a separate class of kernels also raises the question as to how they fit into our current overall understanding of kernel structures. These and other issues of a general nature are outside the scope of the present article, but some specific topics that might interest the second type of reader, such as the exact definition of the replication and generalized collocation property mentioned above, are addressed in section 2.

Third, consider a reader who is very applications- and results-oriented. Such a reader may be disappointed to discover that there is not a table containing numerical examples and results; however, a discussion of point mass and point dipole DIDACKS results and their associated applications settings can be found in section 5, where global nonlinear least-squares (NLLSQ) results are emphasized. Due to the variety and nature of regional gravity data, as well as other issues [21], no known easily replicated example provides generic benchmark results, which is normally an expectation for these tabulated examples. Moreover, point source fitting problems are part of a general class of problems that are "notoriously" ill-conditioned and problematic [4, pp. 214–222] so that each problem encountered should be tackled on its own terms, which means that one or two simple table examples cannot serve to provide adequate implementation guidance. Unfortunately, a thorough discussion of associated implementation strategies is outside the scope of the present article. Also observe that a concrete example provides a replication check that can serve as a consistency test for implementors, but this point is largely superfluous here since the DIDACKS approach exhibits the generalized collocation property with respect to point sources, and thus, when implemented correctly, any point mass or dipole fit replicates the point field data that was used to produce it in the first place to within allowed round-off error, and thus any implementation serves as its own self-consistency test.

Fourth and finally, consider a reader whose primary interest is in the theory and application of Laplacian inverse source theory. Since there are many shared implementation pitfalls common to both point source field modeling and inverse source estimation problems, the comments just made in the last paragraph are also relevant in this context. While specific mathematical tools and implementation strategies are not discussed here, readers with solid applications experience should be able to make direct use of the formalism presented. These readers may also be interested in the topic of continuous parameterized distributions, which is raised in section 2. Finally, it is worth noting that other source region shapes can be entertained within the contexts of the two considered geometries, since the only real requirement is that source regions be bounded and contained within the compliment of the unbounded harmonic field region.

2. Generalized linear least-squares setting. This section addresses the generalized linear least-squares (GLLSQ) plan of approach and the associated functional space setting. There are numerous approaches closely aligned to the GLLSQ method adopted here, such as the Galerkin and Raleigh–Ritz-based techniques mentioned in the introduction; however, the acronym GLLSQ is introduced to imply an implicit change of perspective. In particular, connections to generalized collocation, as discussed later in the section and explicitly formalized by the GLLSQ collocation condition, are implied as well as an approach that is distinct from the usual linear least-squares (LLSQ) ones where sampling and discretization are introduced. Connections to GC are also implied.

Both LLSQ and GLLSQ approaches minimize some cost function $\Phi' = \|v - w\|^2$, where for the problems of interest $v(\vec{X})$ is a point source potential model and $w(\vec{X})$ is some given canonical (or truth) reference potential. For a point mass fit with N_k point masses, v has the form

$$v(\vec{X}) = G \sum_{k=1}^{N_k} \frac{m_k}{|\vec{X} - \vec{X}'_k|} \ , \tag{2.1}$$

where G is the Newtonian gravitational constant $\approx 6.6742 \times 10^{-11}\mathrm{m}^3\mathrm{s}^{-2}\mathrm{kg}^{-1}$ [10, p. 3]. As previously noted, $\vec{X} \in \Omega_j$ and $\vec{X}'_k \in \Omega_{S_j}$, which is a bounded and closed subregion of the open region Ω'_j, so that the (kernel) basis functions occurring in (2.1) are always bounded. Further $\vec{X}'_{k'} \neq \vec{X}'_k$ for all $k' \neq k$ is always assumed.

Five conventions are adopted here. First, in physics texts the potential function v is interpreted as potential energy, and (2.1) has a negative sign since all gravitational bodies attract and the resulting force is given by the negative of the gradient of the potential. In geophysics these sign conventions are different and consistent with (2.1), but in either case this should cause little difficulty since in fitting problems all that is required is that the overall sign conventions for v and w be consistent. Second, it is assumed that gravitational force is always acting on a unit test mass [10, p. 4], and thus it will be treated as having the units of acceleration [10, p. 45]. Third, physical geodesists distinguish between gravity field quantities, which include the effects of the Earth's rotation, and gravitational quantities like (2.1), which do not [10, p. 44]. This is a distinction physicists generally do not make, since rotational effects can be easily tracked and accounted for as required. The physicist's lead is followed here, and this distinction is ignored. Fourth, both positive and negative masses will be considered a possibility, since this is the usual convention adopted in point mass fitting approaches. Specifically, for gravity modeling and estimation problems each m_k can clearly be viewed as a mathematical parameter that can assume either sign. This convention also allows for the ready adaptation of material developed here to other areas where both signs can occur. (Even regional geoexploration inverse mass density estimation problems can be handled by assuming that all smoothed density estimates are with respect to an average or ambient density.) Fifth, it is useful to introduce scaled versions of the above potential functions in order to absorb the factor of G: $V = v/G$ and $W = w/G$. Thus the cost function to be minimized becomes (with $\Phi := \Phi'/G^2$)

$$\Phi = \|V - W\|^2 = \|V\|^2 - 2\,(V, W) + \|W\|^2 \ . \tag{2.2}$$

(Notice that scaling a cost function leaves the minima unchanged.)

Next consider the philosophy behind the norm selection process. As discussed later in section 5, the minimization philosophy of matching the observations as closely as possible has generally been chosen for point mass fitting problems. This philosophy is, however, not necessarily sound in all or even most cases. For modeling problems a reference model, which is assumed to be accurate, is given, and one wishes to match this reference as closely as possible in some physical sense. Here the desire is to minimize the possible error differences that will result when this given reference model is replaced by a new point mass (or point source) model, which invariably occurs in some sort of software emulation of a physical situation. Thus instead of "matching the observables as closely as possible," a sounder strategy is to "minimize the type of errors that will lead to the greatest errors in the end product." From Newton's second law, since these end-product errors here are most often the direct result of gravity errors it is clear that the difference in the given gravity reference field and the developed point mass gravity model should be minimized, say in a squared residual sense at a large number of appropriate sample points. As this distribution of sample points becomes uniformly dense over the entire global region of interest, the following key integral condition results:

$$\text{Minimize } \Phi = \iiint_{\Omega_j} |\vec{\nabla} V - \vec{\nabla} W|^2 \, dV \; = \; \mathrm{D}[W-V,\, W-V,\, 1,\, \Omega_j] \; . \tag{2.3}$$

Temporarily leaving aside fundamental issues, such as how to turn the RHS of (2.3) into a proper norm structure, consider the general form of the linear equation sets that result from minimizing this type of cost function. For concreteness consider the minimization process in $\mathbb{R}^3$ half-space (Ω_1). Since the RHS of (2.3) is already proportional to the field energy, it is natural to consider the half-space ($j = 1$) energy norm:

$$\|V - W\|^2_{\mathrm{E}_1} := \frac{1}{8\pi} \iiint_{\Omega_1} |\vec{\nabla} V - \vec{\nabla} W|^2 \; dV \tag{2.4}$$

(a factor of 8π has been inserted since it often occurs for various field energy expressions in appropriate units). Because $\|V - W\|^2_{\mathrm{E}_1} = \|V\|^2_{\mathrm{E}_1} + \|W\|^2_{\mathrm{E}_1} - 2(V,\, W)_{\mathrm{E}_1}$, the energy inner-product

$$(V,\, W)_{\mathrm{E}_1} := \mathrm{D}[V,\, W,\, 1,\, \Omega_1]/8\pi$$

is also needed. In particular, if V is specified through (2.1), with $\ell_k := |\vec{X} - \vec{X}'_k|$, and if W is an appropriate reference field, then

$$\|V - W\|^2_{\mathrm{E}_1} = \|W\|^2_{\mathrm{E}_1} - 2\sum_{k=1}^{N_k} m_k (\ell_k^{-1},\, W)_{\mathrm{E}_1} + \sum_{k=1}^{N_k}\sum_{k'=1}^{N_k} m_k\, m_{k'} (\ell_k^{-1}, \ell_{k'}^{-1})_{\mathrm{E}_1} . \tag{2.5}$$

Taking the partial of (2.5) with respect to $m_{k''}$ (for $k'' = 1,\, 2,\, 3,\, \ldots,\, N_k$), setting the result to zero, then dividing by two yields a linear equation set that can be easily inverted for the mass values, provided that $(\ell_k^{-1},\, \phi)_{\mathrm{E}_1}$ can be easily computed for $\phi = W$ and $\phi = 1/\ell_{k'}$. The relationship which makes this possible is (1.1). By introducing $T_{k,k'} = (\ell_k^{-1}, \ell_{k'}^{-1})_{\mathrm{E}_1}$ and $A_k = (W, \ell_k^{-1})_{\mathrm{E}_1}$, the linear equation set can be written as

$$\sum_{k'=1}^{N_k} T_{k,k'}\, m_{k'} = A_k . \tag{2.6}$$

For the spherical exterior, matters proceed in much the same fashion except that a weight function, $\mu_0 = 1/r$, must be introduced into (2.3). Not only is this weighting required to turn ℓ_k^{-1} into a replicating kernel, but since regions closer to the Earth's surface are normally of greater interest for geophysical applications than regions further away, it is also desirable.

Obviously with regards to applications, (2.6) is pivotal and, as such, warrants at least an informal examination. Before proceeding it is useful to clarify the differences between reproducing, replication, and generalized collocation kernels. When a kernel, like ℓ_k^{-1}, is not symmetric since its arguments are in different domains and there is a relationship such as (1.1) or (1.2a) that allows for closed-form inner-product expressions, it will be called a replication kernel and be said to have the point replication property. Since reproducing kernels are necessarily symmetric they cannot be considered replicating kernels, so this terminology distinguishes replication and reproducing kernels. Alternatively, the term *generalized collocation property* will be taken to be a generalization of a point data and/or collocation matching condition, and as such includes the possibility of not only reproducing kernels but also replicating kernels. As discussed below, it also allows for the possibility that resulting inner products may be obtainable by numerical means (after assuming that an underlying replication property also holds)—so long as the resulting inner products, A_k, occurring on the RHS of (2.6) can be matched. (The A_k's may also represent empirically obtained data.)

As a concrete example of situations where numerical integration frequently enters, consider fits based on the continuous analogue of (2.1) where the potential is due to some parameterized density function ρ:

$$V(\vec{X}) = \iiint_{\Omega_S} \frac{\rho(\vec{X}', \alpha)}{|\vec{X} - \vec{X}'|}\, dV'. \tag{2.7}$$

Here $\alpha = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{N_k})^T$. If ρ consists of a linear superposition of density basis functions ψ_k (i.e., $\rho = \sum_{k=1}^{N_k} \alpha_k \psi_k(\vec{X}')$), then minimizing $\|W - V\|^2$ yields a linear equation set similar to (2.6), where the A_k's and $T_{k,k'}$'s must generally be computed numerically; however, when (1.1) or (1.2a) is used, then great simplifications result and continuous distributions are tractable. Besides parameterized volume distributions, parameterized surface and line distributions are also obviously possible. One interesting choice for volume density basis functions ψ_k is the use of finite element method (FEM) basis functions. Thus consider the case where $\vec{q}_k'$ are taken to be a set of node points over Ω_S and where the $\psi_k(\vec{X}')$ are chosen to be a set of localized FEM basis functions with the property that $\psi_{k'}(\vec{q}_k') = 1$ if $k' = k$ and $\psi_{k'}(\vec{q}_k') = 0$ otherwise. The α_k determined directly from the analogue of (2.6) then represent the density strengths at the node points $\vec{q}_k'$.

Any additional structure that can help to clarify the possibilities inherent in (2.6) is desirable. Toward that end, briefly consider a suggestive theorem from RKHS theory. When a reproducing kernel, with specified kernel points $\vec{Q}_k \in \Omega$ replacing the values of $\vec{X}_k'$ in (2.1), is used for a basis functions expansion of V that is analogous to (2.1), and when (2.2) is replaced by the corresponding cost function based on the reproducing kernel norm, then minimization of this cost function results in a closed-form linear equation set that is just like (2.6)—except that the inner products corresponding to A_k take on the simpler form $W(\vec{Q}_k)$. This type of reproducing kernel fit also satisfies a minimum norm collocation property: the function with the smallest associated norm that matches the prescribed data set (i.e., the values $W(\vec{Q}_k)$) is the one which results from solving the analogue of (2.6) [19, pp. 207–220]. This minimum norm property is a well-known functional analysis result [17], and it insures that a reproducing kernel fit will simultaneously match the given point data and minimize both $\|V\|$ and $\|V - W\|$ for the associated norm. This fact is of interest here since it strongly suggests that if a replicating kernel expansion for V is used in (2.6), then generally any specified values for A_k are recovered and that D[V, V, 1, Ω_1] or D[V, V, μ_0, Ω_0] also is simultaneously minimized. While this may fail to happen due to auxiliary restrictions placed on the basis functions or on the overall space of admissible functions, the main way that it can fail to happen is if the kernel basis functions themselves are not linearly independent. These possibilities are not usually addressed in connection with general discussions of the minimum norm collocation property, but in many settings linear independence may not always be transparent, especially if combinations resulting from linear operators acting on kernel basis functions are allowed. (These possibilities can be seen from, among other things, the consequences of the fact that various restricted classes of functions, such as polynomials of fixed degree, may have a reproducing kernel.) When (2.6) is invertible the source parameters (m_k) are uniquely determined, and in some sense one can say that the solution to the inverse point source problem has been obtained. To preserve and extend these inverse source interpretational possibilities, the conservative stance is adopted here of requiring that all admitted basis function sets be invertible and that the solutions to (2.6) replicate the specified A_k values. This condition is called the generalized linear least squares collocation (GLLSQC) condition. There are two obvious ways to enforce this condition: either on a computational case-by-case basis or by proving general theoretical results about classes of particular basis functions. While it is not comparatively well known outside geophysics, a demonstration exists that shows that point mass basis functions are independent in $\mathbb{R}^n$ for finite $n > 1$ [24]. Thus, in a theoretical sense the GLLSQC condition holds for point mass fits, but on a case-by-case basis some care is generally required in solving (2.6) due to ill-conditioning. (As a matter of practice, either a singular value decomposition or a Householder triangulation algorithm implemented to an appropriate number of significant digits should be used.) Finally, it should be readily apparent that while continuous distributions may well satisfy the GLLSQC condition, counterexamples can be easily constructed, with perhaps the simplest example resulting from the consideration of several concentric homogenous spherical shells.

With regard to the overall functional analysis setting, the standard course taken nowadays is to adopt some type of general Banach space setting (such as one type or another of Sobolev space), where the completeness of Cauchy sequences is presupposed. The limit of sequences of functions composed from basis functions that satisfy the GLLSQC condition may not satisfy it; hence, admitting limits of sequences can lead to unwanted interpretational difficulties here. Obviously, to solve (2.6) it is only necessary that a finite linear span of basis functions be admitted, which requires only an inner-product structure. In accord with the conservative stance outlined above, a structured pre-Hilbert space setting is adopted since a pre-Hilbert space setting presupposes an inner-product structure, but it does not make the usual assumptions about admissible sequences of functions. (This way of specifying functions of interest is somewhat dated, but prior to the mid-1950s it was the prevailing way of addressing Dirichlet inner-product spaces and was used, for example, in [3].) The adjective "structured" here means that further auxiliary conditions are imposed on the class of admissible functions. One such condition is that any set of basis functions considered must satisfy the GLLSQC condition stated above. Another requirement is that all functions, f, must be harmonic over Ω_j (including $\partial\Omega_j$). Additional structure is needed to insure that (weighted) Dirichlet integrals over unbounded domains can be regarded as defining a positive definite norm. This can be accomplished by assuming one (or all) of the following four largely equivalent requirements:

(i) $\mathrm{D}[f,\, f,\, 1,\, \Omega_j]$ is bounded, and f tails off at least as fast as $1/r$ as $r \longrightarrow \infty$.

(ii) The first Sobolev norm of f over Ω_j is bounded:

$$\mathrm{D}[f,\, f,\, 1,\, \Omega_j] + \iiint_{\Omega_j} |f|^2 \, dV < \infty.$$

(iii) f is a potential function generated by a (well-behaved) localized source distribution in $\Omega_{S_j} \subset \Omega_j' \setminus \partial\Omega_j'$.

(iv) A well-behaved series representation for f always exists in the form of a spherical harmonic expansion of f in powers of $1/r$, for the exterior of some sphere contained inside Ω_j'.

Notice that (i) is the historical approach [3] and that (ii) is equivalent to (i) for $\Omega_j = \Omega_0$ since both the integrals occurring in the first Sobolev norm must be separately bounded. Also for all f not identically zero, $0 < \mathrm{D}[f,\, f,\, \mu_0,\, \Omega_0] < \mathrm{D}[f,\, f,\, 1,\, \Omega_0]/R_0$, where positivity can be easily proved using several standard properties of harmonic functions. (Specifically, if $|f| \neq 0$ at some point in Ω_0, then taking the line integral of $\vec{\nabla} f$ from this point to a point at infinity, one can infer that $|\vec{\nabla} f| > 0$ for at least one point along this line. Then from the mean-value theorem and the maximum-modulus theorem, $\vec{d} \cdot \vec{\nabla} f > 0$ must occur throughout some neighborhood near $\partial\Omega_0$ for one fixed direction $\vec{d}$ or another, and from this it follows immediately that $\mathrm{D}[f,\, f,\, \mu_0,\, \Omega_0] > 0$.) It is clear that (iii) and (iv) are largely equivalent to (i) or (ii) (although (iii) and (iv) do not require harmonicity as a separate condition). Requirement (iii) is more or less tantamount to the assumption that a collection of point sources is being modeled.

3. Half-space ($\boldsymbol{\Omega_1}$) relationships. For the class of admissible functions just described with bounded source region, the following half-space analogue of Poisson's solution to the Dirichlet boundary value problem holds [5, p. 268]:

$$W(\vec{X}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{W(x'', y'', 0)\, z \; dx''\, dy''}{[(x - x'')^2 + (y - y'')^2 + z^2]^{3/2}}, \tag{3.1}$$

where $z > 0$ and $\vec{X}'' \in \partial\Omega_1$. Using the facts that the surface integral at infinity is zero for the unbounded region Ω_1, that $\partial/\partial n := -\partial/\partial z$, and that $dS = dx\, dy$ for the boundary plane ($\partial\Omega_1$) and applying Green's first identity[1] yields

$$\begin{aligned}(\ell_k^{-1},\, W)_{\mathrm{E}_1} &= \frac{1}{8\pi} \iiint_{\Omega_1} \vec{\nabla}\ell_k^{-1} \cdot \vec{\nabla} W \; dV \\ &= -\frac{1}{8\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[W \frac{\partial}{\partial z} \frac{1}{|\vec{X} - \vec{X}_k'|} \right] \Bigg|_{z=0} dx\, dy.\end{aligned} \tag{3.2}$$

Next observe that

$$\left[\frac{\partial}{\partial z} \frac{1}{|\vec{X} - \vec{X}_k'|} \right] \Bigg|_{z=0} = \frac{z_k'}{[(x - x_k')^2 + (y - y_k')^2 + (z_k')^2]^{3/2}}\ , \tag{3.3}$$

where $z_k' < 0$. Combining (3.1), (3.2), and (3.3) produces

$$(\ell_k^{-1},\, W)_{\mathrm{E}_1} = W(x_k',\, y_k',\, -z_k')/4, \tag{3.4}$$

[1] $\iiint_\Omega (\phi \nabla^2 \psi + \vec{\nabla}\psi \cdot \vec{\nabla}\phi)\; dV = \iint_{\partial\Omega} \phi \frac{\partial \psi}{\partial n}\, dS.$

Table 3.1
Point data/source correspondences for $\mathbb{R}^3$ half-space.

Observation type	Source type	Invariance property
Potential	Point mass	Scalar
Gravity	Point dipole	Vector
Gravity gradient	Point quadrupole	2nd rank tensor

which shows (1.1). This relation can be used to evaluate the terms $T_{k,k'}$ and A_k occurring in (2.6):

$$T_{k,k'} = \frac{1}{4\sqrt{(x'_k - x'_{k'})^2 + (y'_k - y'_{k'})^2 + (z'_k + z'_{k'})^2}}\,, \quad A_k = \frac{W(x'_k,\, y'_k,\, |z'_k|)}{4}\,. \tag{3.5}$$

Two final points are relevant. First, dipole and other higher order multipoles can also be easily fit. In general, potentials for these point sources can be written as $\sum_k \sum_i S_{ki}\mathcal{L}_{ki}({\ell_k}^{-1})$, where S_{ki} are the associated source strengths and the $\mathcal{L}_{ki}$ are appropriate linear differential operators that can be expressed as a sum of partials of various orders with respect to x, y, or z. For example, an electrostatic or point mass dipole term is proportional to $\vec{D}_k \cdot \vec{\nabla}\ell_k^{-1}$. Since $\vec{\nabla}\ell_k^{-1} = -\vec{\nabla}'_k \ell_k^{-1}$, where $\vec{\nabla}'_k := (\partial/\partial x'_k,\, \partial/\partial y'_k,\, \partial/\partial z'_k)^T$, and components of $\vec{X}'_k$ serve only as parameters when they occur inside the inner product here, all $\vec{X}$ dependent derivative factors operating on ℓ_k^{-1} that occur inside inner-products can be replaced with $\vec{X}'_k$ derivative factors; hence, these differential operators can be moved inside or outside the inner-products entirely as desired so that all required inner products for $T_{k,k'}$ and A_k can be easily evaluated in closed form. For half-space, these possibilities and the associated measurable point quantities are displayed in Table 3.1. Analogous possibilities exist for the spherical exterior case.

Second, to improve the condition number of the matrix $\boldsymbol{T}$ in (2.6), it is often desirable to use normalized basis functions $\hat{\varphi}_k$ in place of ℓ_k^{-1}:

$$\hat{\varphi}_k := \frac{\widetilde{N}_k}{\ell_k}\,, \quad \text{where} \quad \widetilde{N}_k := \frac{1}{\|\ell_k^{-1}\|} \tag{3.6}$$

in the general norm setting. Introducing $\widetilde{T}_{k,k'} := (\hat{\varphi}_k,\, \hat{\varphi}_{k'}) = \widetilde{N}_k\, \widetilde{N}_{k'}\, T_{k,k'}$, $\tilde{A}_k := (W,\, \hat{\varphi}_k) = \widetilde{N}_k\, A_k$, and $\widetilde{m}_k := m_k/\widetilde{N}_k$ allows (3.5) to be reexpressed as

$$\sum_{k'=1}^{N_k} \widetilde{T}_{k,k'}\, \widetilde{m}_{k'} = \tilde{A}_k. \tag{3.7}$$

For the $\mathbb{R}^3$ half-space energy norm

$$\begin{aligned} \widetilde{N}_k &= 2^{\frac{3}{2}}\sqrt{|z'_k|}\,, \\ \widetilde{T}_{k,k'} &= \frac{2\sqrt{z'_k z'_{k'}}}{\sqrt{(x'_k - x'_{k'})^2 + (y'_k - y'_{k'})^2 + (z'_k + z'_{k'})^2}}\,, \\ \tilde{A}_k &= \frac{\sqrt{|z'_k|}}{\sqrt{2}}\, W(x'_k\; y'_k,\, -z'_k)\,, \end{aligned} \tag{3.8}$$

so that (3.7) can easily be inverted to determine the values of $\widetilde{m}_k$ and thus m_k. While the use of normalized basis functions is not always required for point mass fits, their use in mixed type point source fits is always highly recommended due to the diverse associated physical scales that occur there and the attendant large condition numbers for $\boldsymbol{T}$. (An experiment was performed on the results from one of the global NLLSQ combined point mass/dipole fits discussed in section 5. At the specified source locations a linear fit was done both with and without normalized basis functions. The ratio of the two resulting condition numbers was over 10^{20}.)

4. Spherical exterior ($\boldsymbol{\Omega_0}$) relationships. The goal of this section is the derivation of the two relevant spherical exterior DIDACKS relationships, (1.2a) and (1.3a). Matters are more complex for this case than they were for the half-space case, but all of the supporting issues raised in sections 2 and 3 can obviously be carried over here, so they are not repeated.

Let f and g be two admissible functions as discussed in section 2 and consider the integral norm [21]:

$$(f,\, g)_{\mathrm{I}} := -\frac{R_0^2}{4\pi}\iint_\sigma \mathcal{D}_r(rf\,g)\;d\,\sigma = -\frac{R_0^2}{4\pi}\iint_\sigma \big[\mathcal{D}_r(f\,r\,g)\big]\Big|_{r=R_0} d\,\sigma, \tag{4.1}$$

where $\mathcal{D}_r := \frac{\partial}{\partial\, r}$. The last expression on the RHS here follows from the evaluation convention of (1.4). The label "integral norm" was chosen by the author since, as discussed below, the integrals required for point mass fitting in Ω_0 can be evaluated in closed form and there is little chance of confusing this norm with the usual norm of square integrable functions.

Applying Green's first identity to the Ω_0 energy inner-product and noting that the surface integral at infinity vanishes, while $dS = R_0^2\, d\,\sigma$ and $\partial\,/\partial\, n := -\mathcal{D}_r$ for the bounding inner exterior surface of Ω_0, yields

$$\begin{aligned}(f,\, g)_{\mathrm{E}_0} &:= \frac{1}{8\pi}\iiint_{\Omega_0} \vec{\nabla} f\cdot\vec{\nabla} g\;dV \\ &= -\frac{R_0^2}{8\pi}\iint_\sigma g\,\mathcal{D}_r f\;d\,\sigma = -\frac{R_0^2}{16\pi}\iint_\sigma \big[\mathcal{D}_r(fg)\big]\Big|_{r=R_0} d\,\sigma.\end{aligned} \tag{4.2}$$

From (4.1) and (4.2) it follows immediately that

$$(f,\, g)_{\mathrm{I}} = 4\,R_0(f,\, g)_{\mathrm{E}_0} - R_0^2(f,\, g)_\sigma. \tag{4.3}$$

After the second DIDACKS relation, (1.2a), is addressed, it will be shown that the integral norm is proportional to the weighted Dirichlet integral, which will complete the proof of the third DIDACKS relation, (1.3a).

First, observe that the following two equations can be shown through a relatively straightforward evaluation of their respective LHS and RHSs:

$$\left[\frac{r}{\ell_k}\right]\Bigg|_{r=R_0} = \frac{R_0^2}{r_k'}\left[\frac{1}{|\vec{X}-\vec{P}_k|}\right]\Bigg|_{r=R_0} \tag{4.4}$$

and

$$\left[\mathcal{D}_r\left(\frac{r}{\ell_k}\right)\right]\Bigg|_{r=R_0} = -\frac{R_0^2}{r_k'}\left[\mathcal{D}_r\,\frac{1}{|\vec{X}-\vec{P}_k|}\right]\Bigg|_{r=R_0}, \tag{4.5}$$

where $\vec{P}_k$ is given by (1.2b) with $\vec{X}'_k = \vec{X}'$ and $r'_k := |\vec{X}'_k|$. Employing (4.4) and (4.5) yields

$$\begin{aligned}(4.6)\qquad -\frac{R_0^2}{4\pi}\iint_\sigma \left[\mathcal{D}_r\left(\frac{rW}{\ell_k}\right)\right]\bigg|_{r=R_0} d\sigma \\ = -\frac{R_0^4}{4\pi r'_k}\iint_\sigma \left[\frac{(\mathcal{D}_r W)}{|\vec{X}-\vec{P}_k|} - W\,\mathcal{D}_r\,\frac{1}{|\vec{X}-\vec{P}_k|}\right]\bigg|_{r=R_0} d\sigma.\end{aligned}$$

Recall that W is harmonic for $r > R_0$. Applying Green's second identity[2] to (4.6) yields

$$-\frac{R_0^2}{4\pi}\iint_\sigma \left[\mathcal{D}_r\left(\frac{rW}{\ell_k}\right)\right]\bigg|_{r=R_0} d\sigma = -\frac{R_0^2}{4\pi r'_k}\iiint_{\Omega_0} W\,\nabla^2\left(\frac{1}{|\vec{X}-\vec{P}_k|}\right) dV.$$

Then using $\nabla^2(1/|\vec{X}-\vec{P}_k|) = -4\pi\delta(\vec{X}-\vec{P}_k)$, where δ is the Dirac delta function [11, p. 35], gives

$$(4.7)\qquad -\frac{R_0^2}{4\pi}\iint_\sigma \left[\mathcal{D}_r\left(\frac{rW}{\ell_k}\right)\right]\bigg|_{r=R_0} d\sigma = \frac{R_0^2}{r'_k} W\left(\vec{P}_k\right)$$

or finally, with $P_k = |\vec{P}_k|$:

$$(4.8)\qquad \left(\ell_k^{-1},\, W\right)_{\mathrm{I}} = P_k\, W\left(\vec{P}_k\right).$$

Various other ways exist to prove (4.8). One way is to substitute spherical harmonic expansions for W and ℓ_k^{-1} into the LHS of (4.7). A second way is to first observe that $(\ell_k^{-1}, \ell_{k'}^{-1})_{\mathrm{I}}$ can be evaluated in closed form by using spherical coordinates on the LHS of (4.7). Next, this result can be generalized by substituting the integral form of Poisson's equation (not Poisson's integral) for W into the LHS of (4.8) and then reintroducing the closed-form expression just obtained for $(\ell_k^{-1}, \ell_{k'}^{-1})_{\mathrm{I}}$, which results in the RHS of (4.8) reexpressed in terms of the integral form of Poisson's equation. As previously noted, these various steps were the ones originally followed by the author and account for the nomenclature. Other proofs have also been discovered.

This leaves the proof of the relationship between the weighted Dirichlet integral and the integral norm. Recalling the limit convention implicit for integration over σ, (1.4), and expanding the RHS of (4.1) yields

$$(4.9)\qquad (f,\, g)_{\mathrm{I}} = -\frac{R_0^2}{4\pi}\iint_\sigma f\,g\;d\sigma - \frac{R_0^2}{4\pi}\iint_\sigma r\mathcal{D}_r(f\,g)\;d\sigma.$$

Next temporarily ignore the common factor of $R_0^2/4\pi$ and consider the two integrals on the RHS of (4.9). Using the fact that $\mathcal{D}_r f = (\vec{X}/r)\cdot\vec{\nabla} f$, the first integral on the RHS of (4.9) can be written as

$$\begin{aligned}(4.10)\qquad -\iint_\sigma f\,g\,d\sigma &= \iint_\sigma \left\{\int_{r=R_0}^\infty \mathcal{D}_r\,(f\,g)\,dr\right\} d\sigma = \iiint_{\Omega_0} r^{-2}\,\mathcal{D}_r\,(f\,g)\;dV \\ &= \iiint_{\Omega_0} (g\,r^{-3})\,(\vec{X}\cdot\vec{\nabla} f)\;dV + \iiint_{\Omega_0} (f\,r^{-3})\,(\vec{X}\cdot\vec{\nabla} g)\;dV.\end{aligned}$$

[2] $\iint_{\partial\Omega}\left(\phi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\phi}{\partial n}\right) dS = \iiint_\Omega(\phi\nabla^2\psi - \psi\nabla^2\phi)\;dV.$

Green's second identity in the following form will be useful in reexpressing the second term on the RHS of (4.9):

$$\iiint_{\Omega_0} \vec{\nabla}\psi\cdot\vec{\nabla}\phi \; dV = -\iint_{\sigma} r^2\,\psi\,(\mathcal{D}_r\phi)\,d\sigma, \tag{4.11}$$

where ϕ is harmonic in Ω_0, but ψ need not be, and both must vanish sufficiently fast as $r \longrightarrow \infty$. Using this identity at the appropriate place twice in the following expression (first with $\psi = f/r$ and $\phi = g$ and then with f and g reversed), the second term on the RHS of (4.9) can be rewritten as

$$\begin{aligned} &-\iint_{\sigma} r f\,(\mathcal{D}_r g)\,d\,\sigma - \iint_{\sigma} r\,g\,(\mathcal{D}_r f)\,d\,\sigma \\ &\quad = \iiint_{\Omega_0} \vec{\nabla} g\cdot\vec{\nabla}\,(f/r)\;\; dV + \iiint_{\Omega_0} \vec{\nabla} f\cdot\vec{\nabla}\,(g/r)\;\; dV. \end{aligned} \tag{4.12}$$

Finally using $\vec{\nabla}(f/r) = r^{-1}\vec{\nabla} f - r^{-3}\vec{X} f$ in (4.12) and substituting this result along with (4.10) into (4.9) produces

$$(f,\, g)_{\mathrm{I}} = \frac{R_0^2}{2\pi}\iiint_{\Omega_0} r^{-1}\,\vec{\nabla} f\cdot\vec{\nabla} g \; dV. \tag{4.13}$$

With the aid of (4.3), (4.8), and (4.13), the second and third DIDACKS relationship (i.e., (1.2a) and (1.3a)) follow immediately. From (4.13) and the discussion at the end of section 2 the integral norm is positive definite.

DIDACKS dipole and other higher order multipole implementations differ for the half-space and spherical settings in one significant way. While closed-form expressions for all the required inner-products for an integral-norm point-mass fit can be evaluated just as discussed in section 3, for higher order multipole fits all the derivatives of the potential for all the lower orders are also required in the spherical case because taking partials with respect to the components of $\vec{X}'_k$ yields additional terms on the RHS of (4.8). This means, for example, that a dipole fit requires not only point gravity information, but point potential information as well. It is thus natural in this case to perform not only a point dipole fit, but a combined point mass/dipole fit. With this understanding the spherical case analogue of Table 3.1 is readily obtained.

When an integral norm higher order multipole fit is desired and the lower order derivatives of W are not available, then it still may be possible to do the fit [21]. For example, if a dipole fit is desired and information about W is not available, but the gradient of W is known along various intersecting lines, then potential data in the form $W(\vec{X}) = W_0 + \delta W(\vec{X})$, where W_0 is an unknown constant, can be assumed anywhere along these lines since the form of $\delta W(\vec{X})$ can be found by numerical line integration. For any assumed trial value of W_0, a DIDACKS fit can be performed. The results of this fit can then, in turn, be substituted into a standard LLSQ type of cost function that is the analogue of (5.1) and is based on minimizing gravity computations at various sample points. If gravity data itself is plentiful, then an outer-loop optimization process can be based on minimizing this new cost function, where W_0 is treated as an unknown NLLSQ parameter. In this outer-loop process sample point gravity differences are minimized throughout the fit region of interest (which may be only a small part of Ω_0). When its choice is not obvious, R_0 can also be treated as a parameter and optimized in the same fashion.

5. Broader point mass fitting context. The goal of this section is to illuminate certain aspects of relevant geophysics and point mass fitting background material by way of a brief synopsis (as such, this section is somewhat subjective). Before considering the associated geophysical context it is useful to clarify the distinction between NLLSQ and GLLSQ or LLSQ problems and to briefly consider how the pertinent DIDACKS applications history fits into these categories. In what follows, the distinction between GLLSQ and LLSQ is generally dropped and only the acronym LLSQ is used, so that both classes can be referred to jointly. If all the source locations are known, then linear equation sets result for the source strengths, and in the DIDACKS approach these linear equation sets are exact due to (1.1) and (1.2a). Alternatively, if both the locations and strengths are to be determined, then an NLLSQ problem results since source locations enter as nonquadratic parameters in the cost function. While only LLSQ problems have been addressed so far in this article, accurate low degree and order spherical harmonic (tesseral) NLLSQ fits have been obtained by the author, and this was, in fact, the first area of DIDACKS applications in the early 1980s [21]. These NLLSQ fits are discussed further below. While rather varied approaches have been used by different researchers for point mass–based geoexploration inverse source applications, for the gravity modeling and estimation problems dealt with here far fewer point mass–based approaches have been employed; nevertheless, not only have the associated research efforts been internationally diverse, but the corresponding literature is also extensive, so that only a small part of it can be considered here (see [26] for additional history and references). LLSQ and NLLSQ point mass gravity models can also be produced to serve as synthetic gravity models with realistic attributes, but DIDACKS applications in this area have not been considered, so these applications are not discussed below. Finally, two conventions are adopted in the remainder: (a) a spherical harmonic expansion to degree and order N will be called an $N \times N$ field or expansion, and (b) both field modeling and estimation problems will be referred to as field reconstruction problems.

Next, consider the relevant geophysical aspects. Approximately a quarter of a century ago one could divide the Earth's gravity field into three parts corresponding to their respective data sources:

(I) Global spherical harmonic field data derived directly from satellite tracking data and historically considered accurate to around the degree and order 8 to 12 range. Here this part of the field is called "low degree and order," and it is taken to be 12×12 and below.

(II) An intermediate field taken here to be the part above 12×12 and below 120×120, which before the evolution of more advanced radar equipped satellites could not be accurately determined.

(III) Regional measurements of geophysical surface quantities known as gravity anomaly and vertical deflections. (These data sets contain part (I) and (II) contributions unless they are factored out.)

Parts (I) and (II) together will be taken as comprising the global part of the gravity field. As discussed below, much higher accuracy is desirable for part (I) than part (II) or part (III) data. Currently very accurate spherical harmonic expansions to a much higher degree and order are available on the Internet, so the distinction between the above three data sets is not as distinct as it once was, but to understand the context of the methods discussed in this section it is useful to keep this historical data partition in mind. (Expansions beyond 360×360 are common, and the recent Gravity Recovery and Climate Experiment (GRACE) [10, 20] has already established new global gravity accuracy benchmarks up through 110×110.)

Next consider the origins of part (III) data. Regional data measurements are often made at sea, and it is in this context that the concepts of gravity anomaly and vertical deflection components are best understood. Suppose that the Earth were composed of a homogenous liquid with the same overall mass and volume; then due to rotational effects it would take on an ellipsoidal shape. Geophysicists set up a mathematical model of this configuration that they call the reference ellipsoid. The associated field is called the normal gravity field and is the predominant part of the field. If transient effects (like tides and ocean waves) are accounted for, then under the influence of gravity the oceans form an equal-potential surface (otherwise water would flow from one part to another until an equilibrium was reestablished). If normal gravity is subtracted off and if these transient effects are properly taken into account, then the measurement taken by a gravimeter on a ship is called the gravity anomaly, Δg, and it is a scalar since the measurement is taken along a plumb line. The angular displacement of this plumb line from the vertical is called vertical deflection, which can be resolved into north-south and east-west components. Due to the equipotential effect just noted, this measurement is displaced from the specified reference ellipsoid by a distance that is called the geoid height, but the measurement is recorded as if it had been made at a point on the reference ellipsoid itself [10]. Geoid height and scalar potential values are connected by Bruns formula [10]. Geoid height itself can be determined by a radar-equipped satellite that has a known location. Gravity anomaly is typically measured in units of milligal, where 1 milligal $= 1 \times 10^{-5}\text{m/s}^2$. A 1 milligal error is considered more-or-less acceptable for regional gravity anomaly data processing [10, p. 274], and various underpinning geophysical relationships are derived with an inherent approximation consistent with this 1 milligal requirement [9, 10]. Much higher accuracy is desirable for part (I) than part (II) or part (III) data since gravitational effects are cumulative for most uses and since low degree and order errors tend not to cancel out. GRACE and other modern fields have error levels considerably under a milligal for part (I) data, and for this part of the field it is also desirable to have point mass models with errors somewhat under a milligal.

Historically, there have been three primary motivations for performing NLLSQ fits to the low degree and order spherical harmonic part of the Earth's gravity field: (a) to find a more efficient computational scheme for gravity evaluations, (b) to gain some insight into the distribution of matter in the Earth's interior, (c) to conduct goal-oriented pure research. Due to the computational ease and speed of low degree and order spherical harmonic gravity evaluations that has resulted from computer hardware and software advances, (a) has long ceased to be a realistic reason for using point mass fits, and this point is totally irrelevant now. Other measurement programs and advances have also emerged to address the issues raised by (b), but in fact it was never clear that the small number of point masses used in NLLSQ fits could provide truly significant mantle or deep core density information, which leaves only (c). NLLSQ point mass fitting presents a very challenging problem that can conceivably serve as a sort of test bed for developing techniques to tackle other ill-conditioned problematic NLLSQ problems; moreover, aside from these NLLSQ aspects, it is an intrinsically interesting potential theory problem that has associated cross-fertilization possibilities.

The first step in attacking any LLSQ or NLLSQ problem is to set up a cost function. A commonly chosen minimization philosophy for the NLLSQ point mass fitting problem is that of matching the observed quantities as closely as possible. Since there is a classical result in geophysics (Stokes' integral) that says that if the gravity anomaly is known over the entire reference ellipsoid, then the external field quantities can be reconstructed, standard approaches to performing global NLLSQ low degree and order point mass fits often have been based on matching the gravity anomaly at N_i different sample points, $\vec{X}_i$, specified on the reference surface. This technique will be called the "classical point mass" approach. If $\Delta g_{PM}(\vec{X}_i)$ denotes the anomaly generated by a collection of N_k point masses and $\Delta g_{ref}(\vec{X}_i)$ represents the truth anomaly value at the same point, this classical NLLSQ point mass approach can be framed through the requirement that the following cost function be minimized:

$$\Phi = \sum_{i=1}^{N_i} \left(\Delta g_{PM}(\vec{X}_i) - \Delta g_{ref}(\vec{X}_i) \right)^2 , \tag{5.1}$$

where $N_i \gg N_k$. The usual prescribed number of point masses, N_k, is approximately 50 for an NLLSQ 9×9 fit, while $N_k \approx 80$ for a 12×12 fit. The resulting gravity anomaly error standard deviations (sigma) achieved by applying this classical point mass fitting approach has typically been from 3/4 to several milligals. These approaches normally have a much more sizable error in the smaller degree and order terms than is desirable.

The global NLLSQ point mass fitting problem itself is inherently nonlinear, and the point masses must be quite deep to obtain good results, which heightens numerical difficulties and associated convergence problems. Given this state of affairs, coupled with the facts that the smaller degree and order errors cannot be easily removed using this classical approach and that it contains inherent sampling and discretization error, there is an error floor of around 2/3 milligal that these approaches historically have not been able to overcome. While the associated regional LLSQ approaches that have been tried are remarkably diverse, far fewer low degree and order NLLSQ approaches have been employed—nevertheless the diversity of attempted NLLSQ point mass approaches in the literature is much wider than the above discussion indicates, but still the accuracy levels and deficiencies of the classical point mass approach presented above are thought to be representative of these other existing NLLSQ attempts as well. In section 2 it was argued that the usual philosophy of matching measured quantities is not the correct one and that an energy basis or weighted energy basis is clearly called for in approaching both LLSQ and NLLSQ gravity modeling problems.

As noted above, accurate global NLLSQ low degree and order spherical harmonic point mass fits using the DIDACKS approach were first obtained almost a quarter of a century ago by the author, and in the past as better spherical harmonic data has became available more accurate fits have been obtained [21]. This has been an ongoing effort, and the current NLLSQ 50 point mass fit to the 9×9 part of a recent field has a sigma error of about 0.035 milligal, while the corresponding error in an NLLSQ 80 point mass fit to the 12×12 part of the field is 0.030 milligal. As one might expect, additional masses here can offer marked improvements in accuracy but at a loss of efficiency. Combined point mass/dipole fits (cf. section 4), which are combinations of point masses and dipoles at the same location, have also been performed over the years [21]. Only about half as many of these combined point sources are required for an accurate NLLSQ fit: 22 for a 9×9 field and 35 for a 12×12 field. These NLLSQ DIDACKS fits have all been based on the integral norm introduced in section 4. Besides the basic DIDACKS formalism, these fits also use additional specialized NLLSQ techniques developed by the author to handle the existence of numerous false minima at various physical scales.

While the fits should be feasible, neither LLSQ nor NLLSQ DIDACKS fits to the complete global intermediate part of the field (part (II)) have been attempted, although efficient and accurate DIDACKS fits for various regional and local gravity modeling and estimation problems using both LLSQ and partial NLLSQ implementations have been obtained by the author [21]. Regional point mass gravity field reconstruction (part (III)) based on the "classical point mass" approach epitomized by (5.1) has also often been successfully done over the years by various researchers; however, some have reported disappointing results. This is not surprising since considerable patience is often required in order to obtain the best or even good results with the DIDACKS approach due to source placement issues. Thus while both DIDACKS and the classical point mass fitting approach can be viewed as alternatives to GC for regional applications, they both clearly share the same sensitivity to and dependence on point mass placement. This sensitivity to point mass placement is also apparent from the low degree and order NLLSQ DIDACKS results, since NLLSQ iterations have obviously reduced the errors by several orders of magnitude over those of initial trial configurations. GC is designed to not be as sensitive to the placement of the kernel measurement points (which is the corresponding placement issue for it), and thus GC generally requires less patience and skill. Alternative part (II) and (III) grid-based point mass approaches are referenced and discussed in [26].

It is thus useful to briefly describe GC in order to compare and contrast it with the DIDACKS approach. Unlike DIDACKS, GC theory generally assumes that all available data is used. GC as commonly practiced differs from standard collocation techniques utilized by applied mathematicians in five basic ways:

1. An SRK basis is assumed, and emphasis is primarily focused on a statistical (covariance) interpretation given to these kernels.
2. Laplace's equation in $\mathbb{R}^3$ is always assumed to hold over the region of interest.
3. This field region is assumed to be either half-space or the exterior of a sphere (but the half-space setting is used only occasionally for localized regional distributions, so it is not stressed).
4. The study of kernels is broken down into the standard empirically modeled statistical covariance kernels and analytical collocation (where closed-form kernels are studied)—analytical collocation is generally used only when the functional form of the kernel is understood to be a workable representation of the statistical covariance function.
5. Field measurement errors are allowed.

Practitioners of GC have cataloged most, if not all, useful kernel possibilities allowed by these five differences. The main interpretational basis of GC is the GC or minimum norm property, previously mentioned in connection with the GLLSQC condition in section 2: under the assumption of errorless measurements, from all possible candidate functions that reproduce the given point measurements, GC selects the one that is smoothest—that is, the one with the smallest norm, where the norm is determined by the underlying covariance function [19, pp. 207–220]. It is this minimum norm property, in concert with the emphasis on statistical covariance functions indicated by the first point above, that makes GC approaches less sensitive to kernel point placement issues; however, as indicated above, this also implies a corresponding loss of fitting responsiveness, and thus, for example, it would be hard to argue that GC is capable of the overall level of economy and efficiency indicated above for the NLLSQ DIDACKS low degree and order tesseral fits. Further observe that, just as in RKHS theory, GC also satisfies a least-squares norm property [19], but in practice this property is usually ignored by physical geodesists.

For LLSQ gravity reconstruction problems GC has historically been the first procedure of choice, and for these problems one could argue that when good covariance data is available, these techniques are safe and easy to use; however, as classically practiced, GC techniques do have notable limitations, which are not shared by the DIDACKS approach:

(A) NLLSQ applications cannot be treated.

(B) GC techniques are customarily applicable only to the Earth's gravitational field environment since they require covariance information that is often available only in this context.

(C) The application of GC requires a certain level of familiarity, and thus it is almost never adapted to problems outside the geophysical realm, even when appropriate covariance data can be gathered. (There are, however, various other areas that use somewhat related kernel techniques [22].)

(D) Inverse source estimation problems cannot be entertained.

In summary, for gravity reconstruction problems with accurate selected data sets where either approach can be used, results using the DIDACKS approach can be either much better or much worse than one might normally expect with GC approaches, since the ingenuity, implementation skill, and patience of the practitioner have a much more significant bearing on the outcome for DIDACKS approaches. GC is and probably will always remain the primary mathematical tool employed for raw gravity data processing and related uses where efficiency is not the primary concern, since its behavior in these arenas is well understood and it can handle measurement errors naturally.

6. Mathematical connections to geophysical collocation. As indicated in section 1, (1.2a) can be related to a line of preexisting research performed by Krarup in conjunction with his studies of GC [15]. Also an independent parallel line of point source research exists that maintains direct connections to GC itself and can be considered a direct outgrowth of Krarup's original work. This alternative collocation-based point source scheme is briefly considered after the connections of DIDACKS theory to Krarup's work are addressed. Since this alternative scheme and Krarup's work are primarily based on GC for spherical exteriors, only this geometry will be considered in what follows.

Krarup first introduced the weighted integral occurring on the RHS of (4.13) in conjunction with his study of GC [15, pp. 62–65]. The SRK corresponding to this norm (the "Krarup kernel," K_{κ}) has the form

$$K_{\kappa}(\vec{P},\, \vec{Q}) = \frac{R_0}{\sqrt{R_0^2 - 2\,\vec{P}\cdot\vec{Q} + (P\,Q/R_0)^2}}\ , \tag{6.1}$$

which Krarup and his followers extensively studied and applied. In (6.1), $Q = |\vec{Q}|$. To understand how point mass fitting enters, observe that this SRK can be recast in the form $R_0^2/(P|\vec{Q}-\vec{X}'|)$, or $|\vec{X}'|/|\vec{Q}-\vec{X}'|$ since $\vec{X}' = \vec{P}R_0^2/P^2$, which is proportional to a point mass potential at a fixed location. Making the ansatz $\vec{X}' \longrightarrow \vec{X}'_k$ and $\vec{Q} \longrightarrow \vec{X}$ allows one to transform a collocation fit into a point mass fit if the mass locations are restricted to be at a fixed depth $r'_k = |\vec{X}'_k| =$ constant, where r'_k is simply treated as an overall constant of proportionality. The relevant history of this and related ideas is briefly touched on below. Here one possible next step, which Krarup and his followers did not apparently take since it entails abandoning symmetric kernel forms altogether, is to generalize this procedure to independent variable depths by absorbing the factors

$P_k := R_0^2/r'_k$ and R_0 into each of the collocation fitting parameters separately and then reinterpreting the resulting collocation fit as a point mass fit, which yields a fit based on (1.2a) as the end product.

Conversely, DIDACKS point mass fits based on the integral norm can be reinterpreted as collocation fits [21]. The resulting fits are based on what is called the reciprocal distance covariance function [19, p. 182], which corresponds to R_0^2/PQ times the Krarup kernel specified by (6.1). Notice that while the integral norm can be reexpressed in terms of the "Krarup norm" by (4.13), it is useful to retain the integral norm as a distinct entity specified by (4.1) since (a) it is a surface integral rather than a volume integral, (b) there are some consequences of this form, such as (4.3), that are not apparent from the weighted Dirichlet integral form itself, and (c) the Krarup norm is associated exclusively with the Krarup kernel in SRK form and is primarily linked to GC theory where, as previously noted, the goals and practices are quite different.

As summarized in [16] and as just indicated, starting in the early 1980s a distinct line of point source research based on Krarup's original work above was developed by Marchenko and others (see [7]). This research is based on maintaining connections to symmetric global GC covariance kernel forms. For the spherical case, assuming the usual GC covariance properties, the general form of the allowed global covariance kernel, $C(\vec{P}, \vec{Q})$, can be written as

$$C(\vec{P}, \vec{Q}) = \sum_{n=0}^{\infty} k_n \left(\frac{R_0^2}{PQ}\right)^{n+1} \mathrm{P}_n(\cos\psi)\,, \quad \text{where} \quad \cos\psi = (\vec{P}\cdot\vec{Q}/P\,Q)\,, \tag{6.2}$$

where the P_n are standard (unnormalized) Legendre polynomials and the k_n are constants [19, p. 181]. Here, as before, R_0 is the radius of the spherical region. Mathematically, a kernel specified by (6.2) can, in general, be simply considered an SRK with an added layer of statistical interpretation. Solving (1.2b) for $\vec{X}' = \vec{X}'(\vec{P})$ and substituting the result into (6.2) yields a kernel that is a function of one interior point and one exterior point. For certain applications, when (6.2) is reexpressed in terms of a smaller radius $R_B < R_0$, the resulting sphere is known as a Bjerhammar sphere. For particular choices of k_n if (6.2) can be rewritten as a closed-form expression and if this expression has the correct form, such as (6.1), then the result can be reexpressed as a point mass or other linear combinations of point or line source potentials. As just noted in connection with (6.1), additional parameters will occur in the resulting expressions that do not appear in the source potentials themselves, but if the locations of the sources are restricted, these parameters can be explained away as common constant factors. A Bjerhammar sphere is generally used since it allows for an independent adjustment of the overall depth of the interior source points. By maintaining connections to the symmetric kernel forms that are allowed by (6.2), a statistical covariance interpretation is possible.

As one might expect there are notable differences between this approach and DIDACKS since this alternative approach adopts the basic philosophy of maintaining connections to GC and DIDACKS does not. In particular for a specified domain, from RKHS theory one can deduce that the choice of inner product (or norm) and reproducing kernel must be in one-to-one correspondence [19]; hence, the choice of multipole form to be fit and norm are directly linked. A list of known covariance kernel and multipole point source correspondences for this approach can be found in [16]. For example, in [7] some of the consequences of the Krarup norm choice, along with its half-space approximation, are examined for point mass fits. By employing (6.2), this approach allows for an underpinning statistical covariance interpretation, which is important for many geophysical applications; nevertheless, implicitly retaining symmetric kernel forms also adds a layer of additional complexity, which has the effect of greatly complicating the theory and of limiting the possible choices of dipole and higher multipole forms. For uses outside geophysics there are also clearly interpretational difficulties that limit its use.

In DIDACKS theory all attempts at retaining connections to SRKs are dropped, and the primary emphasis is placed on the (weighted) energy or integral norm and the associated kernel form ℓ^{-1}. As noted earlier, this makes consideration of dipoles and higher multipoles trivial since the required inner products are readily obtained. Moreover, since the DIDACKS approach maintains the same norm choice for all source types there are few interpretational issues—especially with regards to multipole fits of all orders. The types of geophysical areas where one of these two approaches or GC should be preferred over the others clearly warrant further study and consideration since, to date, there have been no researchers proficient in applying all three algorithms. (While GC has had many practitioners and this alternative collocation based approach has had a few practitioners, so far DIDACKS development and use has been limited to the author's involvement.)

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] A. El Badia and T. Ha-Duong, *An inverse source problem in potential analysis*, Inverse Problems, 16 (2000), pp. 651–663.

[3] S. Bergman and M. Schiffer, *Kernel Functions and Elliptic Differential Equations in Mathematical Physics*, Academic Press, New York, 1953.

[4] R. J. Blakely, *Potential Theory in Gravity & Magnetic Applications*, Cambridge University Press, New York, 1996.

[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, John Wiley & Sons, New York, 1962.

[6] L. Garding, *The Dirichlet problem*, Math. Intelligencer, 2 (1979), pp. 42–52.

[7] H. Hauck and D. Lelgemann, *Regional gravity field approximation with buried masses using least-norm collocation*, Manuscripta Geodaetica, 10 (1985), pp. 50–58.

[8] W. A. Heiskanen and F. A. Vening Meinesz, *The Earth and Its Gravity Field*, McGraw–Hill, New York, 1958.

[9] W. A. Heiskanen and H. Moritz, *Physical Geodesy*, W. H. Freeman, San Francisco, 1967.

[10] B. Hofmann-Wellenhof and H. Moritz, *Physical Geodesy*, 1st ed., Springer-Verlag, New York, 2005.

[11] J. D. Jackson, *Classical Electrodynamics*, 3rd ed., John Wiley & Sons, New York, 1999.

[12] P. Kearey, M. Brooks, and I. Hill, *An Introduction to Geophysical Exploration*, 3rd ed., Blackwell Science, Malden, MA, 2002.

[13] O. D. Kellogg, *Foundations of Potential Theory*, Dover Publications, New York, 1953.

[14] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*, 2nd ed., Dover Publications, New York, 2003.

[15] T. Krarup, *A Contribution to the Mathematical Foundation of Physical Geodesy*, Publ. 44, Danish Geodesic Institute, Copenhagen, Denmark, 1969.

[16] A. Marchenko and D. Lelgemann, *A classification of reproducing kernels according to their functional and physical significance*, International Geoid Service (IGeS) Bulletin, Milan, 8 (1998), pp. 49–52.

[17] L. Máté, *Hilbert Space Methods in Science and Engineering*, Adam Hinger, Bristol, England, 1989.

[18] A. F. Mona, *Dirichlet's Principle: A Mathematical Comedy of Errors and Its Influence on the Development of Analysis*, Oosthoek, Scheltema & Holkema, Utrecht, The Netherlands, 1975.

[19] H. Moritz, *Advanced Physical Geodesy*, Abacus Press, Tunbridge Wells, Kent, England, 1980.

[20] C. Reigber, R. Schmidt, F. Flechtner, R. König, U. Meyer, K.-H. Neumayer, P. Schwintzer, and S. Y. Zhu, *An Earth field model complete to degree and order* 150 *from GRACE: EIGEN-GRACE*02*S*, J. Geodyn., 39 (2005), pp. 1–10.

[21] A. E. Rufty, *Point mass, dipole, and quadrupole gravity modeling for FBM systems support*, Naval Surface Warfare Center Dahlgren Division Technical Digest, 1997 Issue, pp. 100–108.

[22] R. Schaback and H. Wendland, *Kernel techniques: From machine learning to meshless methods*, Acta Numer., 15 (2006), pp. 1–97.

[23] N. H. Sleep, K. Fujita, and K. Fujita, *Principles of Geophysics*, Blackwell Science, Malden, MA, 1997.

[24] D. Stromeyer and L. Ballani, *Uniqueness of the inverse gravimetric problem for point mass models*, Manuscripta Geodaetica, 9 (1984), pp. 125–136.

[25] W. M. Telford, L. P. Geldart, and R. E. Sheriff, *Applied Geophysics*, 2nd ed., Cambridge University Press, New York, 1990.

[26] M. Vermeer, *Mass point geopotential modelling using fast spectral techniques; Historical overview, toolbox description, numerical experiment*, Manuscripta Geodaetica, 20 (1995), pp. 362–378.

[27] H. L. Weinert, ed., *Reproducing Kernel Hilbert Spaces Applications in Statistical Signal Processing*, Benchmark Papers in Electrical Engineering and Computer Science 25, Hutchinson Ross Publishing, Stroudsburg, PA, 1982.

SIAM J. APPL. MATH.
Vol. 68, No. 1, pp. 222–238

MOISTURE TRANSPORT AND DIFFUSIVE INSTABILITY DURING BREAD BAKING[*]

H. HUANG[†], P. LIN[‡], AND W. ZHOU[§]

Abstract. In this paper we study multiphase models for simultaneous heat and mass transfer processes during bread baking. Our main objective is to provide an explanation and a remedy to the observed erroneous and/or divergent results associated with an instantaneous phase change model used in the literature. We propose a reaction-diffusion model based on the Hertz–Knudsen equation, where phase change is not instantaneous but determined by an evaporation/condensation rate. A splitting scheme is designed for the reaction-diffusion model so that a link between these two models can be established and the nonintuitive numerical instability associated with the instantaneous phase change model can be identified and eliminated. The evaporation/condensation rate is estimated by comparing results of the reaction-diffusion model with experimental observations reported in the literature. For evaporation/condensation rate beyond the estimated value, oscillatory solution with multiple regions of dry and two-phase zones is observed. We show that these are caused by an instability intrinsic to the model (which we call diffusive instability) using linear stability analysis and numerical tests.

Key words. diffusive instability, finite difference method, heat and mass transfer, linear stability analysis, multiphase modeling, phase change, reaction-diffusion equation

AMS subject classifications. 76R50, 80A20, 76D27, 65M06, 65M12, 76T99, 35K57

DOI. 10.1137/060653329

1. Introduction.

The bread baking process is difficult to model, partly due to the fact that simultaneous heat and mass transfer are involved during the process. During baking, heat transfer in dough is a combination of conduction/radiation from band or tins to the dough surface, convection from air to the dough surface in the absence of baking tins, conduction in the continuous liquid/solid phase of the dough, and evaporation-condensation in the gas phase of the dough.

De Vries, Sluimer, and Bloksma [9] described a 4-step mechanism for the heat transport inside dough: (1) water evaporates at the warmer side of a gas cell that absorbs latent heat of evaporation; (2) water vapor then migrates through the gas phase; (3) when meeting the cooler side of the gas cell, water vapor condenses and becomes liquid; (4) finally, heat and water are transported by conduction and diffusion through the gluten gel to the warmer side of the next cell. The water diffusion mechanism becomes important to heat transfer, because dough tends to be a poor conductor that limits the heat transfer rate via conduction.

The above described mechanism of diffusion, together with evaporation and condensation in dough, was subsequently adopted by Tong and Lund [22] and Thorvalds-

[*] Received by the editors February 28, 2006; accepted for publication (in revised form) July 9, 2007; published electronically October 19, 2007.
http://www.siam.org/journals/siap/68-1/65332.html

[†] Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada (hhuang@yorku.ca). This author's work was partially supported by research grants from NSERC and MITACS of Canada.

[‡] Department of Mathematics, National University of Singapore, Singapore 117543 (matlinp@nus.edu.sg). This author's work was partially supported by academic research grant R-146-000-053-112 from the National University of Singapore.

[§] Department of Chemistry, Food Science and Technology Programme, National University of Singapore, Singapore 117543 (chmzwb@nus.edu.sg). This author's work was partially supported by academic research grant R-143-000-181-112 from the National University of Singapore.

son and Janestad [20]. With this mechanism, liquid water moved towards the loaf center as well as the surface by evaporation and condensation, reducing the partial water vapor pressure due to the temperature gradient. As a result, crumb temperature change was accelerated.

In the model by Zanoni, Peri, and Pierucci [23], an evaporation front inside the dough was assumed to always be at 100°C. This evaporation front progressively advanced towards the center as the bread's temperature increased. Crust was formed in the bread portion above the evaporation front. With similar parameters, Zanoni, Peri, and Pierucci [24] developed a 2-dimensional axisymmetric heat diffusion model. The phenomena were described separately for the upper and lower parts (crust and crumb). The upper part (crust) temperature was determined by equations including heat supply by convection, conductive heat transfer towards the inside, and convective mass transfer towards the outside. The lower part (crumb) temperature was determined by Fourier's law. In addition to the Cartesian coordinate models, a 1-dimensional cylindrical coordinate model was also established by De Vries, Sluimer, and Bloksma [9].

Among the various models, the internal evaporation-condensation mechanism well explains the fact that heat transfer in bread during baking is much faster than that described by the conduction alone in dough/bread. It also supports the observation that there is an increase in the liquid water content at the center of the bread during the early stage of baking (Thorvaldsson and Skjoldebrand [21]) rather than a monotonous decrease resulted from having liquid water diffusion and surface evaporation only. Therefore, a promisingly good model for bread baking might be a multiphase model which consists of three partial differential equations for the simultaneous heat transfer, liquid water diffusion, and water vapor diffusion, respectively, together with two algebraic equations describing water evaporation and condensation in the gas cells. Indeed, Thorvaldsson and Janestad [20] used this multiphase model to describe a 1-dimensional case where a baking tin was absent.

In [20], a slab of bread crumb being baked in a conventional oven was considered. Their model assumes that vapor and liquid water diffuse separately and phase change (evaporation and condensation) occurs instantaneously; i.e., vapor content V is directly proportional to saturated vapor content V_s at any given time, provided that there is enough liquid water available. The authors showed reasonably good agreement between the numerical results predicted by the model and their experimental measurements. However, further investigations using the same model revealed that instability occurs as time step size was refined [18, 25].

Zhou [25] demonstrated that numerically solving the model equations presented a big challenge. Although various schemes of finite difference methods and finite element methods were applied, the corresponding solutions were all shown to be highly sensitive to the time step size, and satisfactory results were yielded only over a limited range of time step sizes. Erroneous and/or divergent results were produced when the time step size was either too small or too large. While it is reasonable to expect that large time steps may lead to instability, failure of the numerical methods due to small time step sizes is nonintuitive.

The main objective of this paper is to provide an explanation as well as a remedy to this highly nonintuitive outcome of the instantaneous phase change model proposed in [20]. In order to identify the source of instability in the instantaneous phase change model, we construct a reaction-diffusion model by assuming that phase change follows a modified Hertz–Knudsen equation, which can be derived using statistical mechanics principles [13]. We show that the two models are related from a numerical point of view when a special time stepping scheme is used for the reaction-diffusion model.

The instantaneous phase change model is equivalent to the reaction-diffusion model with a variable evaporation rate reciprocal to the time step size. Therefore, reducing the time step size in the instantaneous phase change model increases the evaporation rate in the reaction-diffusion model.

The reaction-diffusion model allows us to study the effects of the evaporation rate in more detail by separating two types of instabilities, i.e., instability associated with a particular numerical scheme and diffusive instability associated with the model. Combining numerical tests and linear stability analysis, we show that diffusive instability is an intrinsic feature of the model, which does not disappear as time step size is reduced. As a result, the two-phase region where vapor and liquid water coexist may become unstable. For the parameter values used in [20, 18, 25], the constant state solution in the two-phase region is linearly unstable even though the solution is stable when phase change and diffusion are considered separately. Furthermore, the rate of growth depends on the evaporation rate. A larger evaporation rate leads to faster growth of the disturbances. Oscillation in solutions can occur in the two-phase region before the dry region completely takes over.

We show that the reaction-diffusion model does not lead to numerical instability when sufficiently small time step size is used. For the instantaneous phase change model, reducing the time step size is equivalent to increasing the evaporation rate while keeping the product of time step size and evaporation rate fixed. Therefore, systematic refinement of the time step size induces faster growth of error, which eventually causes numerical instability observed in the computations in [18, 25].

In order to put the problem we study in this paper in a broader context, we shall give a brief overview of some relevant applications whose mathematical description has certain similarity to the bread baking models discussed here. A general feature of these problems is the coupling of thermal diffusion and phase change. We refer the readers to [3, 5, 6, 7, 10] for treatments of phase change and heat transfer phenomena in general. A closely related application is given in [16, 15], where a moisture transport model for the wetting and cooking of a cereal grain is considered. The temperature is decoupled from the moisture model, and it is used as a parameter in their model. In [12, 11], the condensed phase combustion or gasless combustion is considered. The model has applications in synthesizing certain ceramics and metallic alloys and involves a reaction-diffusion system for temperature and fuel concentration, where the diffusion coefficient of fuel concentration is set to be zero. In [2], a model for the aggregate alkali reaction in fluid leaching processes, which is similar to the gasless combustion problem, is studied numerically. All these models are similar to the one studied in [20, 18, 25], with some differences in the way the phase change is handled and in the parameters and coefficients. Numerical techniques used in those studies are also similar, where either finite difference and finite element methods or pseudospectral methods are used for spatial discretization.

The rest of the paper is organized as follows. In section 2, we describe a reaction–diffusion model based on the Hertz–Knudsen equation. The relationship between our model and the instantaneous phase change model is explored in section 2.2 when numerical procedure is discussed. Numerical results for the reaction-diffusion model are given in section 2.3. In section 3 we carry out linear stability analysis of our new model. We finish the paper with a conclusion and a short discussion on future directions in section 4.

2. A reaction-diffusion model. Following [20, 18, 25] we assume that the bread slab can be treated as a 1-dimensional homogeneous porous medium ($0 < x <$

L) with density ρ, specific heat capacity c_p, and thermal conductivity k. While in reality both c_p and k depend on water content, they are assumed to be constants in [20], where density ρ is given as a linear function of the liquid water content. The main variables are temperature T, liquid water content W, and water vapor content V inside the bread with respect to the total weight (both are dimensionless variables defined as the percentage of liquid water and vapor mass with respect to that of the mixture including the bread). Thus the total vapor and liquid water masses per unit of volume can be computed as ρV and ρW, respectively. Vapor and liquid water can be generated via evaporation and condensation, respectively, determined by the saturated vapor concentration V_s (or saturation pressure P_s), which is temperature dependent. Vapor and liquid water can also be transported via diffusion, with coefficients D_v and D_w, respectively.

The governing equations for T, V, and W are

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k \frac{\partial T}{\partial x}\right) + \lambda \Gamma, \tag{2.1}$$

$$\frac{\partial V}{\partial t} = \frac{\partial}{\partial x}\left(D_v \frac{\partial V}{\partial x}\right) - \frac{\Gamma}{\rho}, \tag{2.2}$$

$$\frac{\partial W}{\partial t} = \frac{\partial}{\partial x}\left(D_w \frac{\partial W}{\partial x}\right) + \frac{\Gamma}{\rho}. \tag{2.3}$$

Here Γ is the rate of the phase change (mass per unit volume per unit time), given by the modified Hertz–Knudsen equation (see [13], where this equation is used for studying evaporation phenomena)

$$\Gamma = E(1-\phi)\sqrt{\frac{M}{2\pi R}}\frac{(P_v - cP_s)}{\sqrt{T}}, \tag{2.4}$$

where E is the condensation/evaporation rate, ϕ is the porosity of the bread slab, M is the molecular weight of water and R is the universal gas constant, P_v and P_s are the vapor pressure and saturation pressure, and c is a phase change constant. For simplicity we have included the pore surface area in E, which is inversely proportional to the pore size of the porous sample. Even for a unit pore surface, the value of evaporation rate E is subject to debate [13]. In this study, we estimate the value of E by comparing numerical results of the current model with experimental observations [20]. We also use the fitted saturated vapor pressure data [19] by an exponential function

$$P_s = P_{s,0}\exp[\kappa(T - T_0)] - P_{s,1}. \tag{2.5}$$

Assuming the ideal gas law,

$$P_v = \frac{R\rho V T}{\phi M}, \quad P_s = \frac{R\rho V_s T}{\phi M}, \tag{2.6}$$

we can rewrite (2.4) as

$$\Gamma = \frac{E(1-\phi)\rho}{\phi}\sqrt{\frac{RT}{2\pi M}}\left(V - cV_s\right). \tag{2.7}$$

It is worth noting that evaporation can occur only when there exists a sufficient amount of liquid water. Thus, if $V < cV_s$ and $W = 0$, then no evaporation occurs and $\Gamma = 0$.

As in [20, 18, 25], we assume that initially the bread has certain moisture content W_0 (liquid water) at room temperature T_0. At time zero, the bread is placed in an oven preheated at temperature T_a by a radiator with temperature T_r. The boundary conditions are mixed conditions at $x = 0$:

$$-k\frac{\partial T}{\partial x} = h_r(T_r - T) + h_c(T_a - T), \tag{2.8a}$$

$$-\frac{\partial V}{\partial x} = h_v(V_a - V), \tag{2.8b}$$

$$-\frac{\partial W}{\partial x} = h_w(W_a - W), \tag{2.8c}$$

and they are symmetric conditions at $x = L$:

$$\frac{\partial T}{\partial x} = \frac{\partial V}{\partial x} = \frac{\partial W}{\partial x} = 0. \tag{2.8d}$$

Here h_r, h_c, h_v, and h_w are the radiative and convective heat transfer coefficients and the vapor and liquid water mass transfer coefficients, respectively. V_a and W_a are the vapor and liquid water content in the oven air.

Finally, some of the coefficients are estimated experimentally and following expressions [20] will be used in this paper. The diffusion coefficient for water D_w is approximated as a constant, while for vapor diffusion, we have $D_v = \bar{D}_v T^2$. Heat transfer coefficients depend on temperature and sometimes water content, and they are approximated by $h_v = \bar{h}_v T^{-3}$, $h_w = \bar{h}_{w0}T + \bar{h}_{w1}W + \bar{h}_{w2}TW + \bar{h}_{w3}W^2$, and $h_r = \sigma(T_r^2 + T^2)(T_r + T)/(\epsilon_p^{-1} + \epsilon_r^{-1} - 2 + F_{sp}^{-1})$, where F_{sp} is the shape factor and ϵ_p and ϵ_r are the emissivities of the bread and radiator, respectively. Finally, the density of the mixture is approximated by $\rho = \bar{\rho}_0 + \bar{\rho}W$. The readers are referred to [20] for more details on relevant parameter values.

2.1. Nondimensionalization. We now proceed to nondimensionalize the governing equations by choosing the following scaling:

$$\theta = \frac{T - T_0}{\bar{T}}, \quad \tau = \frac{t}{\bar{t}}, \quad \xi = \frac{x}{L}, \quad \rho' = \frac{\rho}{\bar{\rho}}.$$

Here $T_0 = 298$ K is the initial temperature and $\bar{T} = T_r - T_0$ is the difference between the radiator and initial temperature. We choose the diffusive time scale for the temperature equation

$$\bar{t} = \frac{\bar{\rho} c_p L^2}{k}.$$

Drop the prime for simplicity and we have the following equations:

$$\theta_\tau = \frac{1}{\rho}\theta_{\xi\xi} + S, \tag{2.9}$$

$$V_\tau = D_1[(\theta + \theta_0)^2 V_\xi]_\xi - \alpha S, \tag{2.10}$$

$$W_\tau = D_2 W_{\xi\xi} + \alpha S, \tag{2.11}$$

where the nondimensional density is given as $\rho = W + \rho_0$ with $\rho_0 = \bar{\rho}_0/\bar{\rho}$. The other dimensionless parameters are defined as

$$D_1 = \frac{\bar{D}_v \bar{\rho} \bar{T}^2 c_p}{k}, \quad D_2 = \frac{D_w \bar{\rho} c_p}{k}, \quad \alpha = \frac{\bar{T} c_p}{\lambda}.$$

When there is sufficient amount of liquid water, the source term for phase change is given by

$$S = \beta\sqrt{\theta + \theta_0}\,(V - cV_s)\,, \tag{2.12}$$

where

$$\beta = \lambda E(1-\phi)\frac{\bar{\rho}L^2}{\phi k}\sqrt{\frac{R}{2\pi M\bar{T}}},$$

and the saturated vapor content is

$$V_s = \frac{V_{s,0}e^{\gamma\theta} - V_{s,1}}{(\rho_0 + W)(\theta + \theta_0)},$$

with parameters

$$V_{s,0} = \frac{P_{s,0}M}{\bar{\rho}\bar{T}R}, \quad V_{s,1} = \frac{P_{s,1}M}{\bar{\rho}\bar{T}R}, \quad \gamma = \kappa\bar{T}.$$

The boundary conditions at $x = 0$ ($\xi = 0$) and $x = L$ ($\xi = 1$) are nondimensionalized similarly. At $\xi = 0$, we have

$$\theta_\xi = (h_3 + h_4)(\theta - 1), \tag{2.13a}$$

$$V_\xi = h_1(V - V_a), \tag{2.13b}$$

$$W_\xi = h_2(W - W_a), \tag{2.13c}$$

where

$$\begin{aligned} h_1 &= h_v L, \\ h_2 &= h_w L, \\ h_3 &= \frac{\sigma\bar{T}^3 L}{k}\frac{[(1+\theta_0)^2 + (\theta+\theta_0)^2](\theta + 1 + 2\theta_0)}{\epsilon_p^{-1} + \epsilon_r^{-1} - 2 + F_{sp}^{-1}}, \\ h_4 &= \frac{h_c L}{k}. \end{aligned}$$

Here we have assumed that the air temperature T_a equals the radiator temperature T_r. At $\xi = 1$, we have

$$\theta_\xi = V_\xi = W_\xi = 0. \tag{2.13d}$$

2.2. Numerical method. We now turn our attention to numerical procedures by describing a splitting scheme for the reaction-diffusion model (cf. [4]). We will revisit the instantaneous phase change model and establish a connection between our model and the instantaneous phase change model. Numerical solutions of our new model will be presented at the end of the section.

2.2.1. Reaction-diffusion model. Splitting scheme. We adopt a splitting method for the reaction-diffusion model by separating diffusion from phase change (reaction) and solving the equations in two steps.[1] We first solve the vapor and liquid

[1] From the numerical method point of view, there is also a benefit to using the splitting. Since we can obtain an explicit relation between W and V in the phase change stage, W can be computed using V without treating the source term S in the W equation (or with that relation the source term in the W equation can be written in terms of W and is dissipative). As a result, the algorithm is more stable. We note that time splitting schemes have been used in the literature for reaction-diffusion equations, and we refer interested readers to [1, 8] and the references therein.

water equations by

$$V_\tau = -\alpha S, \tag{2.14}$$

$$W_\tau = \alpha S. \tag{2.15}$$

The second equation for liquid water can be replaced by an algebraic constraint

$$V + W = V^0 + W^0,$$

where V^0 and W^0 are the values at the beginning of the splitting step. Diffusion of vapor and liquid water as well as temperature are solved in the second step by

$$\theta_\tau = \frac{1}{\rho_0 + W}\theta_{\xi\xi} + S, \tag{2.16}$$

$$V_\tau = D_1[(\theta + \theta_0)^2 V_\xi]_\xi, \tag{2.17}$$

$$W_\tau = D_2 W_{\xi\xi}. \tag{2.18}$$

Time stepping scheme. The numerical procedure in semidiscrete form can be described as follows.

1. Vapor content is first solved using

$$\frac{V^c - V^{(n)}}{\Delta\tau} = -\alpha S_1, \tag{2.19a}$$

   where

$$S_1 = \beta\sqrt{\theta^{(n)} + \theta_0}(\bar{V} - cV_s^{(n)}). \tag{2.19b}$$

   Here superscript c denotes the solution updated due to phase change alone and $\bar{V}$ is the arithmetic average, i.e.,

$$\bar{V} = \frac{V^c + V^{(n)}}{2}. \tag{2.20}$$

2. Evaporation can occur only if there exists a sufficient amount of the liquid water. In other words, liquid water content must remain nonnegative. If the available water $W^{(n)}$ is less than the amount $V^c - V^{(n)}$ computed by (2.19), then the amount of liquid water becomes zero. Otherwise, it is given by $W^{(n)} - V^c + V^{(n)}$. Therefore,

$$W^c = \max\{W^{(n)} + V^{(n)} - V^c, 0\}. \tag{2.21}$$

3. As a consequence, vapor content needs to be corrected using the constraint

$$V^c = W^{(n)} + V^{(n)} - W^c. \tag{2.22}$$

4. To account for the possibility of running out of liquid water, phase change needs to be corrected using

$$S_2 = \frac{W^c - W^{(n)}}{\alpha\Delta\tau}. \tag{2.23}$$

5. $V^{(n+1)}$ and $W^{(n+1)}$ are updated using the diffusion equations

$$\frac{V^{(n+1)} - V^c}{\Delta\tau} = D_1[(\theta^{(n)} + \theta_0)^2 V_\xi^{(n+1)}]_\xi, \tag{2.24}$$

$$\frac{W^{(n+1)} - W^c}{\Delta\tau} = D_2 W_{\xi\xi}^{(n+1)}. \tag{2.25}$$

6. Temperature is solved by

$$\frac{\theta^{(n+1)} - \theta^{(n)}}{\Delta\tau} = \frac{1}{\rho_0 + W^{(n+1)}} \theta_{\xi\xi}^{(n+1)} + S_2, \tag{2.26}$$

where the rate of phase change S_2 is given by (2.23).

The time discretizations described above are standard, and it is straightforward to verify that the scheme is consistent and formally the order of accuracy is $\Delta\tau$ for reasonably chosen $\bar{V}$, such as the arithmetic average used in this paper.

2.2.2. Instantaneous phase change model in [20]. The instantaneous phase change model used and studied in [20, 18, 25] is essentially a discrete time model which can be described as follows (in the nondimensional form).

1. Vapor and water contents are solved using

$$V^* = \min\{cV_s^{(n)}, V^{(n)} + W^{(n)}\}, \quad W^* = V^{(n)} + W^{(n)} - V^* \tag{2.27a}$$

or

$$W^* = \max\{W^{(n)} + V^{(n)} - cV_s^{(n)}, 0\}, \quad V^* = V^{(n)} + W^{(n)} - W^*. \tag{2.27b}$$

2. Rate of phase change is computed using

$$S_1 = \frac{W^* - W^{(n)}}{\alpha\Delta\tau}. \tag{2.28}$$

3. Vapor content due to diffusion is updated by

$$\frac{V^{**} - V^*}{\Delta\tau} = D_1[(\theta^{(n)} + \theta_0)^2 V_\xi^{**}]_\xi. \tag{2.29}$$

4. Vapor and water contents are updated again using

$$V^{(n+1)} = \min\{cV_s^{(n)}, V^{**} + W^*\}, \quad W^{**} = V^{**} + W^* - V^{(n+1)} \tag{2.30a}$$

or

$$W^{**} = \max\{W^* + V^{**} - cV_s^{(n)}, 0\}, \quad V^{(n+1)} = V^{**} + W^* - W^{**}. \tag{2.30b}$$

5. Rate of additional phase change is computed using

$$S_2 = \frac{W^{**} - W^*}{\alpha\Delta\tau}. \tag{2.31}$$

6. $W^{(n+1)}$ is updated using the diffusion equations

$$\frac{W^{(n+1)} - W^{**}}{\Delta\tau} = D_2 W_{\xi\xi}^{(n+1)}. \tag{2.32}$$

7. Temperature is solved by

$$\frac{\theta^{(n+1)} - \theta^{(n)}}{\Delta\tau} = \frac{1}{\rho_0 + W^{(n+1)}} \theta_{\xi\xi}^{(n+1)} + S_1 + S_2. \tag{2.33}$$

2.2.3. Simplified instantaneous phase change model. The rate of phase change given by (2.31) is typically small compared to that from (2.28). Therefore, we can simplify the model by eliminating steps (2.30) and (2.31) and using the following procedure.

1. Vapor and water contents are solved using

$$W^* = \max\{W^{(n)} + V^{(n)} - cV_s^{(n)}, 0\}, \quad V^* = V^{(n)} + W^{(n)} - W^*. \tag{2.34}$$

2. Rate of phase change is computed using

$$S = \frac{W^* - W^{(n)}}{\alpha\Delta\tau}. \tag{2.35}$$

3. Vapor content due to diffusion is updated by

$$\frac{V^{(n+1)} - V^*}{\Delta\tau} = D_1[(\theta^{(n)} + \theta_0)^2 V_\xi^{(n+1)}]_\xi. \tag{2.36}$$

4. $W^{(n+1)}$ is updated using

$$\frac{W^{(n+1)} - W^*}{\Delta\tau} = D_2 W_{\xi\xi}^{(n+1)}. \tag{2.37}$$

5. Temperature is solved by

$$\frac{\theta^{(n+1)} - \theta^{(n)}}{\Delta\tau} = \frac{1}{\rho_0 + W^{(n+1)}}\theta_{\xi\xi}^{(n+1)} + S. \tag{2.38}$$

2.2.4. Model comparison. Note that the diffusion part of the reaction-diffusion model (2.24)–(2.26) is the same as that for the simplified instantaneous model (2.36)–(2.38). The only difference between the reaction-diffusion model and the instantaneous phase change model lies in the way the phase change is computed, or more precisely the way vapor content is computed. We now show that the two models are related to each other in the following sense.

In the instantaneous phase change model, assuming there is a sufficient amount of water,[2] we set

$$V^* = cV_s. \tag{2.39}$$

Again assuming there is a sufficient amount of water, using the reaction-diffusion model and $\bar{V}$ defined in (2.20), we obtain

$$V^c - V^{(n)} = -\nu(V^c + V^{(n)} - 2cV_s^{(n)}),$$

where

$$\nu = \frac{\Delta\tau\alpha\beta\sqrt{\theta + \theta_0}}{2}.$$

When $\nu = 1$, or

[2] When all the available water is evaporated, it is straightforward to show that the two models are equivalent since $V^c = V^* = V^{(n)} + W^{(n)}$ and $W^c = W^* = 0$.

$$\beta = \frac{2}{\Delta\tau\alpha\sqrt{\theta+\theta_0}}, \tag{2.40}$$

we obtain

$$V^c = cV_s^{(n)}.$$

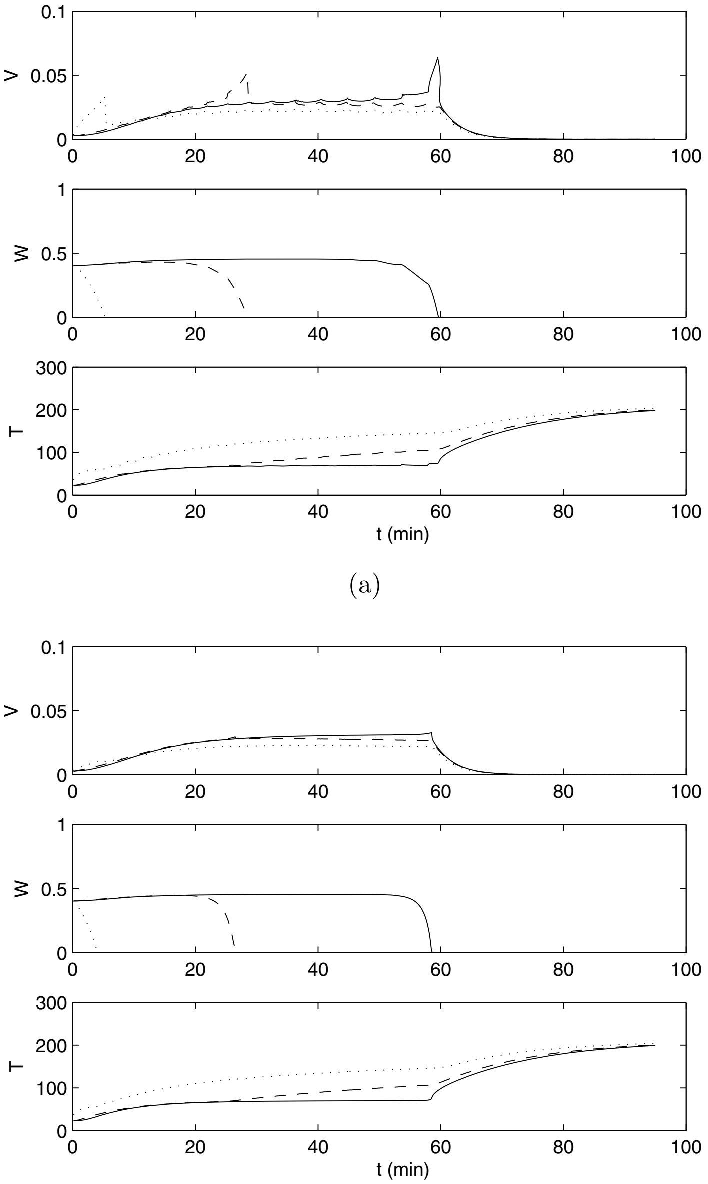
Therefore, the instantaneous phase change model is numerically equivalent to the reaction-diffusion model with a variable β inversely proportional to the time step size $\Delta\tau$. This suggests that the instantaneous phase change model can be viewed as an *inconsistent* discretization of the reaction-diffusion model, which provides a possible explanation for the observed divergence in the numerical solution as $\Delta\tau \to 0$.

2.3. Numerical results. The numerical solutions are obtained using parameter values given in [20, 25]. The corresponding dimensionless parameters are $D_1 = 4.37$, $D_2 = 1.9\times10^{-3}$, $h_1 = 5.05(\theta+\theta_0)^{-3}$, $h_2 = 2.6\times10^{-3}\theta+2.7\times10^{-3}W-7.4\times10^{-4}\theta W-7.7\times10^{-3}W^2$, $h_3 = 5.13\times10^{-2}(\theta+1+2\theta_0)[(1+\theta_0)^2+(\theta+\theta_0)^2]$, $h_4 = 7.14\times10^{-2}$, $V_{s,0} = 3.51\times10^{-5}$, $V_{s,1} = 1.45\times10^{-5}$, $\alpha = 0.286$, $\beta = 1.735\times10^5 E$, $\gamma = 9.662$, $\rho_0 = 5.99\times10^{-1}$, $\theta_0 = 1.61$.

To simulate the moisture transport during bread baking we need to determine the evaporation or condensation rate E or its nondimensionalized quantity β. This is done by choosing its value so that the numerical solution matches the experimental results in [20]. The estimated value is $\beta = 781$ and the corresponding evaporation rate is $E = 4.5\times10^{-3}$. In Figure 2.1, the numerical results based on this evaporation rate is given for two time step sizes $\Delta t = 10$ and 0.2 seconds using a coarse grid (16 grid points in x) and a fine grid (128 grid points in x), respectively. The numerical procedure remains stable as the time step size is reduced for a fixed grid size in x, contrary to the instantaneous phase change model. We note that the nonsmoothness of the solution in Figure 2.1(a) is due to the coarse spatial and temporary grids used in the computation. When we refine the spatial grid, solution becomes smooth, as shown in Figure 2.1(b).

Recall from (2.40) that reducing time step size in the instantaneous phase change model is equivalent to increasing the value of E in the reaction-diffusion model. Therefore, it will help us to understand the mechanism of the instability associated with the instantaneous phase change model by carrying out computations using a larger value of E. In Figure 2.2, computational results using the reaction-diffusion model with $E = 0.0045$ and $E = 0.045$ are presented. Here we plot the snapshots of water content at various times in order to show what can happen with the model if we increase E. The results are obtained using 128 grid points for the x variable and a time step size of 0.2 seconds. We have also done further refinement tests, but the results remain virtually the same, indicating convergence of the numerical solutions. Since evaporation rate is independent of the time step size, we can use a much smaller time step size while keeping the evaporation rate unchanged, which is not possible for the instantaneous phase change model. From Figure 2.2, it can be seen that the solution for the larger evaporation rate starts to oscillate. However, this is not due to numerical instability.

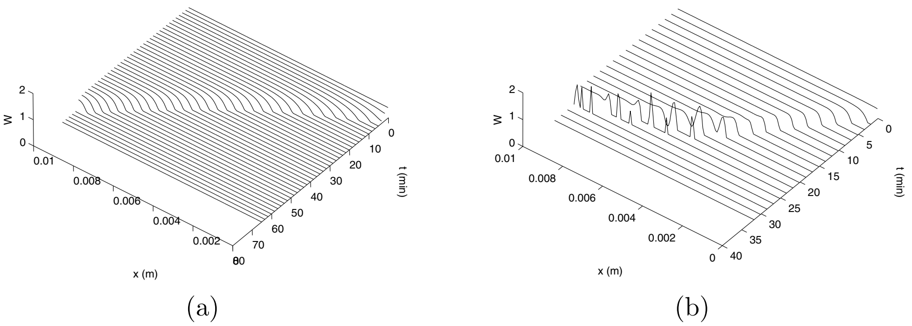
We have also experimented with the instantaneous phase change model, and the results confirmed the observations in [18, 25]. Instead of repeating those results here, we refer interested readers to [18, 25] for more details. In the next section we will provide an explanation for the observed oscillation in the solution of the reaction-diffusion model associated with relatively large value of E (or small $\Delta\tau$ in the instantaneous phase change model when the numerical procedure is still stable).



(a)

(b)

Fig. 2.1. *Computational results of V, W, and T obtained by using the reaction-diffusion model:* (a) *Time and spatial step sizes are $\Delta t = 10$ seconds and $\Delta x = 2^{-4} \times 10^{-2}$ cm;* (b) *Time and spatial step sizes $\Delta t = 0.1$ seconds and $\Delta x = 2^{-8} \times 10^{-2}$ cm. Here the dotted lines represent the quantities at the surface of the slab ($x = 0$), the dashed lines are for those located in the middle of the domain ($x = L/2$), and the solid lines represent the values at the center of the slab ($x = L$). The time is measured in minutes in order to compare our results with those from previous studies in* [20, 18, 25].



(a) (b)

FIG. 2.2. *Snapshots of the water content computed using the reaction-diffusion model for two values of the evaporation constant:* (a) $E = 0.0045$*;* (b) $E = 0.045$*. For the smaller value of* E*, the water content remains monotonic in* x *and the interface between the dry and wet regions is clearly defined. For the larger value of* E*, disturbance from the interface grows and eventually leads to multiple dry-wet regions.*

3. Linear stability analysis. We now turn our attention to the stability of the solution of the nondimensional model (2.9)–(2.11) near a steady state in an infinite domain. It is easy to see that the constant state V_0, W_0, and $\theta = 0$ satisfies the equations as long as

$$V_0 = \frac{c\,(V_{s,0} - V_{s,1})}{\theta_0(\rho_0 + W_0)}.$$

To examine the stability of this constant state solution, we carry out linear stability analysis by assuming that

$$\theta = \hat{\theta}\exp(s\tau + im\xi), \quad V = V_0 + \hat{v}\exp(s\tau + im\xi), \quad W = W_0 + \hat{w}\exp(s\tau + im\xi).$$

The equations for $\hat{\theta}$, $\hat{v}$, and $\hat{w}$ are

$$s\hat{\theta} = -\frac{m^2}{\rho_0 + W_0}\hat{\theta} + \bar{\beta}\left(-V_1\hat{\theta} + \hat{v} + \frac{V_0}{\rho_0 + W_0}\hat{w}\right), \tag{3.1}$$

$$s\hat{v} = -\theta_0^2 D_1 m^2 \hat{v} - \alpha\bar{\beta}\left(-V_1\hat{\theta} + \hat{v} + \frac{V_0}{\rho_0 + W_0}\hat{w}\right), \tag{3.2}$$

$$s\hat{w} = -D_2 m^2 \hat{w} + \alpha\bar{\beta}\left(-V_1\hat{\theta} + \hat{v} + \frac{V_0}{\rho_0 + W_0}\hat{w}\right), \tag{3.3}$$

where $\bar{\beta} = \sqrt{\theta_0}\beta$ and

$$V_1 = \frac{c\,(V_{s,0} - V_{s,1})}{\theta_0(\rho_0 + W_0)} - \frac{cV_{s,0}\gamma}{\theta_0(\rho_0 + W_0)}.$$

Rearranging (3.1)–(3.3) in the matrix form, we obtain

$$M\mathbf{y} = s\mathbf{y},$$

where

$$\mathbf{y} = \begin{pmatrix} \hat{\theta} \\ \hat{v} \\ \hat{w} \end{pmatrix}, \quad M = \begin{pmatrix} -\frac{m^2}{\rho_0+W_0} - \bar{\beta}V_1 & \bar{\beta} & \bar{\beta}\frac{V_0}{\rho_0+W_0} \\ \alpha\bar{\beta}V_1 & -\theta_0^2 D_1 m^2 - \alpha\bar{\beta} & -\alpha\bar{\beta}\frac{V_0}{\rho_0+W_0} \\ -\alpha\bar{\beta}V_1 & \alpha\bar{\beta} & -D_2 m^2 + \alpha\bar{\beta}\frac{V_0}{\rho_0+W_0} \end{pmatrix}. \tag{3.4}$$

3.1. A special case. We want to find out whether instability would occur in this system. This requires us to find eigenvalues of (3.4). To avoid complicated calculations we consider a special case where the linearized diffusion coefficients for temperature and liquid water are the same, i.e.,

$$\frac{1}{\rho_0 + W_0} = \theta_0^2 D_1. \tag{3.5}$$

For our choice of parameters in this bread-baking problem these two coefficients are of similar sizes. Therefore, we expect that this assumption would not change the nature of the stability of the system. Under this assumption the eigenvalues satisfy the following equation:

$$\begin{aligned}(\theta_0^2 D_1 m^2 + s)\Bigg[s^2 + \left(D_2 m^2 + \theta_0^2 D_1 m^2 + \bar{\beta} V_1 + \alpha\bar{\beta} - \frac{\alpha\bar{\beta}V_0}{\rho_0 + W_0}\right) s \\ + D_2 m^2(\theta_0^2 D_1 m^2 + \bar{\beta}V_1 + \alpha\bar{\beta}) - \frac{\alpha\bar{\beta}V_0}{\rho_0 + W_0}\theta_0^2 D_1 m^2\Bigg] = 0.\end{aligned} \tag{3.6}$$

3.1.1. Reaction-only. When there is no diffusion, we have $m = 0$, and the eigenvalues are $s_1 = s_2 = 0$ and

$$s_3 = -\alpha\bar{\beta} - \left(\frac{V_1}{V_0} - \frac{\alpha}{\rho_0 + W_0}\right)\bar{\beta}V_0.$$

For our problem, $\alpha = 0.2864$, $\gamma = 9.662$, $\rho_0 = 0.5986$, $\theta_0 = 1.6116$, and $W_0 = 0.4$. Thus, $s_3 < 0$ and the constant solution is stable.

3.1.2. Reaction-diffusion. We can easily see that the eigenvalue associated with the first factor of (3.6) is negative. Note that the coefficient of s in the second factor is

$$C_1 = D_2 m^2 + \theta_0^2 D_1 m^2 + \bar{\beta}V_1 + \alpha\bar{\beta} - \frac{\alpha\bar{\beta}V_0}{\rho_0 + W_0} > 0.$$

The signs of two eigenvalues associated with the second factor are determined by the sign of the coefficient of s^0:

$$C_0 = D_2 m^2(\theta_0^2 D_1 m^2 + \bar{\beta}V_1 + \alpha\bar{\beta}) - \frac{\alpha\bar{\beta}V_0}{\rho_0 + W_0}\theta_0^2 D_1 m^2$$

since the eigenvalues from the second factor are given by $-C_1 \pm \sqrt{C_1^2 - 4C_0}$. So if $C_0 < 0$ and $C_1^2 > 4C_0$, we will have a positive eigenvalue which indicates instability of the system. If D_2 is comparable to D_1, then C_0 will be positive and both eigenvalues will be negative and the system is stable. Since in our case D_2 is much smaller than D_1, C_0 will be negative if m is not sufficiently large, in which case we will have one positive eigenvalue which leads to instability.

Remark. It is interesting to note that in this case diffusion is actually destabilizing. The instability discussed here is conceptually related to Turing instability or diffusive instability observed in pattern formation, crystal, and tumor growth; cf. [17, 14]. However, the mathematical model and the nonlinearity related to bread baking are quite different from those applications.

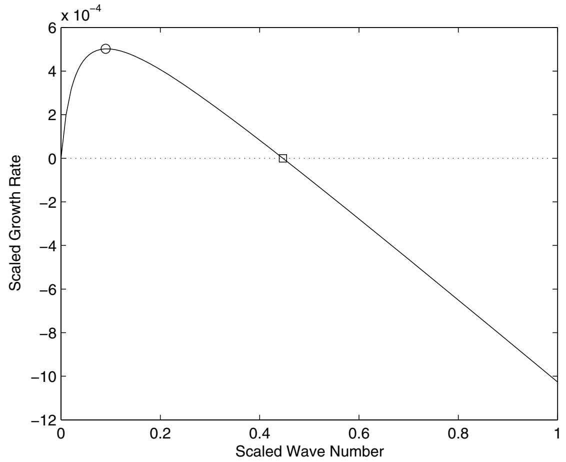
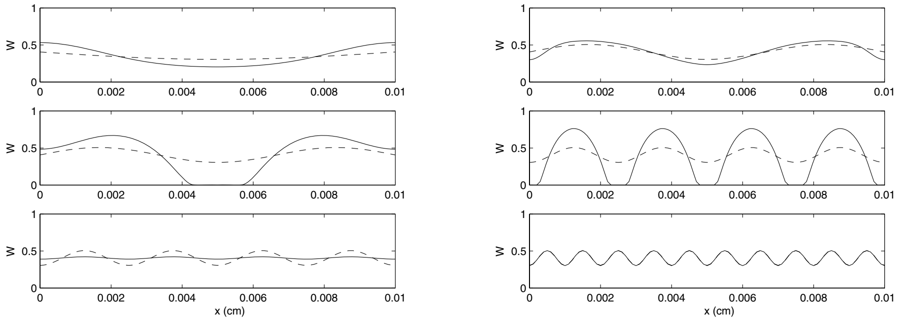


FIG. 3.1. *Plot of the normalized eigenvalue $\bar{s}$ vs. the normalized wave number $\bar{m}$. The circle indicates the most unstable wave number, $\bar{m}_{max} = 0.09$. The square indicates the upper limit of the unstable modes, $\bar{m}_b = 0.4469$.*

3.2. The general case. In the case when diffusion coefficients for temperature and vapor are not the same, analytical expression can still be obtained. However, it is too complicated to provide useful insights. We can, nevertheless, find the eigenvalues numerically by using the MATLAB routine `eig.m`. Again, for relevant parameter values, two of the three eigenvalues have nonpositive real parts. The real part of the third eigenvalue is plotted in Figure 3.1, where both the wave number and the eigenvalue are normalized as $\bar{s} = s/\bar{\beta}$ and $\bar{m} = m^2/\bar{\beta}$. It can be seen that there exists a range of wave numbers between zero and a finite value, indicated by the square in the figure, within which the perturbation will grow. Larger wave number disturbances beyond the critical wave number will decay. Furthermore, there exists a wave number which grows the fastest, indicated by the circle in the figure. Finally, the range of unstable frequencies increases linearly with $\sqrt{\bar{\beta}}$ and the growth rate increases linearly with $\bar{\beta}$. Therefore, a larger evaporation rate E (implies a larger $\bar{\beta}$) leads to a wider range of unstable modes and faster growth rates for all the unstable modes.

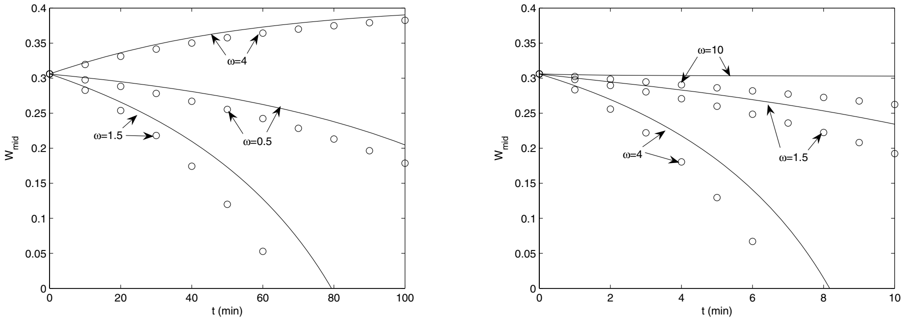
We now verify the results of the linear stability analysis by solving the reaction-diffusion model with Dirichlet conditions which permit the constant solutions $\theta = 0$, $V = V_0$, and $W = W_0$. In Figure 3.2(a), computed liquid water content at the end of 100 minutes is shown. The solutions are obtained by superimposing a small disturbance in the form of $\epsilon \cos[m(\xi - 0.5)]$ to a constant liquid water content W_0 with $m = 2\pi\omega$, $\omega = 0.5$, 1.5, and 4, and $\epsilon = 0.1$. The evaporation rate is $E = 0.0045$ and the range of unstable frequencies is $0 < \omega < 2.9738$. The most unstable mode is given by $\omega \approx 1.3345$, and the rate of growth is approximately $\exp\left(3.5056 \times 10^{-4} t\right)$ measured in dimensionless time. The results show that even though the disturbances with wave numbers below $\omega \approx 3$ grow with time, it takes a relatively long period of time (on the scale of 80 minutes for the fastest growing mode) to develop. On the other hand, the evaporation of all the liquid water takes about 60 minutes (Figures 2.1 and 2.2(a)). Therefore, all liquid water in the bread would have already evaporated before the instability develops and takes effect.

In Figure 3.2(b), computed liquid water content at the end of 10 minutes is shown



(a) (b)

FIG. 3.2. *Computational results of liquid water content subjected to disturbances.* (a) $E = 0.0045$ *with* $\omega = 0.5$, 1.5, *and* 4. (b) $E = 0.045$ *with* $\omega = 1.5$, 4, *and* 10. *The dashed lines indicate the initial states and the solid lines are the final solutions.*



(a) (b)

FIG. 3.3. *Damping rates for* (a) $E = 0.0045$ *and* (b) $E = 0.045$. *The liquid water content in the middle of the domain where the growth is the fastest is plotted for wave numbers* $\omega = 0.5$, 1.5, *and* 4 *in* (a) *and* $\omega = 1.5$, 4, *and* 10 *in* (b). *The circles are predictions by the linear stability analysis.*

for a larger evaporation rate $E = 0.045$, subjected to the same small disturbances. In this case, the range of unstable frequencies is $0 < \omega < 9.4$. The most unstable mode is given by $\omega \approx 4.22$, and the rate of growth is approximately $\exp\left(3.5056 \times 10^{-3} t\right)$ in dimensionless time. Compared to the previous case, this leads to a more rapid growth for the most unstable mode, on the order of 10 minutes in dimensional time, while it takes about 30 minutes for all the liquid water to evaporate (cf. Figure 2.2(b), where the effect of the instability is clearly visible).

The difference in the growth rates for the same disturbance under the two evaporation rates can be seen in Figure 3.3, where liquid water content in the middle of the domain is plotted. The prediction using linear stability analysis is also plotted. Note that the linear stability analysis is valid only near the constant state. Nevertheless, the two sets of data are not too far off. Thus, subject to random perturbation, the initially constant solution in the two-phase region will become unstable. The fastest growing mode eventually causes a multiple region of dry and two-phase zones, as shown in Figure 2.2.

Remark. From the analysis we see that this diffusive instability occurs for any positive β (or E). However, for relatively small β computational results in the previous section are satisfactory. One reason is that the positive eigenvalue is small as β is relatively small and the disturbances grow slowly. In addition, the oscillation (diffusive instability) is visible only when the wave length of the unstable modes is shorter than the domain size in x. This explains why no oscillation is observed in the case of the smaller β (or E).

4. Conclusion. The work in this paper is motivated by the simulation results in [18, 25], based on a multiphase model for bread baking proposed in [20]. This model allows simultaneous heat, vapor, and liquid water transfer by assuming that the phase change is instantaneous. However, previous studies [18, 25] showed that this model produces reasonable solutions only for specific choices of spatial and time step sizes. Reducing spatial and time step sizes usually leads to numerical instability and causes the solution to blow up.

By constructing a reaction-diffusion model and establishing a link between our model and the model in [20], we have identified the source of the instability associated with the model in [20]. We have shown that the instability observed in [18, 25] is a combination of two factors: the numerical instability as well as a diffusive instability. Using our reaction-diffusion model, we showed that the numerical instability can be eliminated by using a sufficiently small time step. This is due to the fact that the reaction-diffusion model separates the numerical instability from the diffusive instability. The diffusive instability, on the other hand, is an intrinsic feature of the model, as demonstrated by linear stability analysis and numerical tests. For relatively large evaporation rate, diffusive instability leads to an oscillatory solution with multiple regions of dry and two-phase zones.

Our analysis of the reaction-diffusion model also reveals that diffusive instability is related to the value of the evaporation rate, which is affected by the properties of the porous medium, such as the surface area of the pore space. This suggests that the phenomenon related to diffusive instability may be realized in the physical process of bread baking. Further experimental investigation is necessary to validate and improve our model for bread baking. On the other hand, the model itself is quite general and may be applicable to similar problems with simultaneous heat and mass transfer processes.

Finally, we wish to point out that a consistent instantaneous phase change model may be derived based on our reaction-diffusion model, using asymptotic analysis and β^{-1} as a small parameter, and this will be pursued in a future study.

Acknowledgments. We would like to thank the referees for many constructive suggestions which have helped to improve the quality of our manuscript. We also wish to acknowledge beneficial discussions with Dr. Jonathan Wylie during the preparation of the manuscript.

REFERENCES

[1] D. Alemani, B. Chopard, J. Galceran, and J. Buffle, *LBGK method coupled to time splitting technique for solving reaction-diffusion processes in complex systems*, Phys. Chem. Chem. Phys., 7 (2005), pp. 3331–3341.

[2] G. Carey, N. Fowkes, A. Staelens, and A. Pardhanani, *A class of coupled nonlinear reaction diffusion models exhibiting fingering*, J. Comput. Appl. Math., 166 (2004), pp. 87–99.

[3] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, Clarendon Press, Oxford, UK, 1959.

[4] P. G. Ciarlet and J.-L. Lions, *Handbook of Numerical Analysis, Vol.* 1, North–Holland, Amsterdam, 1990.

[5] J. Crank, *The Mathematics of Diffusion*, Clarendon Press, Oxford, UK, 1975.

[6] J. Crank, *Free and Moving Boundary Problems*, Oxford University Press, Oxford, UK, 1987.

[7] H. S. Davis, *Theory of Solidifications*, Cambridge University Press, Cambridge, UK, 2001.

[8] S. Descombes, *Convergence of a splitting method of high order for reaction-diffusion systems*, Math. Comp., 70 (2000), pp. 1481–1501.

[9] U. De Vries, P. Sluimer, and A. H. Bloksma, *A quantitative model for heat transport in dough and crumb during baking*, in Cereal Science and Technology in Sweden, N.-G. Asp, ed., STU Lund University, Lund, Sweden, 1989, pp. 174–188.

[10] A. C. Fowler, *Mathematical Models in the Applied Sciences*, Cambridge University Press, Cambridge, UK, 1997.

[11] M. L. Frankel, G. Kovacic, V. Roytburd, and I. Timofeyev, *Finite-dimensional dynamical system modeling thermal instabilities*, Phys. D, 137 (2000), pp. 295–315.

[12] M. Frankel, V. Roytburd, and G. Sivashinsky, *A sequence of period doublings and chaotic pulsations in a free boundary problem modeling thermal instabilities*, SIAM J. Appl. Math., 54 (1994), pp. 1101–1112.

[13] F. Jones, *Evaporation of Water*, CRC Press, Boca Raton, FL, 1991.

[14] D. A. Kessler, J. Koplik, and H. Levine, *Pattern selection in fingered growth phenomena*, Adv. in Phys., 37 (1988), pp. 255–339.

[15] K. A. Landman and C. P. Please, *Modelling moisture uptake in a cereal grain*, IMA J. Math. Bus. Ind., 10 (2000), pp. 265–287.

[16] M. J. McGuinness, C. P. Please, N. Fowkes, P. McGowan, L. Ryder, and D. Forte, *Modelling the wetting and cooking of a single cereal grain*, IMA J. Math. Bus. Ind., 11 (2000), pp. 49–70.

[17] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1989.

[18] G. G. Powathil, *A Heat and Mass Transfer Model for Bread Baking: An Investigation Using Numerical Schemes*, M.Sc. Thesis, Department of Mathematics, National University of Singapore, Singapore, 2004.

[19] R. P. Singh and D. R. Heldman, *An Introduction to Food Engineering*, 3rd ed., Academic Press, London, 2001.

[20] K. Thorvaldsson and H. Janestad, *A model for simultaneous heat, water and vapor diffusion*, J. Food Eng., 40 (1999), pp. 167–172.

[21] K. Thorvaldsson and C. Skjoldebrand, *Water diffusion in bread during baking*, Lebensm.-Wiss. u.-Technol., 31 (1998), pp. 658–663.

[22] C. H. Tong and D. B. Lund, *Microwave heating of baked dough products with simultaneous heat and moisture transfer*, J. Food Eng., 19 (1993), pp. 319–339.

[23] B. Zanoni, C. Peri, and S. Pierucci, *A study of the bread-baking process* I: *A phenomenological model*, J. Food Eng., 19 (1993), pp. 389–398.

[24] B. Zanoni, C. Peri, and S. Pierucci, *Study of bread baking process* II: *Mathematical modelling*, J. Food Eng., 23 (1994), pp. 321–336.

[25] W. Zhou, *Application of FDM and FEM to solving the simultaneous heat and moisture transfer inside bread during baking*, Int. J. Comput. Fluid Dyn., 19 (2005), pp. 73–77.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 239–252

STATISTICAL RECONSTRUCTION OF VELOCITY PROFILES FOR NANOPARTICLE IMAGE VELOCIMETRY*

CHRISTEL HOHENEGGER[†] AND PETER J. MUCHA[‡]

Abstract. Velocities and Brownian effects at nanoscales near channel walls can be measured experimentally in an image plane parallel to the wall by evanescent wave illumination techniques [R. Sadr, M. Yoda, Z. Zheng, and A. T. Conlisk, *J. Fluid Mech.*, 506 (2004), pp. 357–367], but the depth of field in this technique is difficult to modify. Assuming mobility of spherical particles dominated by hydrodynamic interaction between particle and wall, the out-of-plane dependence of the mobility and in-plane velocity are clearly coupled. We investigate such systems computationally, using a Milstein algorithm that is both weak- and strong-order 1. In particle image velocimetry (PIV), image pairs are cross-correlated to approximate the mean displacement of n matched particles between two windows. For comparison, we demonstrate that a maximum likelihood algorithm can reconstruct the out-of-plane velocity profile, as specified velocities at multiple points, given known mobility dependence and perfect mean measurements. We then test this reconstruction for noisy measurements as might be encountered in experimental data. Physical parameters are chosen to be as close as possible to the experimental parameters while we consider three types of velocity profiles (linear, parabolic, and exponentially decaying).

Key words. stochastic differential equations, maximum likelihood estimate, particle image velocimetry, velocity profile, wall effects

AMS subject classifications. 76M35, 76M25, 76D99, 60G99, 65C35

DOI. 10.1137/050648043

1. Introduction.

Fluid velocities in a channel can be measured by illumination and imaging of tracer particles under the assumption that they follow the flow, with corrections possibly applied for effects including, e.g., the near-wall relationship between particle translation and rotation [6]. For laser-Doppler velocimetry, Fuller et al. [3] showed that it is possible to reconstruct the velocity gradient in a laminar flow using light-scattering spectroscopy. This requires the knowledge of the intensity function and the technical capacity of turning the sample to get a proper angle of illumination. At microscales, Meinhart et al. (see [15], [11], [10]) developed an illumination technique, particle image velocimetry (PIV), to replace spectroscopy, where the tracers are illuminated using multiple laser sheets and the velocity profile is computed as means over successive windows using cross-correlation techniques. Again the sample has to be properly illuminated so that particles remain in the focal plane. If it is possible to turn the sample, all components of the mean velocity profile can be obtained.

At nanoscales, including the near-wall region of microchannels, Sadr, Li, and Yoda [12] and Sadr et al. [14] extend PIV to flows illuminated with evanescent waves generated by total internal reflection at the wall. Image pairs are captured on a cam-

*Received by the editors December 20, 2005; accepted for publication (in revised form) June 18, 2007; published electronically October 24, 2007. This work was partially supported by the Department of Energy (DE-FG-02-03ER25567), the National Science Foundation (DMS-0204309), and start-up funds provided by the Institute for Advanced Materials and the Department of Mathematics at the University of North Carolina at Chapel Hill.

http://www.siam.org/journals/siap/68-1/64804.html

[†] Courant Institute of Mathematical Sciences, University of New York, New York, NY 10012 (choheneg@cims.nyu.edu).

[‡] Department of Mathematics & Institute for Advanced Materials, University of North Carolina, Chapel Hill, NC 27599 (mucha@unc.edu).

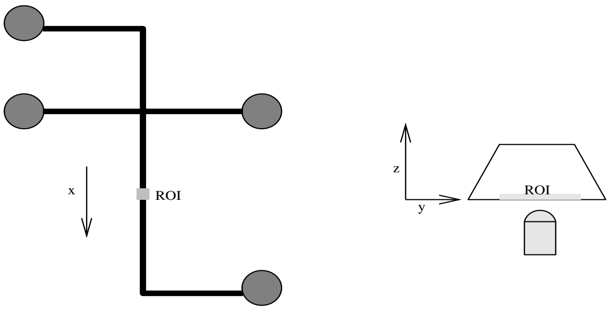


FIG. 1.1. *Experimental setup: Region of interest, flow direction, and wall location.*

era with a time interval ∼ 2ms, and in-image-plane mean velocities are obtained using cross-correlation techniques. Figure 1.1 illustrates the experimental setup around the region of interest and some of the inherent experimental restrictions. While Sadr, Li, and Yoda [12] show that Brownian diffusion can cause additional errors in the measurements as particles drop in and out of the imaged window, only limited information about the velocity along or dependence on the out-of-plane coordinate has been experimentally accessed recently from the brightness of the images and the decay of the illumination function [6], [7], [9]. Questions remain about the accuracy and range of validity of processing based on image intensity, especially in the presence of the highly heterogeneous distribution of fluorescent dye on the tracer particle surfaces, while background noise pollutes the images causing reconstruction of velocity profile based solely on intensity to be extremely challenging (see Li, Sadr, and Yoda [9]).

Another dominant difficulty of these measurements arises from the nonconstant diffusion tensor induced by the proximity of the wall. Both the in-plane and out-of-plane diffusion components strongly depend on the distance from the wall (see Figure 2.1(a)). While this dependence is well understood in terms of the hydrodynamic interaction between particle and wall [1], the effect of such diffusion on the resulting measurements has been only recently addressed experimentally (see, e.g., [14]). Meanwhile, significant effort has been put into extending the range of validity of particle image velocimetry (PIV) and particle tracking velocimetry (PTV) to smaller ranges of particles. For example, Guasto, Huang, and Breuer [4] use a statistical approach assuming nearly constant diffusion to eliminate experimental noise (drop-in/-out, mismatch, particles blinking) and obtain a distribution of velocities. Using a similar idea with nonconstant diffusion, Jin et al. (see [6], [7]) notice in their attempts to assess slip at the wall that a nonnegligible difference exists between the apparent measured mean velocities and the imposed shear rate. Interpretations of such studies are further complicated by the measured velocities representing those across a spatially extended region away from the wall, typically with little mechanism for modifying the extent of such a region.

In this work, we show that it is possible to reconstruct the out-of-plane dependence of the in-plane velocity component as a collection of velocities at specified out-of-plane distances (typically five points), based solely on in-plane images. The unique assumption leading to the statistical reconstruction of the out-of-plane component concerns the out-of-plane distribution of the particles between two window measurements. For simplicity, here we assume that the computational and observational domains are the same, thereby eliminating errors due to particle drop-in/-out for this proof-of-principle demonstration.

The next two sections present the particle model incorporating the most important parameters from the experiments and its numerical simulation with a strong order 1 scheme. In the fourth section we develop the algorithm for the perfect case, in which the mean displacements are known exactly. The reconstruction is illustrated for three different specified velocity test profiles (see Figure 2.1(b)): a linear profile, a parabolic profile, and a profile exponentially decaying to the bulk velocity. In the fifth section we extend the idea to noisy mean displacements obtained through the consideration of a measurement error similar to the one reported for cross-correlation from simulated images (PIV techniques). Finally, we discuss the limitations of the model and its possible improvement.

2. Particle model. We test the algorithmic reconstruction on a simple stochastic model of particle motion. Each particle is assumed to have a fixed radius a ($a = 50\,\mathrm{nm}$ in the demonstrations here), and the hydrodynamic interaction between the wall and a particle is captured by the model for mobility in terms of the out-of-plane coordinate perpendicular to the wall, z. We ignore particle-particle hydrodynamic interactions, which are relatively small for the dilute particle volume fractions of the experiments. The tracer particles are dragged along with the fluid flow; additional interactions between particles and the wall are feasible but not included here.

We consider a system of n ($n = 64$) Brownian particles obeying Stokes drag relations, linearly dependent on the velocity. For time steps Δt bigger than the force relaxation time, Ermak and McCammon [2] show that the displacement Δr_i can be expressed as

$$\Delta r_i = \sum_{j=1}^{3n} \frac{\partial D_{ij}}{\partial r_j} \Delta t + \sum_{j=1}^{3n} \frac{D_{ij} F_j}{k\Theta} \Delta t + W_i(\Delta t), \quad i = 1, \ldots, 3n, \tag{2.1}$$

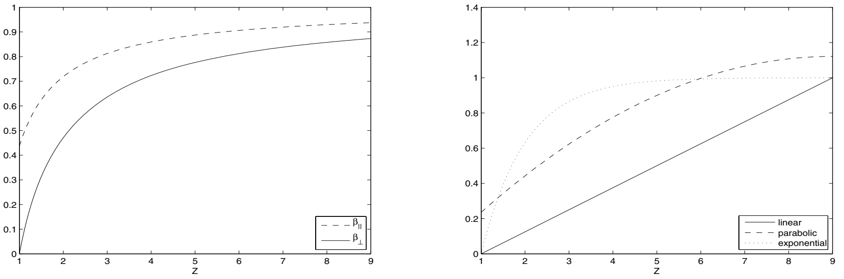
where $W_i(\Delta t)$ is a random displacement with a Gaussian distribution function whose average value is zero and whose variance-covariance matrix is $2\mathbf{D}\Delta t$, $\mathbf{D}$ is the diffusion tensor, $\mathbf{F}$ are the external forces, k is the Boltzmann constant, and Θ is the temperature. The Brownian displacement can be expressed as [2]

$$W_i(\Delta t) = \sum_{j=1}^{i} \sigma_{ij} dW_j, \quad \sigma = \sqrt{\mathbf{2D}}, \quad dW_j = \mathcal{N}(0, \Delta t), \quad j = 1, \ldots, 3n, \tag{2.2}$$

where $\mathcal{N}(\mu, \sigma^2)$ indicates Gaussian random variables of mean μ and variance σ^2.

While we ignore particle-particle interactions, hindered Brownian diffusion due to hydrodynamic particle-wall interactions is an important effect for the near-wall conditions in the experiments. A first approximation of the nonconstant diffusion tensor is obtained by the methods of image singularities for Stokes flows, valid for particle center-to-wall distances, z, that are large compared to the particle radius, a. For our model system here, we include for simplicity only the lowest-order a/z corrections for diffusion components parallel to the planar wall but instead employ the Bevan–Prieve relation [1] for the out-of-plane diffusion perpendicular to the wall, both because of its experimental verification and because it includes the physically impermeable property that the diffusion coefficient goes to zero for a spherical particle touching the wall ($z = a$):

$$\mathbf{D} = \frac{k\Theta}{6\pi\mu a} \begin{pmatrix} 1 - \frac{9}{16}\frac{a}{z} & 0 & 0 \\ 0 & 1 - \frac{9}{16}\frac{a}{z} & 0 \\ 0 & 0 & \frac{6z^2 - 10az + 4a^2}{6z^2 - 3az - a^2} \end{pmatrix} = D_\infty \beta(z), \tag{2.3}$$



(a) Components of $\beta(\frac{z}{a})$

(b) Velocity profiles $\gamma(\frac{z}{a})$

FIG. 2.1. (a) *Dimensionless diffusion coefficients perpendicular $\beta_\perp$ and parallel $\beta_{||}$ to the wall.* (b) *Dimensionless velocity profiles γ for the linear, parabolic, and exponentially decaying test cases.*

for the three components of each individual particle, where $D_\infty = k\Theta/(6\pi\mu a)$ is the Stokes–Einstein relation in the bulk limit far from the wall, the z-component is along the direction perpendicular to the wall, and z is the distance between the center of the particle and the wall.

Here we consider a simulated channel of height H with three different test flows $u_\infty \gamma(z)\mathbf{e_x}$ with "bulk velocity" u_∞, as might be encountered in shear flow, pressure-driven flow, and electroosmotically pumped flow, respectively: linear flow $\gamma(z) = \frac{1}{H-a}(z-a)$, parabolic flow $\gamma(z) = \frac{4}{(2H-a)^2} z(2H-z)$, and an exponentially decaying profile $\gamma(z) = 1 - \exp(1 - z/a)$ (Figure 2.1(b)). For simplicity, we consider the above flows to be the force-free velocity profiles of the tracer particles themselves, with the hydrodynamic balance given for the external forces on the particles $\mathbf{F} = k\Theta\mathbf{D}^{-1} u_\infty \gamma(z)\mathbf{e_x}$. In the physical experiments, additional corrections are required to relate the force-free velocities of the tracers to those of the underlying flow (see, e.g., [6]); we assume such corrections can be imposed if the particle velocities are accurately measured, proceeding with simulations of imposed particle velocities whose velocity profiles we will reconstruct. Here we include only flow along one (x) of the two in-plane directions parallel to the wall, but since the statistical reconstructions below will not process any displacements along the other in-plane direction (y), the methods presented here can be equivalently applied to measure any in-plane flow profile dependent on the out-of-plane (z-) direction.

Our model stochastic ODE Langevin equation for the displacement of an individual particle is then

$$dx = u_\infty \gamma(z)\, dt + \sqrt{2D_\infty \beta_{||}(z)}\, dW_1, \tag{2.4}$$

$$dy = \sqrt{2D_\infty \beta_{||}(z)}\, dW_2, \tag{2.5}$$

$$dz = D_\infty \frac{d\beta_\perp(z)}{dz}\, dt + \sqrt{2D_\infty \beta_\perp(z)}\, dW_3. \tag{2.6}$$

Letting T be the time elapsed between two PIV-window observed images, we set T and the radius a as the characteristic time and length scales, respectively. Letting $x = aX$, $y = aY$, $z = aZ$, and $t = T\tau$ define the dimensionless variables, the resulting dimensionless parameters are $\Pi_1 = \frac{u_\infty T}{a}$ and $\Pi_2 = \frac{D_\infty T}{a^2}$. For our tests reported here, we select $T = 2^{-9}$s, giving $\Pi_2 = 4$ at $\Theta = 300$, with u_∞ selected to give $\Pi_1 = 3$, of a scale typical to those of the experiments. Our dimensionless Langevin model, with $d\mathbf{W} = \mathcal{N}(0, d\tau)$, becomes

$$dX = \Pi_1 \gamma(aZ)\, d\tau + \sqrt{2\Pi_2 \beta_{||}(aZ)} dW_X, \tag{2.7}$$

$$dY = \sqrt{2\Pi_2 \beta_{||}(aZ)}\, dW_Y, \tag{2.8}$$

$$dZ = \Pi_2 \frac{d\beta_\perp(aZ)}{dZ}\, d\tau + \sqrt{2\Pi_2 \beta_\perp(aZ)}\, dW_Z. \tag{2.9}$$

3. Numerical simulation. Equations (2.7), (2.8), and (2.9) form a system of stochastic differential equations of the form $d\mathbf{X} = \mathbf{f}(\mathbf{X}, t)d\tau + \mathbf{g}(\mathbf{X}, t)d\mathbf{W}$. We solve it with a Milstein scheme of weak and strong order of convergence one. The coupling of the system through the Z-component yields a nondiagonal noise in the stochastic differential equation sense. The resulting Milstein scheme is given by [8] (see also [5]):

$$X_{j+1} = X_j + f_{1,j}\Delta\tau + g_{11,j}\Delta W_{1,j} + \frac{1}{2} g_{11,j} \left.\frac{dg_{11}(z)}{dZ}\right|_{Z=Z_j} I_{(3,1)}, \tag{3.1}$$

$$Y_{j+1} = Y_j + f_{2,j}\Delta\tau + g_{22,j}\Delta W_{2,j} + \frac{1}{2} g_{22,j} \left.\frac{dg_{22}(Z)}{dZ}\right|_{Z=Z_j} I_{(3,2)}, \tag{3.2}$$

$$Z_{j+1} = Z_j + f_{3,j}\Delta\tau + g_{33,j}\Delta W_{3,j} + \frac{1}{2} g_{33,j} \left.\frac{dg_{33}(Z)}{dZ}\right|_{Z=Z_j} \Big((\Delta W_{3,j})^2 - \Delta\tau\Big), \tag{3.3}$$

where $f_{i,j} = f_i(Z_j)$, $g_{ii,j} = g_{ii}(Z_j)$, and $I_{(3,i)}$ $(i = 1, 2)$ are the double Itô stochastic integrals defined as $I_{(3,i)} = \int_{t_k}^{t_{k+1}} \int_{t_k}^{t_{k+1}} dW_3 dW_i$. Since $I_{(3,i)}$ has no closed analytical solution, we approximate $I_{(3,i)}$ as the solution of a stochastic differential system:

$$I_{(3,i)} = X_i(t_{k+1}), \quad \text{where} \quad \begin{cases} dX_i = X_3 dW_i, & X_i(t_k) = 0, \\ dX_3 = dW_3, & X_3(t_k) = 0. \end{cases} \tag{3.4}$$

Equation (3.4) is solved using Euler–Maruyama steps, the stochastic equivalent of a forward Euler step, with strong order of convergence $\frac{1}{2}$. To ensure convergence to an accurate solution for the entire system, we choose $\Delta\tau = 2^{-10}$ in (3.1)–(3.3), resolving each Itô integral $I_{(3,i)}$ with 2^{10} time steps in (3.4).

4. Reconstruction with perfect means. We start our proof-of-principle calculations by statistically reconstructing velocity profiles based on perfectly observed mean displacements. By this we mean that the true position of each particle is known and the mean displacement of the n particles between two image-pair windows is computed exactly. Cross-correlation processing of image pairs in PIV extracts, up to various sources of error, the mean displacement of the "matched" particles—those that contribute to both images. If the true displacement of each particle could be experimentally determined, as in particle tracking, then the same reconstruction ideas below do apply, but our various tests indicated that such particle tracking does not improve the results, and may even require greater quantities of data than statistical reconstruction based on mean displacements, presumably because of the statistical reliance below on clearly characterized Brownian displacements.

Let $f_{\Delta X}$ be the probability distribution function of a displacement ΔX. From (2.7) the X-displacement depends on the Z-position. Therefore we define $f_{\Delta X|Z}$ to be the probability density function of ΔX given Z. Then

$$f_{\Delta X} = \int f_{\Delta X|Z} f_Z dZ, \tag{4.1}$$

where f_Z is the probability density function of particles in Z. Because we restrict ourselves to the case where the computation and observation domain are the same, we make the following assumption about the Z distribution:

$$f_Z = \frac{a}{H-a}\chi_{[1,H/a]}, \tag{4.2}$$

where χ_I is the characteristic function on an interval I.

Next we make a fundamental simplifying modeling assumption for the reconstruction: that the particle displacements over the time T between two consecutive windows can be approximated by an Euler step of the form

$$\Delta X \approx \Pi_1\gamma(aZ) + \sqrt{2\Pi_2\beta_{||}(aZ)}dW \quad \text{with } dW = \mathcal{N}(0,1), \tag{4.3}$$

where $\gamma(aZ)$ is the unknown velocity profile. From (4.3) we conclude that

$$f_{\Delta X|Z} = \frac{1}{2\sqrt{\pi\Pi_2\beta_{||}(aZ)}}e^{-\frac{(\Delta X-\Pi_1\gamma(aZ))^2}{4\Pi_2\beta_{||}(aZ)}}. \tag{4.4}$$

Finally, using (4.1), (4.2), and (4.4) we find that

$$f_{\Delta X} = \frac{a}{2\sqrt{\pi\Pi_2}(H-a)}\int_1^{\frac{H}{a}} \frac{1}{\sqrt{\beta_{||}(aZ)}}e^{-\frac{(\Delta X-\Pi_1\gamma(aZ))^2}{4\Pi_2\beta_{||}(aZ)}} dZ. \tag{4.5}$$

Let $\overline{\Delta X}$ be the mean displacement of n matched particles over a window and let f_S be the probability density function of $n\overline{\Delta X}$. Now let f be the joint probability density function of N measured $n\overline{\Delta X}$. A standard result of probability, together with the assumption of independence between two windows measurement, yields

$$f_S = f_{\Delta X} * \cdots * f_{\Delta X} \quad \text{and} \quad f = \prod_{i=1}^{N} f_S, \tag{4.6}$$

where $*$ denotes the convolution. This independence assumption is, of course, incorrect, since consecutive $\overline{\Delta X}$ displacements are correlated by the continuity-in-time of the particles z positions; we nevertheless proceed under this modeling assumption, counting on the effect of the correlations to be sufficiently small.

Figure 4.1 compares the histogram of a $n\overline{\Delta X}$ data set with $Z \in [1, H/a]$ for the parabolic test profile with the probability density function obtained with (4.5) (dashed line). The integral in (4.5) is computed with a Gauss–Legendre quadrature formula under the assumption of a uniform z-distribution. This demonstrates the reasonable validity of assumption (4.3). Going even further, the dotted line in Figure 4.1 represents the probability density function obtained by fitting the data set $n\overline{\Delta X}$ for $Z \in [1, H/a]$ by a single Gaussian. The differences between the integrated Gaussian (dashed line) and the fitted Gaussian (dotted line) are minimal in the height and location of the peak. These minimal distinctions make the desired optimization highly sensitive. Despite these expected difficulties, we nevertheless continue both with our assumption (4.3) and the fundamental ideas of the velocity profile reconstruction.

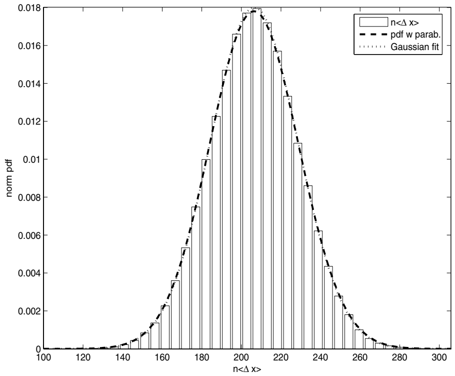


FIG. 4.1. *Comparison of the parabolic profile between the histogram of the distribution of* $n\overline{\Delta X}$*, the probability density function* (4.5)*, and a fitted Gaussian.*

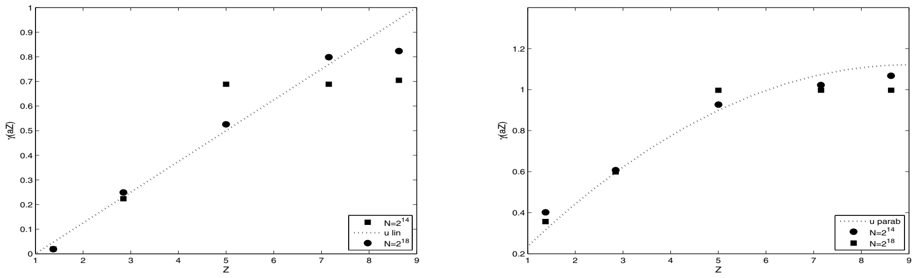
Given N measured mean values $n\overline{\Delta X}$, Π_1, Π_2, and $\beta_{||}(aZ)$, the maximum likelihood estimate of $\gamma(aZ)$ is the value of $\gamma(aZ)$ that makes the observed means most likely. Led by the independence assumption, we define the log-likelihood function

$$\phi(\{\gamma_j\}_{j=1}^{M}) = -\ln f \overset{(4.6)}{=} -\sum_{i=1}^{N} \ln f_S(n\overline{\Delta X}), \tag{4.7}$$

where M is the number of discrete points Z at which we estimate γ. The most likely values for γ_j, $j = 1, \ldots, M$, are obtained by minimizing the log-likelihood function ϕ (4.7) of the M variables $\gamma_1, \ldots, \gamma_M$ for a data set $n\overline{\Delta X}$ of size N.

The statistical reconstruction problem has thus been reduced to two numerical algorithms. First, we evaluate the probability density function f_S in (4.6) by repeated convolution of the probability density function $f_{\Delta X}$ as in (4.5), computing the integral by Gauss–Legendre quadrature for given γ_j values at the Legendre collocation points $Z_j \in [1, H/a]$ for $j = 1, \ldots, M$. We subsequently minimize the function ϕ (4.7) with a direct simplex algorithm penalizing solutions that do not produce an increasing sequence, since we know that the velocity profile is increasing to the bulk velocity away from the wall. We also experimented with the alternative scheme of minimizing ϕ over low-order polynomials for $\gamma(aZ)$ but did not obtain results any more promising than those presented below. Not surprisingly, the minimization routine is highly sensitive to the choice of the initial guess. Therefore, when reconstructing velocity values for a small number of points M, we first search the M-dimensional space for a suitable initial guess by evaluating the function at a fixed number of increasing grid points. When reconstructing velocity values at M points for M larger (say, $M \geq 7$), we interpolate the initial guess from the reconstructed velocity values for smaller M.

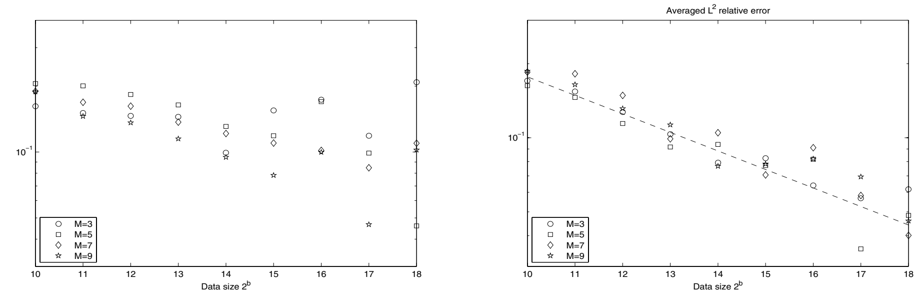
Figure 4.2 illustrates the reconstruction for the linear and parabolic profiles at five points ($M = 5$) for two different data sizes. Since the accuracy of the reconstructed points does not appear to improve when N increases from 2^{14} to 2^{18}, we are motivated to instead consider breaking one block of data up into separate reconstructions over each of B blocks of size 2^b. Figure 4.3 contains semilog plots of the L^2 relative error of the reconstructed γ_j, $j = 1, \ldots, M$, with respect to the true $\gamma(aZ_j)$, $j = 1, \ldots, M$, for individual blocks, the errors averaged over the number of blocks B for four different values of M (3, 5, 7, and 9 points). For the linear profile on the left, we observe the same behavior as in Figure 4.2, namely, that increasing the data size does not



(a) Linear profile

(b) Parabolic profile

FIG. 4.2. *Velocity profile reconstruction at $M = 5$ points for the linear and parabolic test profiles with data set sizes $N = 2^{14}$ and $N = 2^{18}$.*



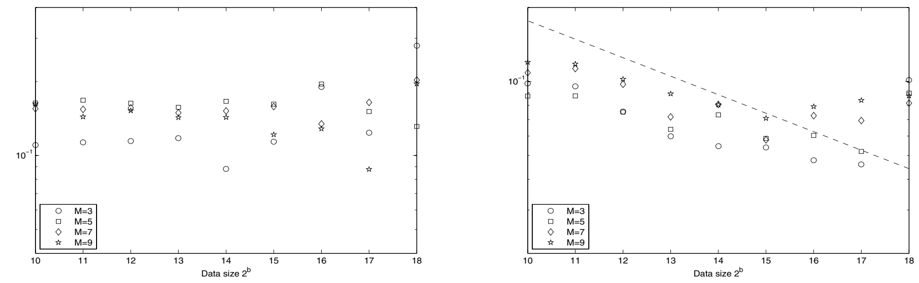
(a) Linear profile

(b) Parabolic profile

FIG. 4.3. *L^2 relative error averaged over the number of blocks B of size 2^b at $M = 3$, $M = 5$, $M = 7$, $M = 9$.*

predictably improve the accuracy of the reconstruction after some point. For the parabolic profile on the right, we find a decay of the relative error in the function of the data size b which appears to be roughly $(2^b)^{-1/4}$ up to another apparent stagnation of the decaying error for data sizes larger than 2^{14} or 2^{15}. We next consider the plot of the L^2-norm of the relative error of the block-averaged reconstructed values $\overline{\gamma_j} = \frac{1}{B}\sum_{k=1}^{B}\gamma_j^k$, $j = 1, \ldots, M$ (where γ_j^k is the reconstructed value at Z_j for the block B_k), with respect to the true $\gamma(aZ_j)$, $j = 1, \ldots, M$ (Figure 4.4). We deduce from the relative errors of the block-averaged values, especially for the parabolic test profile, that errors can be reduced by such averaging over a limited number of blocks. As above, the parabolic profile follows a decay close to $(2^b)^{-1/4}$ up to 2^{15}. We do not at present have any explanation for this particular power law of decay. We conclude that the best reconstruction on a data set of the size $N = 2^{18}$ will be achieved when the average of the reconstructed profile is done over 8 or 16 blocks. We also notice that increasing the number of discrete points to $M = 7$ or $M = 9$ does not produce significantly different normed errors but provides more detail about the calculated profile at the cost of a lengthier computation.

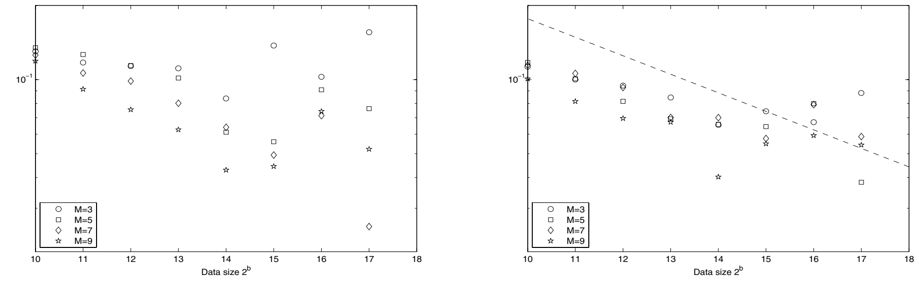
In practice, of course, the goal of the reconstruction is to obtain an approximation of the velocity profile, the true profile being unknown. So, finally, we compare the L^1-norm of the variance of the reconstructed profiles from the individual blocks, plotted



(a) Linear profile

(b) Parabolic profile

FIG. 4.4. *L^2 relative error with γ averaged over the number of blocks B of size 2^b at $M = 3$, $M = 5$, $M = 7$, $M = 9$.*



(a) Linear profile

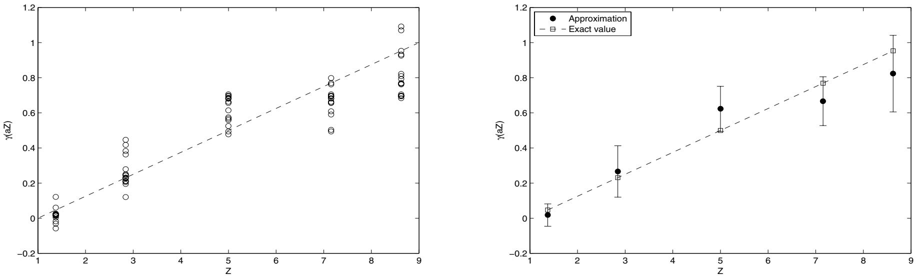
(b) Parabolic profile

FIG. 4.5. *L^1-norm of the variance of γ averaged over the number of blocks B at $M = 3, 5, 7, 9$ for the linear and parabolic profile.*

versus the exponent b in the block size 2^b. Again, we note for both the linear and the parabolic profiles that the variance increases for data sizes bigger than 2^{15}. That is, the L^1-norm of the variance of reconstructed values from individual data blocks appears to trend very similarly to the true L^2-norm errors, and so we propose using the former as a stand-in for the latter in deciding how to block-divide the data in the present setting. Figure 4.5 thereby confirms that a better result can be both obtained and recognized here when averaging over $B = 8$ or $B = 16$ blocks corresponding to blocks of size 2^{15} or 2^{16}. We remark that there are numerous sources of error in the present reconstruction, including errors in the numerical integration, the numerical convolution, and the minimization itself.

Using the result of the block-averaging technique investigated in the previous three error plots (Figures 4.3, 4.4, and 4.5) we can now reconstruct the velocity profile at five points, $M = 5$, for the linear case with $B = 16$ blocks. In Figure 4.6 we examine both the spread of the values obtained for each block and the average $\overline{\gamma_j}$, $j = 1, \ldots, M$, and standard deviation (plotted as 90% confidence interval error bars for the block reconstruction values).

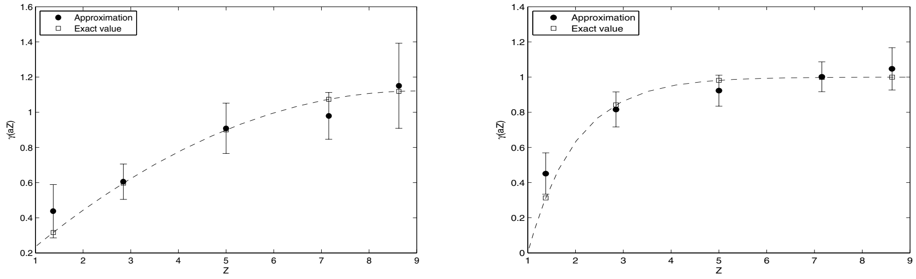
Finally, we apply the block-averaging technique on the parabolic (Figure 4.7(a)) and exponentially decaying (Figure 4.7(b)) test velocity profiles at $M = 5$ points and $B = 16$ blocks. During our proof-of-principle calculations, we sometimes encountered



(a) Scatter plot

(b) Errorbar plot

FIG. 4.6. *Scatter plot of the different block reconstructed values* γ_j^k*,* $j = 1, \ldots, M$*,* $k = 1, \ldots, 16$*, and block-averaged* $\overline{\gamma_j}$*,* $j = 1, \ldots, M$*, with* 90% *confidence interval for the linear test profile with* $B = 16$ *and* $M = 5$*.*



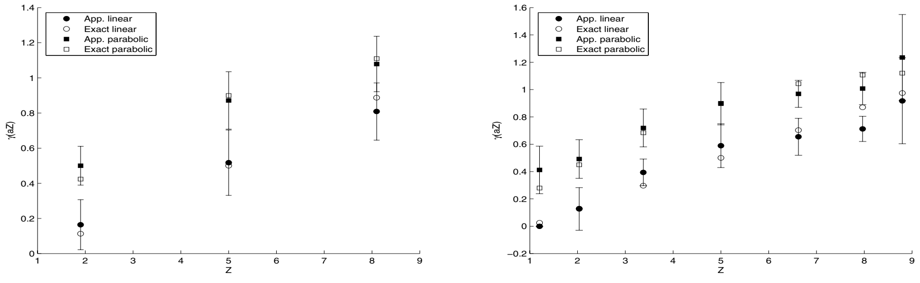
(a) Parabolic profile

(b) Exponentially decaying profile

FIG. 4.7. *Errorbar reconstruction for the parabolic and exponentially decaying profile with* $B = 16$*,* $M = 5$*, and* 90% *confidence interval.*

data sets for which the reconstruction performed particularly poorly, as evidenced by clear jumps in the reconstructed values as might suggest discontinuous velocity profiles. Such poorly performing data was a simple consequence of the state of the random number generator in the simulations; presumably similarly quirky experimental data is not wholly uncommon, and so such reconstructions must therefore, of course, always be questioned, particularly if they indicate highly unlikely results. Finally, we additionally remark that the near-wall region velocity profile is usually assumed to be linear or parabolic, and the exponential case is experimentally unlikely for the present purposes except when the imaged region is large compared to the scale of electroosmotic layers.

To conclude this section we plot the averaged reconstructed mean $\overline{\gamma_j}$, $j = 1, \ldots, M$ (full symbols), compared to their true values (open symbols) at three ($M = 3$) and seven ($M = 7$) points for both the linear (square) and the parabolic (circle) test profiles together in the same figure (Figure 4.8). Figures 4.7 and 4.8 clearly demonstrate that we are able to statistically reconstruct the main behaviors of and distinguish between different profiles (linear, parabolic, and exponentially decaying test profiles) using multiple collocation points ($M = 3, 5, 7$) across the measured region.



(a) $M = 3$

(b) $M = 7$

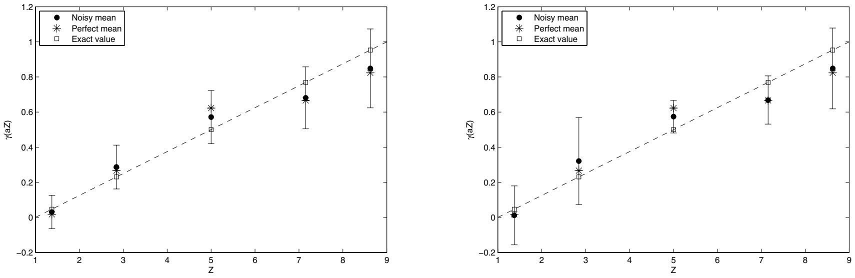
FIG. 4.8. *Block-averaged reconstruction* ($B = 16$) *at* $M = 3$ *and* $M = 7$ *points for the linear (square) and parabolic (circle) test profiles (full symbols) together with their respective exact values (empty symbols), with* 90% *confidence intervals on the reconstructed values.*

5. Reconstruction with cross-correlated velocities. In the previous section we used perfect mean displacements between two image windows. In this section, first we describe the idea behind PIV approximate mean measurements and then how they influence the reconstruction algorithm.

PIV is an illumination technique combined with image processing to obtain components of the mean velocity by measuring the mean displacement over a lag time. At microscales, the sample is illuminated with laser sheets and cross-correlation techniques (see [15], [11], [10]) producing a three-dimensional velocity profile. In nano-PIV (nPIV), total internal reflection fluorescence microscopy is used to image tracer particles [12], [14]. When light undergoes total internal reflection for angle of incidence larger than the critical angle, an evanescent wave is created and propagates parallel to the interface with an exponentially decaying intensity. Zettner and Yoda [16] report errors in the approximation of the mean of the order of 10%, while Sadr, Li, and Yoda [13] estimate that nPIV leads to an error of about 6% in the approximation of the mean x-displacement. We remark that for the well-established technique of μPIV, Meinhart, Wereley, and Santiago [11] conclude that the ensemble-averaged displacements lie within 2% of their true values.

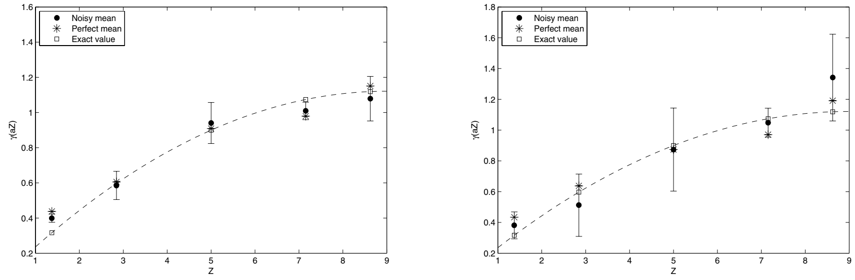
The parameters in our computer simulations are chosen to closely match experimental parameters [12]: the sizes of the region of interest are $\delta x = 25\,\mu\text{m}$, $\delta y = 5\,\mu\text{m}$, and $\delta z = 450\,\text{nm}$, the radius of a particle is $a = 50\,\text{nm}$, and the number of particles is 64. We note that both background image noise and particle drop-in/-out between the two images also degrade the PIV measurement, but we ignore both effects here for simplicity. Therefore the particles are uniformly distributed in the z-direction between two measurements. Once the image matrix is generated, the approximate x- and y-displacement over a window is determined using cross-correlation [15], [11], [10]. The cross-correlation function is the two-dimensional discrete convolution of two image matrices. The location of the maximum peak of the cross-correlation function gives the mean x- and y-displacement between two windows. To gain subpixel accuracy, a Gaussian surface fitting algorithm with 8 to 11 neighbors is typically used.

Because the previously described technique requires significant experimental agility in the choice of the size of the window over which the displacements are obtained and in the ratio of overlapping of the windows, a threshold criteria for eliminating bad displacement vectors has to be adopted (see [7]). Instead of using a computer



(a) PIV data with 5% relative error (b) PIV data with 10% relative error

FIG. 5.1. *Reconstruction with* 5% *and* 10% *approximated PIV means averaged over* $B = 16$ *blocks and compared with reconstruction from perfect means for the linear case.*



(a) PIV data with 5% relative error (b) PIV data with 10% relative error

FIG. 5.2. *Reconstruction with* 5% *and* 10% *approximated PIV means averaged over* $B = 16$ *blocks and compared with reconstruction from perfect means for the parabolic case.*

analogue to PIV techniques with a threshold which will lead to the generation of more data, here we mimic the effect of these additional experimental errors by adding normally distributed relative errors with standard deviations of 5% and 10% with respect to the overall mean of the perfect mean displacements from our simulations. Then we apply the statistical reconstruction algorithm at $M = 5$ with averaging over $B = 16$ blocks on those two distinct noisy data sets to obtain the results of Figure 5.1 for the linear case and Figure 5.2 for the parabolic test profile. The increasing spread in the extent of the confidence intervals with increasing measurement error demonstrated in Figures 5.1 and 5.2 shows that, while error in the measurement of the mean x-displacement on the scale of that described in the PIV literature definitely affects the reconstructed results and confidence intervals, even at 10% relative errors the reconstructed values are promising. Moreover, if experimental uncertainties can be reduced to about 5%, as pursued in the literature [12], [14], then the block-averaging statistical reconstruction here appears to perform essential as well as with perfectly measured displacements, as illustrated in both Figures 5.1(a) and 5.2(a). We remark that the approximation at the last point Z_M is the worst. This might be caused by some numerical artifacts imposed by the artificial upper wall elastic boundary condition, but we have not been able to pinpoint it precisely so far. However, since the goal is to obtain a better approximation of the velocity profile in the very near-wall region, this is not a major drawback.

6. Discussion. We have successfully demonstrated that it is possible to use the correlation between unknown velocity profiles $\gamma(aZ)$ and known wall-hindered diffusion coefficient $D(z)$ to reconstruct the velocity values with reasonable precision at multiple collocation points within the depth of an imaged window, reconstructing the out-of-plane z-dependence using only measured in-plane displacements, with examples from three basic test flows (linear, parabolic, and exponentially decaying).

We emphasize that all previously reported experimental values, except the recently developed multilayer nPIV (see [9]), obtain a single value for the velocity over the entire region of observation, namely the mean located in the center of the field of focus. The present reconstruction algorithm, approximating the behavior of the deterministic velocity at M (typically $M = 3, 5, 7$) points scattered over the imaged region, is thus a significant improvement.

The reconstruction uses block averaging, and the error plots have demonstrated that it is better to approximate the profile individually over data set blocks of size 2^{14} and to average the result over 8 or 16 consecutive blocks than processing all of our simulated data at once. Importantly, this is computationally fast: the mimimum of the likelihood function ϕ in (4.7) is found in less than 30 minutes on a desktop machine. In contrast, each Milstein-scheme simulation used to generate data here required on the order of 10 days on the same processor. This reconstruction does not use any information about the intensity function and offers an alternative approach to the recently developed multilayer PIV techniques [6], [9] which attempt to infer distance from the wall from image intensities. An interesting direction for future development is the possibility of combining the imperfect (from polydispersity) out-of-plane intensity information with the present statistical method.

The amount of data used in the reconstruction process may seem staggering, but a comparison with data actually captured in experiments indicates that such data sets can be achieved in a reasonable time. For example, Guasto, Huang, and Breuer [4] track over 140000 single quantum dots from 900 image pairs to obtain a single approximation over the entire region. Li, Sadr, and Yoda [9] cite a framing rate of about 26Hz leading to a sequence of 100 frames of about 30 particles recorded within 5 seconds. Keeping the same interframe ratio, it will take between 20 minutes and 4 hours to obtain the necessary 2^{18} frames. Moreover, Li et al. report using in their computer simulated multilayer nPIV 2000 frames with 120 particles and 3 windows for each one of their three layers. In other words they use, after having thrown away an unquantified amount of bad data, about 2^{14} mean displacements for 2^7 particles.

The present demonstration assumes that the particles are uniformly distributed between two measurements; once the computational and observation domain are no longer the same, the uniform distribution assumption will be broken due to particle drop-in and drop-out from the window between two measurements. Provided that this distribution can be computed a priori [13], the reconstruction is simply modified to include the nonuniform probability density function of matched particles. The present results are, of course, only a computer-simulated proof of concept, and more physical effects need to be included for proper use on experimental data, perhaps including the effects of background noise in the images, particle polydispersity, and particles dropping in and out of the field of vision.

Acknowledgments. The authors would like to thank Minami Yoda, Haifeng Li, and Reza Sadr in the Woodruff School of Mechanical Engineering at the Georgia Institute of Technology for providing the experimental problem and for fruitful discussions.

REFERENCES

[1] M. A. Bevan and D. C. Prieve, *Hindered diffusion of colloidal particles very near to a wall: Revisited*, J. Chem. Phys., 113 (2000), pp. 1228–1236.

[2] D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interaction*, J. Chem. Phys., 69 (1978), pp. 1352–1360.

[3] G. G. Fuller, J. M. Rallison, R. L. Schmidt, and L. G. Leal, *The measurement of velocity gradients in laminar flow by homodyne light-scattering spectroscopy*, J. Fluid Mech., 100 (1980), pp. 555–575.

[4] J. S. Guasto, P. Huang, and K. S. Breuer, *Statistical particle tracking velocimetry using molecular and quantum dot tracer particles*, Exp. Fluids, 41 (2006), pp. 869–880.

[5] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Rev., 43 (2001), pp. 525–546.

[6] P. Huang, J. S. Guasto, and K. S. Breuer, *Direct measurement of slip velocities using three-dimensional total internal reflection velocimetry*, J. Fluid Mech., 566 (2006), pp. 447–464.

[7] S. Jin, P. Huang, J. Park, J. Y. Yoo, and K. S. Breuer, *Near-surface velocimetry using evanescent wave illumination*, Exp. Fluids, 37 (2004), pp. 825–833.

[8] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1999.

[9] H. Li, R. Sadr, and M. Yoda, *Multilayer nano-particle image velocimetry*, Exp. Fluids, 41 (2006), pp 185–194.

[10] D. F. Liang, C. B. Jiang, and Y. L. Li, *A combination correlation-based interrogation and tracking algorithm for digital PIV evaluation*, Exp. Fluids, 33 (2002), pp. 684–695.

[11] C. D. Meinhart, S. T. Wereley, and J. G. Santiago, *PIV measurements of a microchannel flow*, Exp. Fluids, 27 (1999), pp. 414–419.

[12] R. Sadr, H. Li, and M. Yoda, *Impact of hindered Brownian diffusion on the accuracy of nano-particle image velocimetry data*, Exp. Fluids, 38 (2004), pp. 90–98.

[13] R. Sadr, H. Li, and M. Yoda, *Bias due to hindered Brownian diffusion in near-wall velocimetry*, in Proceedings of the 6th International Symposium on Particle Image Velocimetry, 2005.

[14] R. Sadr, M. Yoda, Z. Zheng, and A. T. Conlisk, *An experimental study of electro-osmotic flow in rectangular microchannels*, J. Fluid Mech., 506 (2004), pp. 357–367.

[15] S. T. Wereley and C. D. Meinhart, *Micron-resolution particle image velocimetry*, in Diagnostic Techniques for Microfluidics, K. Breuer, ed., Springer-Verlag, Berlin, 2005, pp. 51–110.

[16] C. Zettner and M. Yoda, *Particle velocity field measurements in a near-wall flow using evanescent wave illumination*, Exp. Fluids, 34 (2003), pp. 115–121.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 253–270

COUPLED FORWARD-ADJOINT MONTE CARLO SIMULATIONS OF RADIATIVE TRANSPORT FOR THE STUDY OF OPTICAL PROBE DESIGN IN HETEROGENEOUS TISSUES[*]

CAROLE K. HAYAKAWA[†], JEROME SPANIER[‡], AND VASAN VENUGOPALAN[†]

Abstract. We introduce a novel Monte Carlo method for the analysis of optical probe design that couples a forward and an adjoint simulation to produce spatial-angular maps of the detected light field within the tissue under investigation. Our technique utilizes a generalized reciprocity theory for radiative transport and is often more efficient than using either forward or adjoint simulations alone. For a given probe configuration, the technique produces rigorous, transport-based estimates of the joint probability that photons will both visit any specified target subvolume and be detected. This approach enables the entire tissue region to be subdivided into a collection of target subvolumes to provide a phase-space map of joint probabilities. Such maps are generated efficiently using only one forward and one adjoint simulation for a given probe configuration. These maps are used to identify those probe configurations that best interrogate targeted subvolumes. Inverse solutions in a layered tissue model serve to illustrate and reinforce our analysis.

Key words. Monte Carlo methods, radiative transfer, inverse problems, biological applications

AMS subject classifications. 65C05, 85A25, 34A55, 78A70, 92C55

DOI. 10.1137/060653111

1. Introduction.

The use of light for noninvasive, in vivo determination of optical and physiological properties of tissue volumes is established for a host of applications in biomedical optics. In some cases, other imaging modalities, such as x-ray, ultrasound, or MRI, are used in conjunction with optical techniques to identify heterogeneous tissue regions that require further analysis. Knowledge of this structural information can provide information critical to the design of optical probes to target these regions effectively or to provide information regarding both the target region and its surroundings.

With these goals in mind, much effort has been expended in improving the design of optical probes. For example, there have been attempts to enhance the light delivered to specific tissue regions by varying source and detector characteristics such as orientation, size, angle of emission (for sources), angle of acceptance (for detectors), source-detector (s-d) separation, and/or distance between the target volume and the source/detector [6, 8, 14, 16, 21, 22, 31]. These optical probes are configured in an attempt to enhance the light that is both delivered to the targeted volume and subsequently detected at the tissue surface. Clearly, detailed knowledge of the spatial-angular distribution of the detected light field for a given probe configuration would serve to assess the effectiveness of these approaches and provide a basis to compare competing probe designs.

[*] Received by the editors February 23, 2006; accepted for publication (in revised form) July 3, 2007; published electronically November 7, 2007. This research was supported by National Institutes of Health grants P41-RR-00192 (Laser Microbeam and Medical Program) and R01-EB-00345.
http://www.siam.org/journals/siap/68-1/65311.html

[†] Laser Microbeam and Medical Program, Beckman Laser Institute, University of California—Irvine, Irvine, CA 92612-3010, and Department of Chemical Engineering and Materials Science, University of California—Irvine, Irvine, CA 92697-2575 (hayakawa@uci.edu).

[‡] Laser Microbeam and Medical Program, Beckman Laser Institute, University of California—Irvine, Irvine, CA 92612-3010.

Previous studies of radiative transfer in tissue from source to detector have been based mainly on the diffusion approximation to the radiative transport equation [3, 7, 15, 17]. However, the validity of diffusion-based models is compromised when (a) s-d separations are small or (b) the tissue absorption is comparable to or greater than scattering. While analytic [27] and specific Monte Carlo [2] approaches have been investigated, it is unclear how these methodologies would extend to heterogeneous media. Moreover, conventional Monte Carlo simulations provide results with large uncertainties in the detected signals due to the small detector sizes often used in optical probes.

To address this problem in the context of radiative transport, we have developed a novel Monte Carlo method that produces phase-space maps to provide quantitative measures of the ability for a given probe configuration to detect light delivered to specific regions within the tissue. This general approach can be applied to complex, heterogeneous media. The method makes use of coupled forward-adjoint simulations to estimate the joint probability of both visitation of a target region and subsequent detection at the tissue surface. Bayes's theorem [12] is used to decompose this joint probability into the product of an absolute and a conditional probability. These two probabilities are then estimated using separate and efficient simulations. In cases for which the targeted volume is large compared to both source and detector volumes, the gains in efficiency over the use of either a forward or an adjoint simulation alone can be substantial.

In this paper, we describe the foundations of our method as well as its operational details. We then apply the method to investigate how a layered epithelial tissue is interrogated by optical probe designs in which we allow variation in s-d separation. Forward and adjoint simulations are generated for various probe configurations. The simulation results are used to produce maps that provide both qualitative and quantitative information regarding the phase-space distribution of the detected light. This information provides a basis for the comparison of prospective probe designs to determine the merits of each. Accurate recovery of optical properties from heterogeneous tissues via inverse solutions serves to confirm the comparative analysis of candidate probe designs as evaluated by the coupled forward-adjoint Monte Carlo simulations.

2. Method. To determine the probability of detecting light that has visited a targeted volume, one could utilize a conventional Monte Carlo simulation in which one follows photon trajectories from the source to a target volume and then tallies the final photon weight for those photons that are subsequently detected. Alternatively, one could use an adjoint Monte Carlo simulation, in which one follows backward-propagating photons from the detector to the target volume, and then to the source. However, when the source and detector are each small relative to the target volume, sole use of a forward or adjoint simulation engenders low signal-to-noise ratios (SNRs). Such a situation is exceedingly common in biomedical optics.

Our approach is to break the problem into two components and determine separately (a) the probability of source to target trajectories, $P(V)$ ("target visitation"), and (b) the probability of detection conditioned by target visitation, $P(D|V)$ ("detection given target visitation"). The combination of these two probabilities using Bayes's theorem provides the rigorous joint transport probability of "target visitation and detection":

$$P(V \cap D) = P(V) \cdot P(D|V). \tag{2.1}$$

We use a conventional Monte Carlo simulation to determine $P(V)$. However, for

$P(D|V)$, we utilize an adjoint simulation to combat the inherently low SNR in its estimation in the reverse direction. This is done by modifying a generalized reciprocity principle [5, 19, 25, 29] to convert $P(D|V)$ to a coupled forward-adjoint computation at the surface of the target volume.

In the next section, we describe our application of coupled forward-adjoint Monte Carlo methods for the determination of $P(V \cap D)$. This includes a review of classical reciprocity theory and basic equations. We then describe generalized reciprocity, for which classical reciprocity is a special case. Finally, we develop our extension of generalized reciprocity theory to arrive at an estimate of $P(V \cap D)$. This will set the stage for the application of this methodology to problems in biomedical optics.

3. Coupled forward-adjoint Monte Carlo methods.

A series of publications [5, 19, 20, 25, 24, 30, 29] has developed and described the "midway" forward-adjoint coupling method to increase the efficiency of estimating detector responses in radiative transport problems. The idea is to simulate both forward and adjoint Monte Carlo transport and combine the tallies from each at an intermediate surface to estimate the total system response. The midway method is made rigorous by appealing to a generalized reciprocity theory for transport equations [19, 20, 29, 30]. The midway method has been shown to be particularly efficient in problems that involve deep penetration and/or complex streaming pathways taken by the radiation as it moves from source(s) to detector(s) [19, 20, 24, 25].

We modify the midway method in order to apply it to the estimation of the conditional probability $P(D|V)$. Photons are launched at a physical source and are propagated until they exit the phase space. At each interaction within the tissue, the photon weight is reduced according to its survival probability, a technique sometimes referred to as "absorption weighting" [23]. Only photon trajectories that have intersected the target volume $\mathbb{V}$ contribute to the estimate of $P(V)$. These "visiting" photons generate an induced source internal to $\mathbb{V}$ that produces a surface source on $\partial\mathbb{V}$. This surface source is then paired with the adjoint flux on $\partial\mathbb{V}$ in a bilinear integration that produces an estimate of $P(D|V)$. The product of the two probabilities $P(V)$ and $P(D|V)$ then provides the probability that photons will both visit and subsequently be detected from subvolumes within the phase space. We use this product to provide quantitative information to assess the characteristics of potential probe designs.

3.1. Classical reciprocity.

We begin with the integro-differential form of the radiative transport equation (RTE) assumed to hold in the interior of a closed, bounded subset $\mathbb{D}$ of $\mathbb{R}^3$:

$$\nabla \cdot \mathbf{\Omega}\, \Phi(\mathbf{r}, \mathbf{\Omega}) + \mu_t(\mathbf{r})\Phi(\mathbf{r}, \mathbf{\Omega}) = \mu_s(\mathbf{r}) \int_{4\pi} f(\mathbf{\Omega'} \to \mathbf{\Omega})\Phi(\mathbf{r}, \mathbf{\Omega'})\, d\mathbf{\Omega'} + Q(\mathbf{r}, \mathbf{\Omega}), \tag{3.1}$$

where $\Phi(\mathbf{r}, \mathbf{\Omega})$ is the photon flux, $\mu_t(\mathbf{r}) = \mu_s(\mathbf{r}) + \mu_a(\mathbf{r})$ is the total attenuation coefficient, $\mu_s(\mathbf{r})$ is the scattering coefficient, $\mu_a(\mathbf{r})$ is the absorption coefficient, $f(\mathbf{\Omega'} \to \mathbf{\Omega})$ is the single scattering phase function, and $Q(\mathbf{r}, \mathbf{\Omega})$ is an internal (volumetric) source function, with $\mathbf{r} = (x, y, z)$ and $\mathbf{\Omega} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ representing position and unit direction vectors, respectively. A unique solution $\Phi(\mathbf{r}, \mathbf{\Omega})$ is assured for all $\mathbf{r} \in \mathbb{D}$, $\mathbf{\Omega} \in S^2$ by specifying the photon flux $\Phi_{\text{inc}}(\mathbf{r}, \mathbf{\Omega})$ incident on $\partial\mathbb{D}$ from outside the tissue. We introduce an abbreviated form of (3.1):

$$\nabla \cdot \mathbf{\Omega}\Phi + \mathcal{B}\Phi = Q, \tag{3.2}$$

where $\mathcal{B}$ denotes the transport operator less the divergence term.

A typical optical probe is an instrument that both introduces light at the tissue boundary and collects light reemitted from the boundary using or more detectors positioned at fixed distances from the source. Assuming that there are no other (external) sources of light, the unique solution Φ of this RTE can be written as a superposition of the photon fluxes produced by the internal source Q and the boundary source Q_s defined by

$$Q_s(\mathbf{r}_s, \mathbf{\Omega}) = -\mathbf{\Omega} \cdot \mathbf{n}_s \Phi_{\text{inc}}(\mathbf{r}_s, \mathbf{\Omega}) \quad \text{for } \mathbf{r}_s \in \partial\mathbb{D},\ \mathbf{\Omega} \cdot \mathbf{n}_s < 0, \tag{3.3}$$

where $\mathbf{n}_s$ is the outward-pointing unit normal at r_s:

$$\begin{aligned} \Phi(\mathbf{r}, \mathbf{\Omega}) = &\int_{\mathbb{D}\times S^2} G[(\mathbf{r}_0, \mathbf{\Omega}_0) \to (\mathbf{r}, \mathbf{\Omega})] Q(\mathbf{r}_0, \mathbf{\Omega}_0)\, d\mathbf{r}_0\, d\mathbf{\Omega}_0 \\ &+ \int_{\partial\mathbb{D}\times S^2} G_s[(\mathbf{r}_0, \mathbf{\Omega}_0) \to (\mathbf{r}, \mathbf{\Omega})] \Phi_{\text{inc}}(\mathbf{r}_s, \mathbf{\Omega}_s)\, d\mathbf{r}_0\, d\mathbf{\Omega}_0, \end{aligned} \tag{3.4}$$

where G is the volume Green's function and G_s is the surface Green's function for the problem. An alternate, equivalent representation that uses only the volume Green's function is

$$\Phi(r, \Omega) = \int_{\mathbb{D}\times S^2} G[(\mathbf{r}_0, \mathbf{\Omega}_0) \to (\mathbf{r}, \mathbf{\Omega})] Q(\mathbf{r}_0, \mathbf{\Omega}_0)\, d\mathbf{r}_0\, d\mathbf{\Omega}_0, \tag{3.5}$$

where the second term in (3.4) is replaced by the boundary condition

$$\Phi(\mathbf{r}_s, \mathbf{\Omega}_s) = Q_s(\mathbf{r}_s, \mathbf{\Omega}_s) \tag{3.6}$$

and Q_s is defined as in (3.3) [4]. The relationship (3.3) and the equivalence between the representations (3.4) and (3.5) together with (3.6) will be utilized in section 3.3.

The response of either a virtual or a physical detector can then be described in terms of a linear functional of Φ:

$$I = \int_{\mathbb{D}\times S^2} Q^* \Phi\, d\mathbf{r}\, d\mathbf{\Omega}, \tag{3.7}$$

where Q^* characterizes the detector position, size, and acceptance angle. Both the source function Q and detector function Q^* may be described mathematically using characteristic functions associated with the source and detector. For example, if the tissue $\mathbb{D}$ is assumed to occupy the half space characterized in rectangular coordinates by $z > 0$, and a fiber-optic laser source of radius q and unit strength is normally incident at $(0, 0, 0)$, we have

$$Q_s(x_s, y_s, 0, \mathbf{\Omega}) = \begin{cases} 1, & \begin{array}{l} x_s^2 + y_s^2 \le q^2 \quad \text{and} \\ -1 \le \mathbf{n}_{\mathbb{D}} \cdot \mathbf{\Omega} < -\cos\theta_Q, \end{array} \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

where $\mathbf{n}_{\mathbb{D}}$ denotes the outward-pointing unit normal at the bounding surface $z = 0$ and the source is confined to an emission angle θ_Q. A similar description characterizes a typical fiber-optic detector placed elsewhere on the tissue surface $z = 0$, except that it collects light that scatters *into* the half space $z < 0$ at location $(X_{Q^*}, Y_{Q^*}, 0)$ within the detector radius q^* and acceptance angle θ_{Q^*}. Specifically,

$$Q_s^*(x_s, y_s, 0, \mathbf{\Omega}) = \begin{cases} 1, & \begin{array}{l} (x_s - X_{Q^*})^2 + (y_s - Y_{Q^*})^2 \le (q^*)^2 \quad \text{and} \\ 1 \ge \mathbf{n}_{\mathbb{D}} \cdot \mathbf{\Omega} > \cos\theta_{Q^*}, \end{array} \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

It is well known that classical reciprocity theory also permits the detector response I to be expressed as a linear functional of the solution to the RTE that is adjoint to (3.1) [23]:

(3.10)
$$-\nabla \cdot \mathbf{\Omega}\, \Phi^*(\mathbf{r}, \mathbf{\Omega}) + \mu_t(\mathbf{r})\Phi^*(\mathbf{r}, \mathbf{\Omega}) = \mu_s(\mathbf{r}) \int_{4\pi} f(\mathbf{\Omega} \to \mathbf{\Omega}')\Phi^*(\mathbf{r}, \mathbf{\Omega}')\, d\mathbf{\Omega}' + Q^*(\mathbf{r}, \mathbf{\Omega})$$

and

$$I = \int_{\mathbb{D}\times S^2} Q\Phi^*\, d\mathbf{r}\, d\mathbf{\Omega}. \tag{3.11}$$

Upon comparing the detector response representations (3.7) and (3.11), we notice that the roles of the source function Q and detector function Q^* are interchanged in this statement of reciprocity, so that Q acts as a "detector" function for the adjoint formulation and Q^* plays the role of a "source" function for the adjoint equation.

Using operator notation, (3.10) can be written as

$$-\nabla \cdot \mathbf{\Omega}\Phi^* + \mathcal{B}^*\Phi^* = Q^*, \tag{3.12}$$

where $\mathcal{B}^*$ is the operator adjoint to $\mathcal{B}$. For (3.11) to be valid, it is also understood that the boundary condition satisfied by Φ^* on $\partial\mathbb{D}$ is dual to that specified for Φ. For example, in our application $\partial\mathbb{D}$ is the surface of the tissue that is composed of the source region A_Q, the detector region A_{Q^*}, and the complement of these two regions $\partial\mathbb{D}\backslash(A_Q \cup A_{Q^*})$. Here we assume for simplicity that both the source emission angle and the detector acceptance angle are fully open; i.e., $\cos\theta_Q = \cos\theta_{Q^*} = 0$. For this case, the boundary condition at $z = 0$ satisfied by Φ is

$$\Phi(x_s, y_s, 0, \mathbf{\Omega}) = Q_s(x_s, y_s, 0, \mathbf{\Omega}) \quad \text{for } (x_s, y_s) \in A_Q, \tag{3.13}$$

where the right-hand side is defined by (3.8) with $\cos\theta_Q = 0$. The dual boundary condition for Φ^* becomes

$$\Phi^*(x_s, y_s, 0, \mathbf{\Omega}) = Q_s^*(x_s, y_s, 0, \mathbf{\Omega}) \quad \text{for } (x_s, y_s) \in A_{Q^*}, \tag{3.14}$$

where the right-hand side is defined by (3.9) with $\cos\theta_{Q^*} = 0$. From (3.13) and (3.14) we have $\Phi\Phi^* = 0$ on $\partial\mathbb{D}$ establishing that the boundary conditions are dual to each other. Note that (3.13) and (3.14) are incomplete statements of the boundary conditions for $\mathbf{r} \in \partial\mathbb{D}\backslash(A_Q \cup A_{Q^*})$, as they do not include the full range of $\mathbf{\Omega}$. The missing conditions for our case accommodates a tissue-air refractive index mismatch using the Fresnel relations for unpolarized light [28]. This results in a mixed boundary condition comprised of a linear combination of reflecting and nonreentrant conditions, each component of which leads to duality as shown by Aronson [1].

The duality of the governing equations and boundary conditions enables a detector response to be computed *either* in the context of forward Monte Carlo sampling *or* adjoint Monte Carlo sampling via this "classical" reciprocity for the RTE. Usually one of these formulations will lead to a more efficient simulation than the other. However, many problems in biomedical optics utilize *both* small sources *and* small detectors, making *neither* formulation efficient.

However, we can improve the efficiency of classical reciprocity by utilizing a midway surface between source and detector. The midway method combines forward and adjoint sampling that characterizes those photons that have migrated from source to detector through a separating midway surface. Application of generalized reciprocity to the estimation of $P(D|V)$ using this midway method will then lead us to our final evaluation of $P(V \cap D)$. In section 3.2 we explain generalized reciprocity and the midway method. In section 3.3 we extend these ideas to compute the joint probability of visitation and detection $P(V \cap D)$.

3.2. Generalized reciprocity. Let us first consider $\mathbb{V}_M$ to be an arbitrary closed, bounded subset of $\mathbb{D}$ and $\partial\mathbb{V}_M$ its surface. Multiplying (3.2) by Φ^* and (3.12) by Φ, subtracting the latter product from the former, and integrating the difference over all locations and directions within V_M, we get

$$\int_{\mathbb{V}_M \times S^2} \nabla \cdot \mathbf{\Omega}\Phi\Phi^* \, d\mathbf{r} \, d\mathbf{\Omega} = \int_{\mathbb{V}_M \times S^2} [Q\Phi^* - Q^*\Phi] \, d\mathbf{r} \, d\mathbf{\Omega}. \tag{3.15}$$

Use of Green's theorem to replace the volume integral on the left-hand side of (3.15) by a surface integral leads to

$$\int_{\partial\mathbb{V}_M \times S^2} \mathbf{n}_M \cdot \mathbf{\Omega}\Phi\Phi^* \, d\mathbf{r} \, d\mathbf{\Omega} = \int_{\mathbb{V}_M \times S^2} [Q\Phi^* - Q^*\Phi] \, d\mathbf{r} \, d\mathbf{\Omega}, \tag{3.16}$$

where $\mathbf{n}_M$ is the outward-pointing unit vector normal to $\partial\mathbb{V}_M$. Equation (3.16) is often referred to as the global reciprocity theorem [30]. Note that if $\mathbb{V}_M = \mathbb{D}$ and the boundary conditions at the air-tissue interface cause the integral on the left-hand side to vanish (as is the case in our problem), we then arrive at the "classical" statement of reciprocity:

$$\int_{\mathbb{V}_M \times S^2} [Q\Phi^* - Q^*\Phi] \, d\mathbf{r} \, d\mathbf{\Omega} = 0. \tag{3.17}$$

While (3.16) is valid generally, it becomes particularly useful when $\mathbb{V}_M$ encloses *either* the source *or* the detector region. The surface of $\mathbb{V}_M$, $\partial\mathbb{V}_M$, can then be identified as a "midway" surface between source and detector. In this case, every photon that is detected from the source *must* intersect the midway surface.

The function $\Phi\Phi^*$ that occurs in (3.16) has been called a "contributon" response function [5, 19, 20, 24, 25, 29, 30] and used to define a unit of information that characterizes transport from source to detector. The integral of this function appearing on the left-hand side of (3.16) plays a similar role here. It captures the flow of information across the boundary of the midway volume $\mathbb{V}_M$.

If $\mathbb{V}_M$ encloses the source region as shown in Figure 3.1(a) and $Q^* = 0$ in $\mathbb{V}_M$, the left-hand side of (3.16) is positive and equals $\int_{\mathbb{V}_M \times S^2} Q\Phi^*$, which is the adjoint representation of the detector response. If $\mathbb{V}_M$ encloses the detector region, and $Q = 0$ in $\mathbb{V}_M$, the left-hand side of (3.16) is negative and equals $-\int_{\mathbb{V}_M \times S^2} Q^*\Phi$, which is the forward representation of the detector response. Reversing the sense of $\mathbf{n}_M$ by replacing the outward-pointing unit normal with the inward-pointing unit normal changes the sign in the surface integral on the left-hand side of (3.16) and also reverses the sense of enclosure. That is, if $\mathbb{V}_M$ is treated as an enclosure for the source, then $\mathbf{n}_M$ points outward. However, if the complement of $\mathbb{V}_M$ in $\mathbb{D}$, $\mathbb{D}\backslash\mathbb{V}_M$, is treated as an enclosure for the detector, then $\mathbf{n}_M$ points inward.

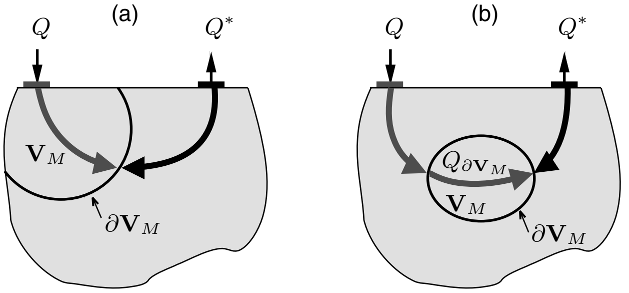


FIG. 3.1. *Geometry of* (a) *generalized reciprocity with the midway surface* $\partial\mathbf{V}_M$ *enclosing the physical source, and* (b) *an extension to generalized reciprocity with an induced boundary source* $Q_{\partial\mathbf{V}_M}$ *out of* $\mathbf{V}_M$ *used to compute conditional probabilities.*

In many situations, this midway surface method proves to be more efficient than using either the forward or adjoint simulation alone [19, 20]. The gains in efficiency will largely be the consequence of the relatively larger "size" of $\mathbb{V}_M$ compared with either the size of the source region, where $Q \neq 0$, or of the detector region, where $Q^* \neq 0$. Here "size" is to be interpreted in a probability sense as opposed to a strict physical size. More precisely, the relevant condition is that the probability of reaching $\mathbb{V}$ from either the source or detector should be larger than the corresponding probabilities in the reverse directions.

3.3. Probability of visitation and detection. The probe design problem requires an understanding of more than the *total* system response at the detector from the original optical source. It requires, in addition, knowledge of the detector response due *only* to those photons that have visited a targeted tissue region $\mathbb{V}$. For our application, $\mathbb{V}$ is a region that encloses *neither* the source *nor* the detector. Thus, to make use of generalized reciprocity, we treat $\mathbb{V}$ as a region that generates a secondary or "virtual" source, induced by the original physical source. This construct allows us to decompose the problem into two subproblems. The first problem deals only with the estimation of $P(V)$ and is handled using conventional Monte Carlo simulation. The second problem handles the estimation of $P(D|V)$ and will be accomplished by a suitable application of the generalized reciprocity relation (3.16).

Accordingly, the midway volume $\mathbb{V}_M$ can be considered as an arbitrary volume enclosing neither source nor detector with surface $\partial\mathbb{V}_M$. A possible geometry is shown in Figure 3.1(b). This arbitrary volume $\mathbb{V}_M$ is one at whose boundary, $\partial\mathbb{V}_M$, information will be collected from both forward and adjoint photons for the estimation of $P(V)$ and $P(D|V)$, respectively. We estimate the first factor of (2.1), $P(V)$, by launching photons from the original source Q characterized in (3.8) and, for those photons that enter $\mathbb{V}_M$, tally the entering weight of each photon. These photons produce estimates of $P(V)$ and generate samples drawn from an induced source $Q_{\partial\mathbb{V}_M}$ impinging on $\partial\mathbb{V}_M$ from *inside* $\mathbb{V}_M$. The boundary surface $\partial\mathbb{V}_M$ then defines the surface for the midway method applied to the problem of estimating the conditional probability, $P(V|D)$.

The details of the required computation deserve elaboration. Denote $\Phi(\mathbf{r}, \mathbf{\Omega})$ as the solution of the boundary-value problem of (3.1) with source described as in (3.8), and let $\Phi_{\mathbb{V}_M}(\mathbf{r}, \mathbf{\Omega})$ denote the restriction of $\Phi(\mathbf{r}, \mathbf{\Omega})$ to $\mathbf{r} \in \mathbb{V}_M$. The photon flux $\Phi_{\partial\mathbb{V}_M}(\mathbf{r}, \mathbf{\Omega})$ for $\mathbf{r} \in \partial\mathbb{V}_M$, $\mathbf{\Omega} \cdot \mathbf{n}_{\partial\mathbb{V}} < 0$, where $\mathbf{n}_{\partial\mathbb{V}}$ = unit normal out of $\mathbb{V}_M$ (*into* $\mathbb{D}\backslash\mathbb{V}$) then generates a boundary source $Q_{\partial\mathbb{V}_M}(\mathbf{r}_s, \mathbf{\Omega}_s) = -\Omega_s \cdot \mathbf{n}_{\partial\mathbb{V}_M} \Phi_{\partial\mathbb{V}_M}(\mathbf{r}_s, \mathbf{\Omega}_s)$ on

$\partial\mathbb{V}_M$. If we merely replace the source function Q by the source function $Q_{\partial\mathbb{V}}(\mathbf{r},\boldsymbol{\Omega})$ and repeat the derivation that led to (3.16), we obtain

$$\int_{\partial\mathbb{V}_M\times S^2}\mathbf{n}_{\partial\mathbb{V}_M}\cdot\boldsymbol{\Omega}\Phi_{\mathbb{V}_M}\Phi^*\,d\mathbf{r}\,d\boldsymbol{\Omega}=\int_{\mathbb{V}_M\times S^2}\left[Q_{\partial\mathbb{V}_M}\Phi^*-Q^*\Phi_{\mathbb{V}_M}\right]d\mathbf{r}\,d\boldsymbol{\Omega}. \tag{3.18}$$

We replace $Q^*(\mathbf{r},\boldsymbol{\Omega})$ by $Q^*(\mathbf{r},-\boldsymbol{\Omega})$ to generate an adjoint flux, $\Phi^*(\mathbf{r},-\boldsymbol{\Omega})$, *inside* the tissue. This, of course, reverses the direction in the arguments of Q^* and Φ^* in (3.18), which then reads

$$\begin{aligned}&\int_{\partial\mathbb{V}_M\times S^2}\mathbf{n}_{\partial\mathbb{V}_M}\cdot\boldsymbol{\Omega}\Phi_{\mathbb{V}_M}(\mathbf{r},\boldsymbol{\Omega})\Phi^*(\mathbf{r},-\boldsymbol{\Omega})\,d\mathbf{r}\,d\boldsymbol{\Omega}\\&\qquad=\int_{\mathbb{V}_M\times S^2}\left[Q_{\partial\mathbb{V}_M}(\mathbf{r},\boldsymbol{\Omega})\Phi^*(\mathbf{r},-\boldsymbol{\Omega})-Q^*(\mathbf{r},-\boldsymbol{\Omega})\Phi_{\mathbb{V}_M}(\mathbf{r},\boldsymbol{\Omega})\right]d\mathbf{r}\,d\boldsymbol{\Omega}\\&\qquad=\int_{\mathbb{V}_M\times S^2}Q_{\partial\mathbb{V}_M}(\mathbf{r},\boldsymbol{\Omega})\Phi^*(\mathbf{r},-\boldsymbol{\Omega})\,d\mathbf{r}\,d\boldsymbol{\Omega}\end{aligned} \tag{3.19}$$

since $Q^*=0$ inside $\mathbb{V}_M$. Estimation of (3.19) is performed using an adjoint simulation and provides the detected response due to the induced source $Q_{\partial\mathbb{V}_M}$, or $P(D|V)$.

The forward simulation of photons exiting an arbitrary target volume $\mathbb{V}_M$ is used to determine $P(V)$ and is matched with the adjoint simulation estimate of $P(D|V)$ at $\partial\mathbb{V}_M$. The joint probability of visitation and detection $P(V\cap D)$ (see (2.1)) is formed by the product of these two factors. The resulting probability characterizes a three body system involving radiative transport from (a) the original source Q to (b) the target volume $\mathbb{V}_M$ and finally to (c) the detector. In what follows, we shall refer to joint probability of visitation and detection of the target volume, $P(V\cap D)$, as *interrogation* of the target volume.

4. Implementation.

Monte Carlo simulations of both the forward RTE equation (3.1) and the adjoint RTE (3.10) are quite conventional [10, 26]. Photon and adjoint photon biographies are generated by alternately sampling from exponential distributions representing intercollision distances and angular deflections sampled from the Henyey–Greenstein phase function [11, 28]. The resulting random walks are followed until they escape the tissue phase space.

We utilize our coupled forward-adjoint methodology to create quantitative maps of the entire tissue that illustrate how the light interrogates various regions in the tissue. We shall refer to these maps as "interrogation maps." To create such maps, the tissue is subdivided into a finite number of voxels, each treated as a target volume $\mathbb{V}$. The matching of photon trajectories between the forward simulation and the adjoint simulation occurs at the boundary of each voxel. The integration shown on the left-hand side of (3.19) requires the pairing of estimates of the photon current $J=\mathbf{n}_{\partial\mathbb{V}_M}\cdot\boldsymbol{\Omega}\tilde{\Phi}$ from the forward simulation with the estimation of the photon flux Φ^* from the adjoint simulation. Upon exiting a voxel $\mathbb{V}$, both the location and orientation of the photon's track are assigned to one of $N_{\partial\mathbb{V}}\cdot N_\mu\cdot N_\phi$ spatial-angular bins:

$$\Delta_{ijk}:\begin{cases}\mathbf{r}\in\partial\mathbb{V}_i, & i=1,\ldots,N_{\partial\mathbb{V}},\\ \frac{2(j-1)}{N_\mu}<(\mu+1)\le\frac{2j}{N_\mu}, & j=1,\ldots,N_\mu,\\ \frac{2\pi(k-1)}{N_\phi}<\phi\le\frac{2\pi k}{N_\phi}, & k=1,\ldots,N_\phi.\end{cases} \tag{4.1}$$

Each solid angle bin is determined by $\mu=\cos\theta$ and ϕ, where θ is the polar angle and ϕ is the azimuthal angle. The north pole for the directional system is taken to be the outward-pointing normal on each voxel side. Trajectories in the forward simulation that *exit* bin Δ_{ijk} are matched with trajectories in the adjoint simulation that *enter* the same angular bin. We determine $P(V \cap D)$ for each voxel $\mathbb{V}$ by summing the product of the tallies of the forward and adjoint photons in the matched spatial-angular pairs over all bins:

$$P(V \cap D) = \sum_{i}^{N_{\partial \mathbb{V}}} \sum_{j}^{N_\mu} \sum_{k}^{N_\phi} J_{ijk} \Phi^*_{ijk} \Delta \partial \mathbb{V}_i \Delta \mu_j \Delta \phi_k. \tag{4.2}$$

In (4.2), J_{ijk} is estimated by tallying the photon weight w_{ijk} per unit area and solid angle upon exiting the voxel surface

$$J_{ijk} = \frac{1}{N_F} \sum_{n=1}^{N_F} \frac{w^{(n)}_{ijk}}{\Delta \partial \mathbb{V}_i \Delta \mu_j \Delta \phi_k}, \tag{4.3}$$

where N_F is the number of photons launched in the forward simulation. The adjoint simulation converts the adjoint current to an adjoint flux via the relation

$$\Phi^*_{ijk} = \frac{J^*_{ijk}}{\mu_j}, \tag{4.4}$$

where μ_j is the polar cosine of the entering photon. The weight w^*_{ijk} of each adjoint photon entering the voxel surface is then used in the estimate of the adjoint current

$$J^*_{ijk} = \frac{1}{N_A} \sum_{n=1}^{N_A} \frac{w^{*(n)}_{ijk}}{\Delta \partial \mathbb{V}_i \Delta \mu_j \Delta \phi_k}, \tag{4.5}$$

where N_A is the total number of adjoint photons launched. For simplicity, we use uniform spatial and angular bins. In practice, however, we anticipate the need to utilize finer binning closer to the tissue surface and at other locations where the distribution of the light field either is highly anisotropic or possesses large spatial gradients. Sufficiently deep in the tissue, where the flux is expected to be nearly isotropic, a coarse uniform angular grid should suffice.

The variance of our $P(V \cap D)$ estimates is derived as specified in the midway method literature [20]. Specifically, the relative variances of the forward current J_{ijk} and the adjoint flux Φ^*_{ijk} are determined by

$$r^2[J_{ijk}] = \frac{\sum_{n=1}^{N_F} \left[w^{(n)}_{ijk} \right]^2}{\left[\sum_{n=1}^{N_F} w^{(n)}_{ijk} \right]^2} - \frac{1}{N_F} \tag{4.6}$$

and

$$r^2[\Phi^*_{ijk}] = \frac{1}{\mu_j^2} \left\{ \frac{\sum_{n=1}^{N_A} \left[w^{*(n)}_{ijk} \right]^2}{\left[\sum_{n=1}^{N_A} w^{*(n)}_{ijk} \right]^2} - \frac{1}{N_A} \right\}, \tag{4.7}$$

respectively. Since the quantities of J_{ijk} and Φ^*_{ijk} are estimated from forward and adjoint random walks that are sampled independently, a first-order approximation of the relative variance of their product is provided by the sum of their relative variances

$$r^2[J_{ijk} \Phi^*_{ijk}] \approx r^2[J_{ijk}] + r^2[\Phi^*_{ijk}]. \tag{4.8}$$

The variance of $P(V \cap D)$ is obtained by summing the variances of the products over all bins:

$$\sigma^2[P(V \cap D)] = \sum_i^{N_{\partial\mathbb{V}}} \sum_j^{N_\mu} \sum_k^{N_\phi} r^2[J_{ijk}\Phi^*_{ijk}]J^2_{ijk}\Phi^{*2}_{ijk}. \tag{4.9}$$

Note that (4.9) provides the variance for *each* target volume. For a fixed number of launched (forward and adjoint) photons, an increase in $N_{\partial\mathbb{V}}$, N_μ, and N_ϕ results in an increase in $r^2[J_{ijk}\Phi^*_{ijk}]$ because there are fewer photons per bin and will tend to increase the relative error in the estimate for each bin. While an increase in the number of bins will tend to increase the variance $\sigma^2[P(V \cap D)]$, it will also reduce the discretization error. Optimal choices for $N_{\partial\mathbb{V}}$, N_μ, and N_ϕ vary depending on the precise location of the voxel in the tissue. For the purposes of the results presented in this initial study, we chose uniform spatial and angular binning. An analysis of the optimal binning allocation strategy is beyond the scope of this paper and, in any event, will be highly problem-dependent.

As a means to eliminate the discretization error, we investigated a method proposed by Cramer [5] in which two sets of trajectories are launched at the voxel boundary in *exactly* opposite directions. Each set is then followed until possible detection at the source or detector. However, due to the small size of the fiber-optic source and detector and the large number of target volumes treated in our application, this method was not particularly efficient and was not employed in this study.

Note that while we are summing over all spatial-angular bins in this study, we could easily provide maps containing information for photons entering and exiting at *any* particular set of orientations or locations. This more refined information would enable an evaluation of the impact of angular variations in the light distribution on the conditional system response. Such angular detail will be especially important for voxels in regions in which the light field is highly anisotropic, for example, in the proximity of collimated sources or interfaces of refractive index mismatch.

5. Numerical results. We apply our methodology to a test case depicting epithelial tissue consisting of a thin upper cellular layer ($0 < z < 0.5$ mm) situated above a much thicker structural (stromal) layer ($z > 0.5$ mm). The goal of this study is to assess the effect of probe s-d separation on the interrogation of each layer. We first examine the forward problem; that is, we generate $P(V \cap D)$ spatial-angular maps for normal tissue. We refer to this as our "background" tissue problem. The purpose of these maps is to indicate the effectiveness of a given probe configuration to detect and isolate transformations in each of the layers associated with the formation of precancerous tissue. Simulated data of measured reflectance is then generated that contains information characterizing physiologically relevant changes in one or both layers. This measured data is then used to predict changes in the layered optical properties via an inverse solution that employs a special perturbation and differential Monte Carlo optimization method developed previously [9, 10]. For our particular study in epithelial tissue, we will discuss possible relationships between the information provided by the $P(V \cap D)$ maps and the quality of the inverse solution results.

5.1. Background tissue forward problem. We first consider a homogeneous background tissue with a refractive index $n = 1.4$ and optical properties typical of normal stromal tissue [13] at an optical wavelength of 849 nm: $\mu_a = 0.034$/mm, $\mu_s = 6.11$/mm, $g = 0.9$. Here g is the average cosine of the Henyey–Greenstein single-scattering phase function commonly used for tissue [28]. The probe configurations considered consist of both source and detector oriented normal to the tissue surface with s-d separations ranging from 1–3 mm. The source and detector have emission and acceptance angles $\theta_Q = \theta_{Q^*} = 15.3°$ relative to their central axis and are 200 μm in radius. To efficiently present the $P(V \cap D)$ results, which constitute a three-dimensional data set, we sum the results of (4.2) along the y-axis and project them onto the x-z plane in 0.1 mm × 0.1 mm pixels, each of which represents a different target volume $\mathbb{V}$.

Figure 5.1 displays the interrogation $P(V \cap D)$ maps for the background tissue. In this analysis, we compare probe features consisting of s-d separations of 1, 2, and 3 mm. In these plots, the color of every voxel represents the absolute (unscaled) conditional probability of detection (conditioned by visiting the voxel in question). We shall refer to this quantity as the "conditional system response." To enable the visualization of a greater dynamic range in $P(V \cap D)$, the colors are represented on a log scale with a spectrum ranging from large (red (10^{-8})) to small (blue (10^{-11})) probability. A dashed white line at a $z = 0.5$ mm delineates the interface between the two layers of interest. Note that in each of these maps, we display the conditional system response for each voxel. The database so constructed provides the raw material for the analysis of competing probe configurations.

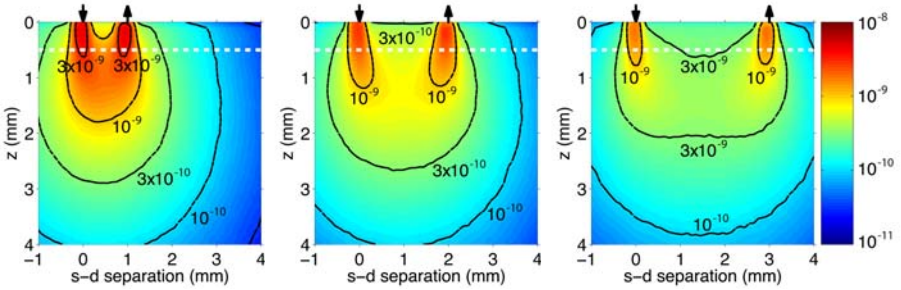


FIG. 5.1. *Interrogation map of the background problem for s-d separations of* 1, 2, *and* 3 *mm (left to right).*

The results of Figure 5.1 can be normalized by the sum of $P(V \cap D)$ over the whole domain $\mathbb{D}$ to produce a true probability density function which we will refer to as an "interrogation density function":

$$p_{V \cap D} = \frac{P(V \cap D)}{\int_{\mathbb{D}} P(V \cap D)}. \tag{5.1}$$

Equation (5.1) provides an appropriate function to assess how a particular region of interest is interrogated. This normalized function allows different probe configurations to be compared on an equivalent basis. For example, we can integrate $p_{V \cap D}$ over the top layer $\mathbb{T}$ or bottom layer $\mathbb{B}$, resulting in $\int_{\mathbb{T}} p_{V \cap D}$ and $\int_{\mathbb{B}} p_{V \cap D}$, respectively. This will provide the relative contribution from each layer to the detected signal in the form of a probability.

Figure 5.2 presents the integration of $p_{V \cap D}$ over the top and bottom layers as a function of s-d separation. These results reveal that roughly four times as much detected signal has interrogated the bottom layer as opposed to the top layer. The bottom layer probabilities increase by 4.5% (from 0.805 to 0.841) as the s-d separation increases from 1 to 3 mm, revealing that the larger s-d separations are more effective in interrogating the bottom layer than the smaller ones. Recall that by "interrogate"

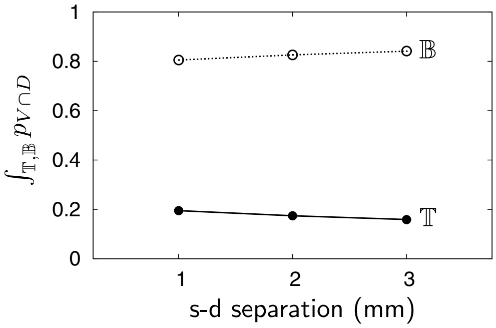


FIG. 5.2. *Interrogation density function* (5.1) *integrated over the top* $\mathbb{T}$ *and bottom* $\mathbb{B}$ *layers as a function of s-d separation.*

we mean that the light both has visited the region and is subsequently detected. This is not surprising, as the detected photons using probes with larger s-d separations will possess longer trajectories that will typically penetrate deeper into the tissue. Consistent with this bottom layer analysis, the top layer probabilities decrease by nearly 20% (from 0.195 to 0.159) as the s-d separation increases from 1 to 3 mm, indicating that the smaller s-d separations provide a considerable improvement in the interrogation of the top layer. This information drawn from our transport model assists in the design of probes for the accurate recovery of optical properties in each layer.

We next solve the inverse problem using similar probe configurations to verify the expected correlation between information given by our $P(V \cap D)$ maps and the ability of specific probe designs to determine optical property changes in one of the layers.

5.2. Perturbed tissue inverse problem. Here we discuss the impact of the interrogation maps on the accurate recovery of optical properties. To utilize information gained from our $P(V \cap D)$ maps, which are generated for various s-d separations, probes with similar features are employed for the recovery of optical properties. However, for the purpose of performing a two-parameter inverse solution, at least two detectors are required. To perform the inversion, we choose to employ six detectors to make the inverse solution more robust with respect to signal noise in the measured reflectance signal. These detectors are 200 μm in diameter and are positioned adjacent to each other to form a linear detector array that spans 1.2 mm. We solve the inverse problem with this detector array immediately adjacent to the source, resulting in s-d separations that span [0–1.2] mm. We also consider the quality of the inverse problem results in seven other configurations by moving the detector array progressively away from the source in increments of 0.2 mm. This provides measurements with eight distinct ranges of s-d separations: [0–1.2] mm, [0.2–1.4] mm, [0.4–1.6] mm, [0.6–1.8] mm, [0.8–2.0] mm, [1.0–2.2] mm, [1.2–2.4] mm, and [1.4–2.6] mm. These configurations of six detectors with a single source provide a spatially resolved measurement of reflected light. Clearly, the progressive movement of the detector array away from the source results in the interrogation of deeper layers, as already seen in Figure 5.2.

We examine two test cases that represent transformations in each layer typical of optical properties changes occurring from the development of precancer: (a) an increase in optical absorption within the lower layer due to the recruitment of increased blood flow and (b) an increase in optical scattering within the upper layer due to local cellular transformations. Simulated spatially resolved reflectance data are generated using two-region Monte Carlo simulations with 2% Gaussian noise added. Our method to determine μ_a and μ_s uses starting values taken from our background (homogeneous) case. Perturbation and differential Monte Carlo methods [9, 10, 18] are used in a two-parameter optimization algorithm to determine the changes to these optical properties, $\hat{\mu}_a = \mu_a + \delta\mu_a$ and $\hat{\mu}_s = \mu_s + \delta\mu_s$, prescribed in one of the layers that best fit the simulated measured data. The solution identifies the layer optical properties that best match the measured data in the least squares sense. Details of the inverse solution method are described elsewhere [9, 10, 18].

5.2.1. Bottom layer $\boldsymbol{\mu_a}$ perturbation. In our first test case we consider a 200% increase to μ_a relative to the background optical properties in the bottom layer. All other optical properties are held fixed. This results in the following set of optical properties: $\mu_s = 6.11$/mm, $\mu_a = 0.034$ in the top layer and $\hat{\mu}_s = 6.11$/mm, $\hat{\mu}_a = 0.068$/mm in the bottom layer. Our two-parameter inverse solution seeks to identify and decouple both $\hat{\mu}_s$ and $\hat{\mu}_a$ successfully.

Figure 5.3 displays the recovered optical properties in the bottom layer as a function of the separation between the source and the linear array of detectors. Error bars representing one standard deviation confidence intervals are shown. The solid and dashed horizontal lines represent the true $\hat{\mu}_s$ and $\hat{\mu}_a$ values in the bottom layer, respectively. The $\hat{\mu}_s$ recovery for all ranges of s-d separation is excellent. The quality of the $\hat{\mu}_a$ estimates improves as the s-d separation increases, as is evidenced by more accurate mean values and smaller confidence intervals. This is consistent with the improved interrogation of the bottom layer at larger s-d separations as predicted by Figure 5.2.

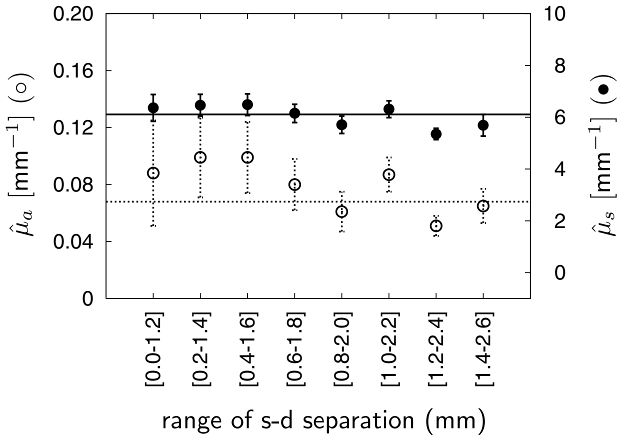


FIG. 5.3. *Recovered bottom layer absorption* (○) *and scattering* (●) *coefficients due to a* 200% μ_a *perturbation in the bottom layer as a function of the range of s-d separations provided by the detector array.*

While the inverse solution results are consistent with the features shown in Figures 5.1 and 5.2, it must be noted that these interrogation maps were generated from the background, not the perturbed system. To focus on changes in the interrogation provided by the perturbed system, we examine a map that displays the relative difference in the interrogation of the perturbed tissue as compared to the background problem (shown in Figure 5.1). Figure 5.4 provides this result, specifically a map of

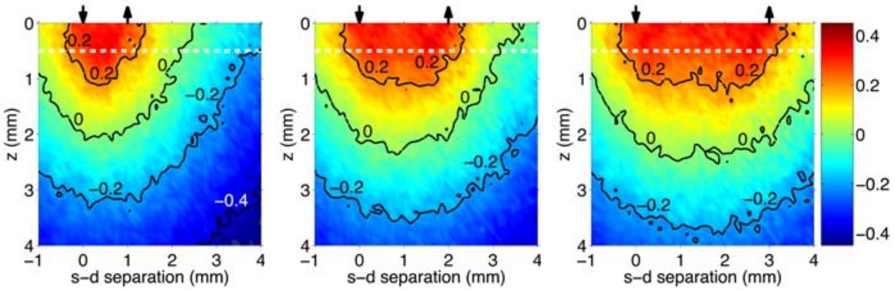


FIG. 5.4. *Plots of the relative difference between the interrogation density function for a* 200% μ_a *perturbation in the bottom layer and the background system for s-d separations of* 1, 2, *and* 3 *mm (left to right).*

$[(\hat{p}_{V\cap D} - p_{V\cap D})/p_{V\cap D}]$ for s-d separations of 1, 2, and 3 mm.

These maps display regions in which the relative difference between the interrogation of the perturbed and background medium is zero (green (0 contour)), increasingly negative (deeper blues (−0.2, −0.4 contours)), and increasingly positive (yellow-orange-red (0.2, 0.4 contours)). Negative values indicate diminished interrogation in the perturbed medium relative to the background system, while positive values indicate enhanced interrogation.

From these maps we observe that the enhancement of interrogation penetrates deeper into the bottom layer with increasing s-d separation. This is consistent with the improved inverse solution results at larger s-d separation. However, this interrogation at larger s-d separation is offset by the increased absorption in the bottom layer of the perturbed system, which depletes the detected signal. This may explain why the $\hat{\mu}_s$ predictions do not improve markedly.

To understand the contributions from each layer, we integrate the data in Figure 5.4 over the top and bottom layers, $(\int_{\mathbb{T}} \hat{p}_{V\cap D} - \int_{\mathbb{T}} p_{V\cap D})/\int_{\mathbb{T}} p_{V\cap D}$ and $(\int_{\mathbb{B}} \hat{p}_{V\cap D} - \int_{\mathbb{B}} p_{V\cap D})/\int_{\mathbb{B}} p_{V\cap D}$, respectively. From these results (shown in Figure 5.5) we see that the probability of interrogating the bottom layer is degraded slightly in the perturbed medium due to the increased absorption in that layer. Despite the detrimental effect of the increased bottom layer absorption, interrogation of the bottom layer still improves with increases of the s-d separation. This again is consistent with the inverse results shown in Figure 5.3. Figure 5.5 also displays the integrated top layer results. Although these results are not pertinent to the bottom layer inverse problem considered here, they may shed light on other inverse problems in which determination of top layer optical properties is desired within a system that is simultaneously undergoing a change in the bottom layer absorption. All top layer values are positive, indicating improved interrogation in this layer of the perturbed medium compared to the background medium due to the increased absorption in the bottom layer.

5.2.2. Top layer $\boldsymbol{\mu_s}$ perturbation. We now examine a second test case involving a 120% increase in μ_s in the top layer relative to the background value. All other optical properties in both layers are held fixed. This results in the following set of optical properties: $\hat{\mu}_s = 7.332$/mm, $\hat{\mu}_a = 0.034$/mm in the top layer and $\mu_s = 6.11$/mm, $\mu_a = 0.034$/mm in the bottom layer. Figure 5.6 displays the recovered optical properties in the top layer as a function of the separation between the source and the linear array of detectors. The results show improved estimates in the mean values of $\hat{\mu}_s$ as well as smaller confidence intervals when the linear detector array is closer to

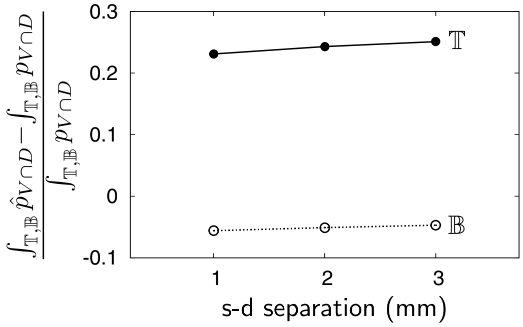


FIG. 5.5. *Relative difference between the interrogation density function for a 200% μ_a perturbation in the bottom layer versus the background, integrated over the top $\mathbb{T}$ and bottom $\mathbb{B}$ layers as a function of s-d separation.*

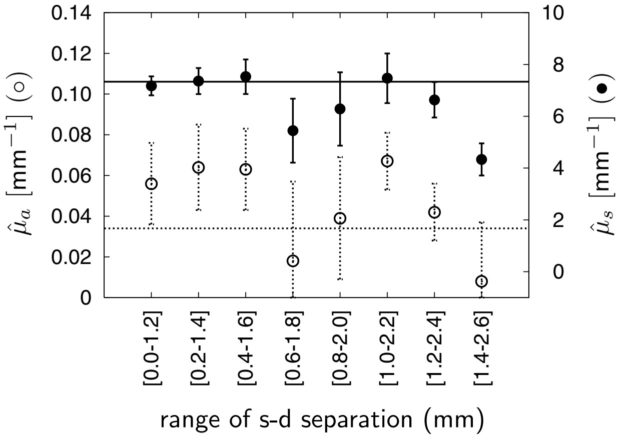


FIG. 5.6. *Recovered top layer absorption (○) and scattering (●) coefficients due to a 120% μ_s perturbation in the top layer as a function of the range of s-d separations provided by the detector array.*

the source. This is in line with the results shown in Figure 5.2 that showed improved interrogation of the top layer at the smaller s-d separations. The recovered mean values of $\hat{\mu}_a$ display no correlation with s-d separation.

For this case it is also useful to examine plots of the relative difference between the perturbed and background medium which are shown in Figure 5.7. The increased scattering in the top layer of the perturbed medium results in enhanced interrogation of this layer. This is especially true when the s-d separation is small, as evidenced by the deep red colors (0.2 contour) in the top layer. However, this enhancement dissipates rapidly as the s-d separation increases and indicates that the increased scattering in the top layer plays a diminishing role in the detected signal. This is easily discerned by focusing attention on the top layer and noticing that while this area is primarily red (0.2 contour) at small s-d separations, it rapidly changes to orange, yellow, and green (0.1 and 0 contours) at larger separations. Moreover, for even larger separations (not shown), this region changes to blue (−0.2 contour). This illustrates that as the s-d separation increases, the increased scattering in the top layer no longer provides an enhanced interrogation of the top layer. This occurs because the photon pathlengths between source and detector increase for larger s-d separations,

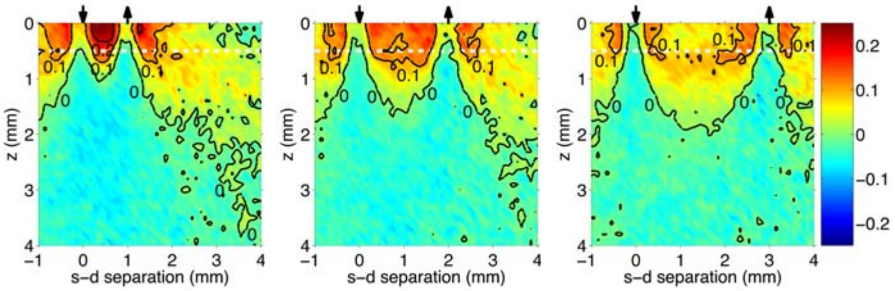


FIG. 5.7. *Plots of the relative difference between the interrogation density function for a* 120% μ_s *perturbation in the top layer and the background system for s-d separations of* 1, 2, *and* 3 *mm (left to right).*

resulting in the depletion of the detected signal by absorption in the top layer. This is the principal cause for the lack of significant improvement in the $\hat{\mu}_a$ predictions with increasing s-d separation shown in Figure 5.6.

The relative difference maps integrated over each layer are shown in Figure 5.8. These plots confirm that the integrated sampling of the top layer is enhanced in the perturbed medium but that this enhancement decreases rapidly with increasing s-d separation. Again, while the integrated bottom layer results are not directly relevant to our top layer inversion, it is interesting to observe that the increased scattering in the top layer has a diminishing effect on the bottom layer interrogation as the s-d separation increases.

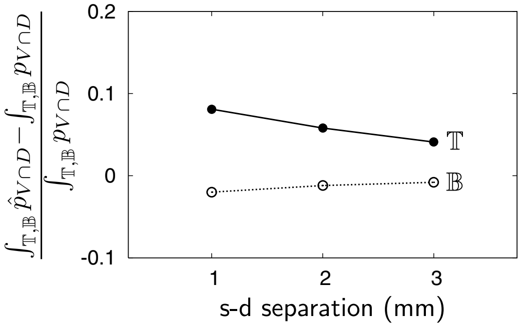


FIG. 5.8. *Relative difference between the interrogation density function for a* 120% μ_s *perturbation in the top layer versus the background, integrated over the top* $\mathbb{T}$ *and bottom* $\mathbb{B}$ *layers as a function of s-d separation.*

It is important to note that for a given medium, the $P(V \cap D)$ maps (Figures 5.1, 5.4, and 5.7) for *all* s-d separations were created from a *single* forward and a *single* adjoint Monte Carlo simulation. To achieve this degree of computational efficiency, we have made use of the symmetry of the layered problem and the probe configuration. This enables the application of linear superposition to align the two simulations for *any* selected s-d separation for the generation of the resulting $P(V \cap D)$ maps. While the use of the coupled forward-adjoint Monte Carlo technique is already more efficient than conventional Monte Carlo simulation, use of symmetry further enhances the computational efficiency of our methodology. A study of the comparative efficiency of conventional Monte Carlo simulation and the coupled forward-adjoint technique developed here is beyond the scope of this paper.

6. Summary and conclusions. We have developed a transport-based technique that determines the joint probability that photons will visit any tissue subvolume and subsequently be detected. Calculation of this conditional system response (system interrogation) is based on an extension of classical reciprocity theory for radiative transport that couples the responses of forward and adjoint Monte Carlo simulations at the boundary of any designated tissue subvolume. These maps of conditional response provide insight as to how s-d configurations affect the spatial distribution of tissue interrogation. While our maps that display the conditional system response were integrated over all angles, it is important to reiterate that angle-specific $P(V \cap D)$ maps can also be generated easily and will be of interest for many applications.

We then applied this computational methodology using data descriptive of a two-layer epithelial/stromal tissue and produced $P(V \cap D)$ maps for varying s-d separations. Moreover, we provided evidence that the maps produced by our coupled forward-adjoint Monte Carlo method provide useful and reliable guidelines for the choice of preferred probe designs, as measured by the successful recovery of optical properties from selected tissue regions.

In biomedical optics applications, the tissue volume targeted for further examination is typically large compared to both the source and detector. In these cases, the coupled forward-adjoint Monte Carlo approach will be especially advantageous from a computational efficiency standpoint. Use of variance reduction methods applied to both forward and adjoint simulations will further increase the efficiency of this new computational method.

REFERENCES

[1] R. Aronson, *Radiative transfer implies a modified reciprocity relation*, J. Opt. Soc. Amer. A, 14 (1997), pp. 486–490.

[2] F. Bevilacqua, J. S. You, C. K. Hayakawa, and V. Venugopalan, *Sampling tissue volumes using frequency-domain photon migration*, Phys. Rev. E (3), 69 (2004), 051908.

[3] D. A. Boas, M. A. O'Leary, B. Chance, and A. G. Yodh, *Detection and characterization of optical inhomogeneities with diffuse photon density waves: A signal-to-noise analysis*, Appl. Opt., 36 (1997), pp. 75–92.

[4] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison–Wesley, Reading, MA, 1967.

[5] S. N. Cramer, *Forward-adjoint Monte Carlo coupling with no statistical error propagation*, Nucl. Sci. Eng., 124 (1996), pp. 398–416.

[6] J. P. Culver, V. Ntziachristos, M. J. Holboke, and A. G. Yodh, *Optimization of optode arrangements for diffuse optical tomography: A singular-value analysis*, Opt. Lett., 46 (2001), pp. 701–703.

[7] S. Feng, F. Zeng, and B. Chance, *Photon migration in the presence of a single defect: A perturbation analysis*, Appl. Opt., 34 (1995), pp. 3826–3837.

[8] E. E. Graves, J. P. Culver, J. Ripoll, R. Weissleder, and V. Ntziachristos, *Singular-value analysis and optimization of experimental parameters in fluorescence molecular tomography*, J. Opt. Soc. Amer. A, 21 (2004), pp. 231–241.

[9] C. K. Hayakawa and J. Spanier, *Perturbation Monte Carlo methods for the solution of inverse problems*, in Monte Carlo and Quasi-Monte Carlo Methods 2002, Springer, Berlin, 2004, pp. 227–241.

[10] C. K. Hayakawa, J. Spanier, F. Bevilacqua, A. K. Dunn, J. S. You, B. J. Tromberg, and V. Venugopalan, *Perturbation Monte Carlo methods to solve inverse photon migration problems in heterogeneous tissues*, Opt. Lett., 26 (2001), pp. 1335–1337.

[11] L. G. Henyey and J. L. Greenstein, *Diffuse radiation in the galaxy*, Astrophys. J., 93 (1941), pp. 70–83.

[12] P. Hoel, S. Port, and C. Stone, *Introduction to Probability Theory*, Houghton Mifflin, Boston, 1971.

[13] R. Hornung, T. H. Pham, K. A. Keefe, M. W. Berns, Y. Tadir, and B. J. Tromberg, *Quantitative near-infrared spectroscopy of cervical dysplasia in vivo*, Human Reproduction,

14 (1999), pp. 2908–2916.

[14] T. Papaioannou, N. W. Preyer, Q. Fang, A. Brightwell, M. Carnohan, G. Cottone, R. Ross, L. R. Jones, and L. Marcu, *Effect of fiber-optic probe design and probe-to-target distance on diffuse reflectance measurements of turbid media: An experimental and computational study at* 337 *nm*, Appl. Opt., 43 (2004), pp. 2846–2860.

[15] M. S. Patterson, S. Andersson-Engels, B. C. Wilson, and E. K. Osei, *Absorption spectroscopy in tissue-simulating materials: A theoretical and experimental study of photon paths*, Appl. Opt., 34 (1995), pp. 22–30.

[16] T. J. Pfefer, L. S. Matchette, A. M. Ross, and M. N. Ediger, *Selective detection of fluorophore layers in turbid media: The role of fiber-optic probe design*, Opt. Lett., 28 (2003), pp. 120–122.

[17] J. C. Schotland, J. C. Haselgrove, and J. S. Leigh, *Photon hitting density*, Appl. Opt., 32 (1993), pp. 448–453.

[18] I. Seo, J. S. You, C. K. Hayakawa, and V. Venugopalan, *Perturbation and differential Monte Carlo methods for measurement of optical properties in a layered epithelial tissue model*, J. Biomed. Opt., 12 (2007), 014030.

[19] I. V. Serov, T. M. John, and J. E. Hoogenboom, *A new effective Monte Carlo midway coupling method in MCNP applied to a well logging problem*, Appl. Radiat. Isot., 49 (1998), pp. 1737–1744.

[20] I. V. Serov, T. M. John, and J. E. Hoogenboom, *A midway forward-adjoint coupling method for neutron and photon Monte Carlo transport*, Nucl. Sci. Eng., 133 (1999), pp. 55–72.

[21] M. C. Skala, G. M. Palmer, C. Zhu, Q. Liu, K. M. Vrotsos, C. L. Marshek-Stone, A. Genfron-Fitzpatrick, and N. Ramanujam, *Investigation of fiber-optic probe designs for optical spectroscopic diagnosis of epithelial pre-cancers*, Lasers Surg. Med., 34 (2004), pp. 25–38.

[22] K. Sokolov, L. T. Nieman, A. Myakov, and A. Gillenwater, *Polarized reflectance spectroscopy for pre-cancer detection*, Technology in Cancer Research and Treatment, 3 (2004), pp. 1–14.

[23] J. Spanier and E. Gelbard, *Monte Carlo Principles and Neutron Transport Problems*, Addison–Wesley, Reading, MA, 1969.

[24] T. Ueki and J. E. Hoogenboom, *Exact Monte Carlo perturbation analysis by forward-adjoint coupling in radiation transport calculations*, J. Comput. Phys., 171 (2001), pp. 509–533.

[25] T. Ueki, J. E. Hoogenboom, and J. L. Kloosterman, *Analysis of correlated coupling of Monte Carlo forward and adjoint histories*, Nucl. Sci. Eng., 137 (2001), pp. 117–145.

[26] L. Wang, S. L. Jacques, and L. Zheng, *Mcml-Monte Carlo modeling of light transport in multi-layered tissues*, Comput. Methods Programs Biomed., 47 (1995), pp. 131–146.

[27] G. H. Weiss, R. Nossal, and R. F. Bonner, *Statistics of penetration depth of photons re-emitted from irradiated tissue*, J. Mod. Opt., 36 (1989), pp. 349–359.

[28] A. J. Welch and M. van Gemert, *Optical-Thermal Response of Laser-Irradiated Tissue*, Plenum Press, New York, 1995.

[29] M. L. Williams, *Generalized contributon response theory*, Nucl. Sci. Eng., 108 (1991), pp. 355–383.

[30] M. L. Williams and W. W. Engle, *The concept of spatial channel theory applied to reactor shielding analysis*, Nucl. Sci. Eng., 62 (1977), pp. 92–104.

[31] C. Zhu, Q. Liu, and N. Ramanujam, *Effect of fiber optic probe geometry on depth-resolved fluorescence measurements from epithelial tissues: A Monte Carlo simulation*, J. Biomed. Opt., 8 (2003), pp. 237–247.

SIAM J. Appl. Math.
Vol. 68, No. 1, pp. 271–291

IDENTIFYING SCATTERING OBSTACLES BY THE CONSTRUCTION OF NONSCATTERING WAVES[*]

D. RUSSELL LUKE[†] AND ANTHONY J. DEVANEY[‡]

Abstract. There are many methods for identifying the shape and location of scatterers from far field data. We take the view that the connections between algorithms are more illuminating than their differences, particularly with regard to the linear sampling method [D. Colton and A. Kirsch, *Inverse Problems*, 12 (1996), pp. 383–393], the point source method [R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Chapman & Hall, London, UK, 2001], and the MUSIC algorithm [A. J. Devaney, *IEEE Trans. Antennas and Propagation*, 53 (2005), pp. 1600–1610]. Using the first two techniques we show that, for a scatterer with Dirichlet boundary conditions, there is a nontrivial incident field that does not generate a scattered field. This incident field, written as an expansion of eigenfunctions of the far field operator, is used in the MUSIC algorithm to image the shape and location of the obstacle as those points z where the incident field is orthogonal to the far field pattern due to a point source located at z. This has two intriguing applications, one for inverse scattering and the other for signal design. Numerical examples demonstrate these ideas.

Key words. inverse scattering, MUSIC, linear sampling, point source method

AMS subject classifications. 35R30, 35P25, 94A08

DOI. 10.1137/060674430

1. Introduction.

The inverse scattering literature abounds with methods to determine the shape of scatterers from far field data. Of principal concern here are the MUSIC algorithm [11], the linear sampling method [7], the point source method [26], and the connections between these methods. The connection between the MUSIC algorithm and Kirsch's factorization method [17] has been detailed by Cheney [5] and Kirsch [18] for scattering from point-like inhomogeneities. More recent studies [1, 16, 12, 13] approach an application of the MUSIC algorithm to scatterers of some specified size, relative to the wavelength, and are based on the finite-dimensional multistatic response matrix for point-like scatterers. Our results complement those of Hazard and Ramdani [15], although they were concerned with the mathematical justification of the decomposition of time-reversal operator (DORT) method [27]. The DORT method also relies on the asymptotic behavior of the time-reversal operator as the scatterers become small. Our goal here is to provide an analysis in the continuum of the inverse problem of determining geometric information about scatterers of arbitrary size and shape that are illuminated by fields of arbitrary frequency.

Our central result, Theorem 3.1, is built upon the linear sampling method of Colton and Kirsch [7] and shows that, on the boundary of a scatterer with Dirichlet boundary conditions, there is a nontrivial incident field that has an arbitrarily small far field pattern. With the help of the point source method of Potthast [26], we show in Corollary 3.2 that such an incident field does not generate a scattered field. Theorem 3.5 combines these results as the foundation for a MUSIC algorithm [11] for determining the shape and location of an obstacle. The technique indicates intriguing

[*] Received by the editors November 8, 2006; accepted for publication (in revised form) July 9, 2007; published electronically November 7, 2007.
http://www.siam.org/journals/siap/68-1/67443.html

[†] Department of Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (rluke@math.udel.edu). Support for this author was provided in part by NSF grant DMS-0712796.

[‡] Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 (devaney@ece.neu.edu).

possibilities for the construction of *nonscattering fields* that might be used to shield obstacles from interrogating waves.

To our knowledge the analysis presented here shows for the first time the feasibility of the MUSIC algorithm for determining the shape and location of Dirichlet obstacles without dependence on the size of the obstacle or the frequency of the incident field. The next section introduces our notation and the background for our main theoretical results presented in section 3. Practical implementations of a MUSIC-type algorithm are discussed in section 4. We illustrate the effectiveness of the algorithm with two examples presented in section 5.

2. Scattering background. We consider acoustic scattering of small-amplitude, monochromatic, time-harmonic waves from one or more impenetrable, sound-soft obstacles embedded in an isotropic homogeneous medium. The obstacles are identified by the domain $\Omega \subset \mathbb{R}^m$, $m = 2$ or 3. The domain Ω is assumed to be bounded with a simply connected exterior and C^2 boundary $\partial\Omega$ and the unit outward normal ν. The governing equation is the Helmholtz equation

$$\left(\triangle + k^2\right) v(x) = 0, \quad x \in \Omega^o \subset \mathbb{R}^m, \tag{2.1}$$

where $\triangle$ denotes the Laplacian, $k \geq 0$ is the wavenumber, $\Omega^o := \mathbb{R}^m \backslash \overline{\Omega}$, and the closure of the open exterior is denoted by the complement of Ω, that is, Ω^c. The surface of the obstacle is assumed to be perfectly absorbing or *sound-soft*. This is modeled with Dirichlet boundary conditions: $v = f$ on $\partial\Omega$, where f is continuous on $\partial\Omega$.

2.1. General incident fields. Let

$$v = v^i + v^s, \tag{2.2}$$

where the total field $v : \Omega^c \to \mathbb{C}$ solves (2.1) on Ω^o with boundary data

$$v(x) := 0 \quad \text{for } x \in \partial\Omega. \tag{2.3}$$

The incident field $v^i : \mathbb{R}^m \to \mathbb{C}$ solves (2.1) on $\mathbb{R}^m$. The scattered field $v^s : \Omega^c \to \mathbb{C}$ solves (2.1) on Ω^o with boundary data

$$v^s(x) = -v^i(x) \quad \text{for } x \in \partial\Omega \tag{2.4}$$

and

$$r^{\frac{m-1}{2}}\Big(\frac{\partial}{\partial r} - ik\Big)v^s(x) \to 0, \quad r = |x| \to \infty, \tag{2.5}$$

uniformly in all directions.

By Green's formula we can express the scattered field on Ω^o by the boundary integral

$$v^s(x) = \int_{\partial\Omega} \left\{ \frac{\partial\Phi(x,y)}{\partial\nu(y)} v^s(y) - \Phi(x,y) \frac{\partial v^s}{\partial\nu}(y) \right\} ds(y), \tag{2.6}$$

where $x \in \Omega^o$ and $\Phi(x,y)$ is the outgoing free-space fundamental solution to (2.1), also referred to as Green's function. As $|x| \to \infty$ one can see that the scattered field has the behavior

$$v^s(x) \;=\; \frac{e^{ik|x|}}{|x|^{\frac{(m-1)}{2}}} \left\{ v^\infty(\widehat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \to \infty, \tag{2.7}$$

where the function v^∞ is the far field pattern on the unit sphere $\mathbb{S} := \{\widehat{x} \in \mathbb{R}^m \mid |\widehat{x}| = 1\}$ given by

$$v^\infty(\widehat{x}) = \beta \int_{\partial\Omega} \left(\frac{\partial e^{-\mathrm{i}k\widehat{x}\cdot y}}{\partial\nu}(y) v^s(y) - e^{-\mathrm{i}k\widehat{x}\cdot y} \frac{\partial v^s}{\partial\nu}(y) \right) ds(y) \tag{2.8}$$

for $\widehat{x} \in \mathbb{S}$ with

$$\beta = \begin{cases} \frac{e^{\mathrm{i}\frac{\pi}{4}}}{\sqrt{8\pi k}}, & m = 2, \\ \frac{1}{4\pi}, & m = 3, \end{cases} \qquad k > 0. \tag{2.9}$$

We define next the Herglotz wave operator $\mathcal{H} : L^2(\mathbb{S}) \to H^1_{loc}(\mathbb{R}^m)$ by

$$(\mathcal{H}g)(x) := \int_{\mathbb{S}} e^{-ik\widehat{\eta}\cdot x} g(-\widehat{\eta}) ds(\widehat{\eta}), \quad x \in \mathbb{R}^m. \tag{2.10}$$

The corresponding Herglotz wave function is denoted $v_g(x) := (\mathcal{H}g)(x)$. Here H^1 denotes the Sobolev space of order 1. The signs in our definition are not standard, but they are chosen to assure consistency between the directions of incident waves and measurement points on the far field sphere. The physical interpretation of the signs is more apparent in a limited aperture setting [20].

LEMMA 2.1 (Herglotz wave functions). *Herglotz wave functions $v_g(x) := (\mathcal{H}g)(x)$ with $g \in L^2$ are dense with respect to the $H^1(\Omega)$-norm in the space of solutions to the Helmholtz equation.*

Proof. The proof is found in Theorem 2.3 of [10]. See also [9, Theorem 2.3]. □

By Lemma 2.1 and the trace theorems for elliptic equations [23], we can construct the density g_z such that $v_{g_z}(x) \approx \Phi(x,z)$ and $\frac{\partial v_{g_z}}{\partial\nu}(x) \approx \frac{\partial\Phi(x,z)}{\partial\nu}$ on $\partial\Omega$ arbitrarily closely for $z \in \Omega^o$ with respect to the $H^{1/2}$- and $H^{-1/2}$-norms, respectively. By (2.6) and (2.8), for $z \in \Omega^o$ we have

$$\begin{aligned} v^s(z) &= \int_{\partial\Omega} \left\{ \frac{\partial\Phi(z,y)}{\partial\nu(y)} v^s(y) - \Phi(z,y) \frac{\partial v^s}{\partial\nu}(y) \right\} ds(y) \\ &\approx \int_{\partial\Omega} \left\{ \frac{\partial v_{g_z}(y)}{\partial\nu(y)} v^s(y) - v_{g_z}(y) \frac{\partial v^s}{\partial\nu}(y) \right\} ds(y) \\ &= \int_{\mathbb{S}} \int_{\partial\Omega} \left\{ \frac{\partial e^{ik(-\widehat{x})\cdot y}}{\partial\nu(y)} v^s(y) - e^{ik(-\widehat{x})\cdot y} \frac{\partial v^s}{\partial\nu}(y) \right\} ds(y) g_z(-\widehat{x}) ds(\widehat{x}) \\ &= \frac{1}{\beta} \int_{\mathbb{S}} v^\infty(\widehat{x}) g_z(-\widehat{x}) ds(\widehat{x}). \end{aligned} \tag{2.11}$$

Note that the boundary of the scatterer is no longer involved in the expression for the scattered field. Moreover, the above approximation does not depend on the boundary condition. At each point $z \in \Omega^o$, by the correct choice of the density g_z, we can, in principle, reconstruct the scattered field. In the case of obstacles with Dirichlet boundary conditions, knowing the scattered field allows one to determine the shape and location of the scatterer as the zeros of the total field, or by constructing an indicator function for the scatterer via the eigenfunction expansion theorem [21]. The problem, however, is that the accuracy of this reconstruction depends on finding a density g that approximates the fundamental solution on the boundary of the scatterer, which we do not know! The point source method is concerned mainly with strategies for constructing the density g (see, for example, [24, 25, 26, 20]). Note also that the density g must contain information about the *evanescent* fields in v^s since none of this information is present in the far field pattern v^∞.

2.2. Plane wave scattering. There are two ways to view the last integral in (2.11) that distinguish many numerical methods in inverse scattering. By the first interpretation the last integral in (2.11) is an integral operator with the far field pattern v^∞ as a kernel. By the second interpretation, the kernel of the operator is the density g, and the operator acts on the far field data v^∞. These two different approaches are best illustrated by considering the case of scattering from incident plane waves

$$u^i(x;\widehat{\eta}) := e^{ik(\widehat{\eta})\cdot x}, \tag{2.12}$$

where the incident field is parameterized by the direction of propagation $\widehat{\eta} \in \mathbb{S}$. The corresponding scattered field and far field patterns are denoted $u^s(x;\widehat{\eta})$ and $u^\infty(\widehat{x},\widehat{\eta})$.

When the scattering is from an incident plane wave with direction $\widehat{\eta}$, we define the *far field operator* $\mathcal{F} : L^2(\mathbb{S}) \to L^2(\mathbb{S})$:

$$\mathcal{F}f(\widehat{x}) := \int_{\mathbb{S}} u^\infty(\widehat{x},\widehat{\eta}) f(\widehat{\eta})\ ds(\widehat{\eta}). \tag{2.13}$$

This operator corresponds to the view of (2.11) as an integral operator with the data u^∞ as the kernel. In this case (2.11) becomes

$$u^s(z;\widehat{\eta}) \approx \frac{1}{\beta}\int_{\mathbb{S}} u^\infty(\widehat{x},\widehat{\eta}) g_z(-\widehat{x}) ds(\widehat{x}) = \frac{1}{\beta}\int_{\mathbb{S}} u^\infty(-\widehat{\eta},-\widehat{x}) g_z(-\widehat{x}) ds(\widehat{x}), \tag{2.14}$$

where the last equality follows from the reciprocity relation

$$u^\infty(\widehat{x},\widehat{\eta}) = u^\infty(-\widehat{\eta},-\widehat{x}). \tag{2.15}$$

The fact that $(\mathcal{F}g_z)(-\widehat{\eta}) \approx \beta u^s(z,\widehat{\eta})$ is a coincidence of having selected the correct function g_z to operate on.

By the second interpretation the last integral in (2.11) is an integral operator with the density g_z as a kernel: $\mathcal{A}_{g_z} : L^2(\mathbb{S}) \to \mathbb{X}(z)$, defined as

$$\mathcal{A}_{g_z} f(z) := \frac{1}{\beta}\int_{\mathbb{S}} g_z(\widehat{x}) f(\widehat{x})\ ds(\widehat{x}). \tag{2.16}$$

Here we have left the image space $\mathbb{X}$ ambiguous because the dependence of the kernel of $\mathcal{A}_{g_z}$ on the points z is not specified. Acting on the far field pattern corresponding to an incident plane wave with direction $\widehat{\eta}$, the operator $\mathcal{A}_{g_z}$ can be seen to be a *backpropagation operator* that propagates the far field back to the scattered field at $z \in \Omega^o$. We will occupy ourselves mostly with the latter interpretation, but our principal tool will be the far field operator of the first interpretation.

Lemma 2.2 (far field operator). *The far field operator $\mathcal{F} : L^2(\mathbb{S}) \to L^2(\mathbb{S})$ is compact. $\mathcal{F}$ is injective with dense range if and only if there does not exist a Dirichlet eigenfunction for Ω which is a Herglotz wave function.*

Proof. Compactness follows from the fact that the kernel is continuous. For the remainder of the statement see [8, Corollary 3.18]. □

The far field operator has a useful factorization in terms of a Herglotz wave function and the mapping of radiating solutions to the Helmholtz equation from the boundary data to the far field pattern, denoted by $\mathcal{B}$.

Lemma 2.3 ($\mathcal{B}$). *Assume that k^2 is not an eigenvalue of $-\triangle$ in Ω. The mapping of radiating solutions to the Helmholtz equation from the boundary data to the far field* *pattern,* $\mathcal{B}: H^{1/2}(\partial\Omega) \to L^2(\mathbb{S})$, *is a compact, injective bounded linear operator with dense range and* range$\mathcal{B}$ = range$(\mathcal{F}^*\mathcal{F})^{1/4}$, *where* $\mathcal{F}^*$ *denotes the adjoint of the far field operator. Moreover, the far field pattern of the outgoing fundamental solution to the Helmholtz equation,* $\Phi^\infty(\cdot; z)$, *is in the range of* $\mathcal{B}$ *if and only if* $z \in \Omega$.

Proof. For the proof, see [17, Theorem 3.6] and [17, Theorem 3.7]. See also [4]. □

For any incident wave v^i restricted to $\partial\Omega$ we have $-\mathcal{B}v^i = v^\infty$, and, in particular, incident fields that can be written as superpositions of plane waves, $v^i = \mathcal{H}g$, yield the desired factorization

$$-\mathcal{B}\mathcal{H}g = \mathcal{F}g. \tag{2.17}$$

We slightly abuse the notation since, by our definitions of $\mathcal{H}$ and v^i, we need to include a trace operator restricting them to the boundary $\partial\Omega$. This should be clear from the context.

3. Nonscattering fields.

The next theorem, modeled after the linear sampling method of [4], shows that there is a nontrivial density $\widehat{g}$ that converges to the null space of the far field operator.

Theorem 3.1 (normalized linear sampling). *Let* Ω *be a domain with smooth boundary and assume that* k^2 *is not a Dirichlet eigenvalue for* $-\triangle$ *on* Ω. *If* $z \in \Omega$, *then for every* $\epsilon > 0$ *there exists a solution* g_z *to*

$$\|\mathcal{F}g_z(\cdot) - \Phi^\infty(\cdot; z)\|_{L^2(\mathbb{S})} < \epsilon \tag{3.1a}$$

such that

$$\lim_{z \overset{\Omega}{\to} \partial\Omega} \|\mathcal{F}\widehat{g}_z\|_{L^2(\mathbb{S})} = 0 \quad \textit{and} \quad \lim_{z \overset{\Omega}{\to} \partial\Omega} \left\|\mathcal{H}\widehat{g}_z - \frac{f_z}{\|g_z\|_{L^2(\mathbb{S})}}\right\|_{H^{1/2}(\partial\Omega)} = 0, \tag{3.1b}$$

where

$$\widehat{g}_z := \frac{g_z}{\|g_z\|_{L^2(\mathbb{S})}} \quad \textit{and} \quad f_z \quad \textit{solves} \quad \mathcal{B}f_z(\cdot) = -\Phi^\infty(\cdot; z). \tag{3.1c}$$

Here $\overset{\Omega}{\to}$ *indicates that the limit is taken by points from within* Ω.

Proof. Our proof is modeled after that of [6, Theorem 2.2]. Since $-\Phi^\infty(\cdot; z) \in$ range$(\mathcal{B})$, by Lemma 2.3 there is a solution f_z to

$$\mathcal{B}f_z(\cdot) = -\Phi^\infty(\cdot; z). \tag{3.2}$$

By Lemma 2.1 and the trace theorem [23], since k^2 is not a Dirichlet eigenvalue for the negative Laplacian on Ω, the Herglotz wave operator is injective with dense range in $H^{1/2}(\partial\Omega)$. Hence for any $\epsilon' > 0$ there is a solution $g_z \in L^2(\mathbb{S})$ to

$$\|\mathcal{H}g_z - f_z\|_{H^{1/2}(\partial\Omega)} \leq \epsilon' \tag{3.3}$$

and hence

$$\left\|\mathcal{H}\widehat{g}_z - \frac{f_z}{\|g_z\|_{L^2(\mathbb{S})}}\right\|_{H^{1/2}(\partial\Omega)} \leq \frac{\epsilon'}{\|g_z\|_{L^2(\mathbb{S})}}. \tag{3.4}$$

Then by the continuity of $\mathcal{B}$ and the factorization (2.17) we have

$$\left\|-\mathcal{F}\widehat{g}_z(\cdot) + \frac{\Phi^\infty(\cdot; z)}{\|g_z\|_{L^2(\mathbb{S})}}\right\| = \left\|\mathcal{B}\mathcal{H}\widehat{g}_z - \mathcal{B}\frac{f_z}{\|g_z\|_{L^2(\mathbb{S})}}\right\|_{H^{1/2}(\partial\Omega)} \leq \frac{\epsilon}{\|g_z\|_{L^2(\mathbb{S})}}, \tag{3.5}$$

where ϵ' is small enough that $C\epsilon' < \epsilon$, where C is the norm of $\mathcal{B}$. Now as $z \to \partial\Omega$, we have $f_z(x) \to -\Phi(x,z)$ for $x \in \partial\Omega$; hence $\|f_z\|_{H^{1/2}(\partial\Omega)} \to \infty$ as $z \to \partial\Omega$. Since f_z is approximated by $\mathcal{H}g_z$, it then follows that $\|\mathcal{H}g_z\|_{H^{1/2}(\partial\Omega)} \to \infty$ as $z \to \Omega$. Note also that $\|\mathcal{H}g_z\|_{H^{1/2}(\partial\Omega)} \le \|\mathcal{H}g_z\|_{H^1(\Omega)}$; thus by the Cauchy–Schwarz inequality we have $\|g_z\|_{L^2(\mathbb{S})} \to \infty$ as $z \to \partial\Omega$. In light of (3.5) this yields

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} \left\| -\mathcal{F}\widehat{g}_z(\cdot) + \frac{\Phi^\infty(\cdot;z)}{\|g_z\|_{L^2(\mathbb{S})}} \right\| = \lim_{z \xrightarrow{\Omega} \partial\Omega} \|\mathcal{F}\widehat{g}_z\| = 0,$$

while by (3.4) we have

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} \left\| \mathcal{H}\widehat{g}_z - \frac{f_z}{\|g_z\|_{L^2(\mathbb{S})}} \right\|_{H^{1/2}(\partial\Omega)} = 0.$$

This completes the proof. □

Note that we make no statement about the behavior of $f_z/\|g_z\|_{L^2(\mathbb{S})}$ as $z \xrightarrow{\Omega} \partial\Omega$; hence it is unclear from (3.1b) what the behavior of $\mathcal{H}g_z$ is in the limit as $z \xrightarrow{\Omega} \partial\Omega$. It is an open problem to characterize the rate at which $\|g_z\| \to \infty$ and $\|f_z\| \to \infty$.

Since the far field pattern is zero if and only if there is no scattered field, the above theorem implies that the incident Herglotz wave function $\mathcal{H}\widehat{g}_z$ does not scatter in the limit as $z \to \partial\Omega$. That is, the following corollary holds.

COROLLARY 3.2 (nonscattering incident fields). *Fix any $\epsilon > 0$ and let g_z satisfy* (3.1a) *with*

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} \|\mathcal{F}\widehat{g}_z\|_{L^2(\mathbb{S})} = 0 \quad \text{and} \quad \lim_{z \xrightarrow{\Omega} \partial\Omega} \left\| \mathcal{H}\widehat{g}_z - \frac{f_z}{\|g_z\|_{L^2(\mathbb{S})}} \right\|_{H^{1/2}(\partial\Omega)} = 0,$$

where

$$\widehat{g}_z := \frac{g_z}{\|g_z\|_{L^2(\mathbb{S})}} \quad \text{and} \quad f_z \quad \text{solves} \quad \mathcal{B}f_z(\cdot) = -\Phi^\infty(\cdot;z).$$

Then the scattered field, $v^s_{\widehat{g}_z}$, corresponding to the incident Herglotz wave function $v^i_{\widehat{g}_z} = \mathcal{H}\widehat{g}_z$, has the behavior

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} v^s_{\widehat{g}_z}(x) = 0 \quad \text{for} \quad x \in \Omega^o, \qquad \text{while} \qquad \lim_{x \xrightarrow{\Omega^o} \partial\Omega} \lim_{z \xrightarrow{\Omega} \partial\Omega} v^i_{\widehat{g}_z}(x) = 0.$$

Our proof relies on the backpropagation interpretation of (2.11) that is central to the point source method.

LEMMA 3.3 (backpropagation). *Assume that k^2 is not a Dirichlet eigenvalue of $-\triangle$ on Ω and let $x \in \Omega^o$. Given any $\delta' > 0$, there exists an $\epsilon'(x) > 0$ such that for all $p_x \in L^2(\mathbb{S})$ satisfying*

$$\left\| \Phi(\cdot,x) - \mathcal{H}p_x(\cdot) \right\|_{H^{1/2}(\partial\Omega)} < \epsilon'(x) \tag{3.6}$$

one has

$$\left| u^s(x,\widehat{\eta}) - (\mathcal{A}_{p_x}u^\infty)(x,\widehat{\eta}) \right| < \delta', \tag{3.7}$$

where $\mathcal{A}_{p_x}$ is defined by (2.16).

Proof. The proof is a special case of [26, Theorem 5.1.2]. See also [20, Theorem 1]. □

Proof of Corollary 3.2. To show the first limit we construct a backpropagation operator $\mathcal{A}_{p_x}$ to approximate $v^s_{\widehat{g}_z}$. For $\delta' > 0$ and $x \in \Omega^o$, by Lemma 3.3 there is an $\epsilon'(x) > 0$ such that $p_x \in L^2(\mathbb{S})$ satisfying (3.6) also satisfies (3.7). The existence of such a p_x follows from the denseness of the Herglotz wave operator. Next denote

$$v^\infty_{\widehat{g}_z} := \mathcal{F}\widehat{g}_z, \tag{3.8}$$

where $\widehat{g}_z$ is the density in Theorem 3.1. By [8, Lemma 3.16] the scattered field corresponding to $v^\infty_{\widehat{g}_z}$ is

$$v^s_{\widehat{g}_z}(x) = \int_{\mathbb{S}} u^s(x; -\widehat{\eta})\widehat{g}_z(-\widehat{\eta})\ ds(\widehat{\eta}); \tag{3.9}$$

hence by (2.16), (3.9), and the Cauchy–Schwarz inequality

(3.10)

$$\begin{aligned}
&\left|v^s_{\widehat{g}_z}(x) - (\mathcal{A}_{p_x} v^\infty_{\widehat{g}_z})(x)\right| \\
&\quad = \left| \int_{\mathbb{S}} u^s(x; -\widehat{\eta})\widehat{g}_z(-\widehat{\eta})\ ds(\widehat{\eta}) - \frac{1}{\beta}\int_{\mathbb{S}} p_x(\widehat{y}) \left(\int_{\mathbb{S}} u^\infty(\widehat{y}; -\widehat{\eta})\widehat{g}_z(-\widehat{\eta})\ ds(\widehat{\eta}) \right)\ ds(\widehat{y}) \right| \\
&\quad = \left| \int_{\mathbb{S}} \widehat{g}_z(-\widehat{\eta}) \left(u^s(x; -\widehat{\eta}) - \frac{1}{\beta}\int_{\mathbb{S}} p_x(\widehat{y}) u^\infty(\widehat{y}; -\widehat{\eta})\ ds(\widehat{y}) \right)\ ds(\widehat{\eta}) \right| \\
&\quad < C\delta' \|\widehat{g}_z\|_{L^2(\mathbb{S})} = C\delta',
\end{aligned}$$

where C is the surface area of the unit sphere. For x and δ' fixed, $\mathcal{A}_{p_x}$ is bounded and linear, independent of $\widehat{g}_z$; thus, since $\lim_{z \xrightarrow{\Omega} \partial\Omega} \|v^\infty_{\widehat{g}_z}\|_{L^2(\mathbb{S})} = 0$, it follows that $\lim_{z \xrightarrow{\Omega} \partial\Omega} |\mathcal{A}_{p_x} v^\infty_{\widehat{g}_z}(x)| = 0$. Hence by (3.10) and the triangle inequality,

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} \left|v^s_{\widehat{g}_z}(x) - (\mathcal{A}_{p_x} v^\infty_{\widehat{g}_z})(x)\right| = \lim_{z \xrightarrow{\Omega} \partial\Omega} |v^s_{\widehat{g}_z}(x)| < C\delta' \tag{3.11}$$

for arbitrary $\delta' > 0$, which completes the proof of the first statement.

To see the corresponding incident field behavior, note that the total field $v^i_{\widehat{g}_z} + v^s_{\widehat{g}_z}$ is continuous and since Ω has Dirichlet boundary conditions,

$$(v^i_{\widehat{g}_z} + v^s_{\widehat{g}_z})(x) = 0 \quad \text{for} \quad x \in \partial\Omega.$$

Thus, given any $\epsilon'' > 0$, there is a $\rho > 0$ such that for all $z \in \Omega$ and $x \in \Omega^o$ with dist $(x, \Omega) < \rho$, we have

$$|v^i_{\widehat{g}_z}(x) + v^s_{\widehat{g}_z}(x)| < \frac{\epsilon''}{2}.$$

Now, since $\lim_{z \xrightarrow{\Omega} \partial\Omega} |v^s_{\widehat{g}_z}| = 0$ pointwise, we have by the triangle inequality, $|v^i_{\widehat{g}_z}(x)| \leq \epsilon''$ for arbitrary $\epsilon'' > 0$ and x near enough to $\partial\Omega$. This completes the proof. □

Remark 3.4. By Theorem 3.1, the fact that $\lim_{x \xrightarrow{\Omega^o} \partial\Omega} \lim_{z \xrightarrow{\Omega} \partial\Omega} v^i_{\widehat{g}_z}(x) = 0$ implies that the gradient of this field is the sole contribution to the $H^{1/2}$-norm on $\partial\Omega$. Also note that we have made no assumptions about the frequency or the size of the scatterers, other than to assume that the wavenumber is not a Dirichlet eigenvalue for the scatterer.

Following [15] we interpret the integral operator on the right-hand side of (2.14) as a time-reversal operator for the multistatic data of an antenna array arranged on the aperture $\mathbb{S}$ emitting time-harmonic fields. A transducer located at $r\widehat{\eta}$ for $r \gg 1$ emits a spherically spreading field $\Phi(r\widehat{\eta}, x)$ which, in the region of the scatterer Ω, is approximately the plane wave $u^i(x, -\widehat{\eta})$. The resulting scattered field is measured in the far field at the antenna element located at $r\widehat{x}$. The recorded data $u^\infty(\widehat{x}, -\widehat{\eta})$ is reversed, or backpropagated, in order to reconstruct the scattered field around the obstacle. This multistatic data array is thus the discrete realization of the far field operator $\mathcal{F}$. The connection between the MUSIC algorithm and Kirsch's factorization method for scattering from an inhomogeneous medium has been detailed in [5, 18]. We will have more to say about the discrete operator in section 4.2, where we investigate the spatial resolution as a function of the far field sampling frequency and the number of incident fields.

Denote the singular system of $\mathcal{F}$ by $(\sigma_n, \xi_n, \psi_n)$, where

$$\mathcal{F}\xi_n = \sigma_n\psi_n \quad \text{and} \quad \mathcal{F}^*\psi_n = \sigma_n\xi_n \tag{3.12}$$

with singular values $|\sigma_n| > |\sigma_m|$ for $m > n$, and left and right singular functions ψ_n and ξ_n, respectively. Then, by (2.14), for the correct g_z we have

$$\mathcal{F}^* u^s(z, \cdot) \approx \Psi\Sigma\Xi^* g_z,$$

where Ψ and Ξ are the singular operators corresponding to ψ_n and ξ_n, respectively, and Σ is a diagonal operator with the singular values σ_n on the diagonal.

By Lemma 2.2, $\mathcal{F}$ has at most a countable number of discrete eigenvalues with zero as the only possible cluster point. In fact, zero is an eigenvalue if and only if k^2 is an eigenvalue of the negative Laplacian on the interior of Ω with corresponding eigenfunction a Herglotz wave function. Such k, if they exist, form a discrete set [23]. Thus, the null space of $\mathcal{F}$ is almost always trivial, though the eigenvalues decay exponentially by the analyticity of the kernel of $\mathcal{F}$.

The MUSIC algorithm is based on the observation that the set of Green's functions

$$\Phi^\infty(\widehat{\eta}; z) := \lim_{r\to\infty} \Phi(r\widehat{\eta}, z) = \beta e^{\mathrm{i}k(-\widehat{\eta})\cdot z}, \tag{3.13}$$

for z near $\partial\Omega$ and all $\widehat{\eta} \in \mathbb{S}$, are nearly orthogonal to the *noise subspace* of $\mathcal{F}$. We discuss what we mean by the noise subspace in more detail in the next section. In precise terms we have the following theorem.

THEOREM 3.5 (MUSIC). *Let Ω be a domain with smooth boundary and assume that k^2 is not a Dirichlet eigenvalue for the negative Laplacian on Ω. Let $(\sigma_n, \xi_n, \psi_n)$, $n \in \mathbb{N}$, be the singular system for the far field operator $\mathcal{F}$ with $|\sigma_n| \le |\sigma_m|$ for $n > m$. Given any $\gamma > 0$, there is a vector $a \in l^2$ with $\|a\|_2 = 1$ and $\rho > 0$ such that for any $x \in \Omega^o$ satisfying* dist $(x, \Omega) < \rho$ *we have*

$$\sum_{n=1}^{\infty} \left| a_n \langle \xi_n, \Phi^\infty(\cdot; x)\rangle_{L^2(\mathbb{S})} \right| < \gamma. \tag{3.14}$$

Proof. Let g_z and $\widehat{g}_z$ satisfy (3.1a)–(3.1b). By Corollary 3.2, there are $\delta > 0$ and $\rho > 0$ such that

$$|v^i_{\widehat{g}_z}(x)| < \gamma \quad \text{whenever} \quad \text{dist}\,(z, \partial\Omega) < \delta \ (z \in \Omega) \quad \text{and} \quad \text{dist}\,(x, \Omega) < \rho \ (x \in \Omega^o). \tag{3.15}$$

The density $\widehat{g}_z$ can be written as a linear combination of the singular functions ξ_n [8, Theorem 4.8]:

$$\widehat{g}_z = \sum_{n=1}^{\infty} \widehat{a}_n \xi_n, \qquad \text{where} \qquad \widehat{a}_n = \frac{1}{\sigma_n}(\widehat{G}_z, \psi_n) \qquad \text{and} \qquad \widehat{G}_z := \mathcal{F}\widehat{g}_z. \tag{3.16}$$

This and (3.15) yield

$$\left| \sum_n \widehat{a}_n \left\langle \xi_n, u^i(x, -\cdot) \right\rangle \right| < \gamma \qquad \text{for } \operatorname{dist}(x, \Omega) < \rho. \tag{3.17}$$

Next, we construct a new density $\widetilde{g}_z$ from $\widehat{g}_z$ by rotating the coefficients $\widehat{a}_n$ in the complex plane in such a way that the sum corresponding to (3.17) is of the magnitudes of the individual terms. Define

$$\widetilde{g}_z = \sum_{n=1}^{N} \widetilde{a}_n \xi_n, \qquad \text{where} \quad \widetilde{a}_n := e^{\mathrm{i}\theta_n} \widehat{a}_n \text{ for } \theta_n := -\left(\arg(\left\langle \xi_n,\ u^i(x, -\cdot) \right\rangle) + \arg(\widehat{a}_n)\right). \tag{3.18}$$

Note first that $\|\widetilde{g}_z\|_{L^2(\mathbb{S})} = \|\widetilde{g}_z\|_{L^2(\mathbb{S})}$ and $\|\mathcal{F}\widetilde{g}_z\|_{L^2(\mathbb{S})} = \|\mathcal{F}\widehat{g}_z\|_{L^2(\mathbb{S})}$. As in the proof of Corollary 3.2 we construct a backpropagation operator $\mathcal{A}_{p_x}$ to approximate $v^s_{\widetilde{g}_z}$. For $\delta' > 0$ and $x \in \Omega^o$, by Lemma 3.3 there is an $\epsilon'(x) > 0$ such that $p_x \in L^2(\mathbb{S})$ satisfying (3.6) also satisfies (3.7). Next denote

$$v^{\infty}_{\widetilde{g}_z} := \mathcal{F}\widetilde{g}_z.$$

Again by [8, Lemma 3.16] the scattered field corresponding to $v^{\infty}_{\widetilde{g}_z}$ is

$$v^s_{\widetilde{g}_z}(x) = \int_{\mathbb{S}} u^s(x; -\widehat{\eta}) \widetilde{g}_z(-\widehat{\eta})\ ds(\widehat{\eta}); \tag{3.19}$$

hence by (2.16), (3.19), and the Cauchy–Schwarz inequality,

$$\begin{aligned} &\left| v^s_{\widetilde{g}_z}(x) - (\mathcal{A}_{p_x} v^{\infty}_{\widetilde{g}_z})(x) \right| \\ &\quad = \left| \int_{\mathbb{S}} \widetilde{g}_z(-\widehat{\eta}) \left(u^s(x; -\widehat{\eta}) - \frac{1}{\beta} \int_{\mathbb{S}} p_x(\widehat{y}) u^{\infty}(\widehat{y}; -\widehat{\eta})\ ds(\widehat{y}) \right)\ ds(\widehat{\eta}) \right| \\ &\quad < C\delta' \|\widetilde{g}_z\|_{L^2(\mathbb{S})} = C\delta', \end{aligned} \tag{3.20}$$

where C is the surface area of the unit sphere. For x and δ' fixed, $\mathcal{A}_{p_x}$ is bounded and linear, independent of $\widetilde{g}_z$; thus, since $\lim_{z \xrightarrow{\Omega} \partial\Omega} \|v^{\infty}_{\widetilde{g}_z}\|_{L^2(\mathbb{S})} = \lim_{z \xrightarrow{\Omega} \partial\Omega} \|v^{\infty}_{\widehat{g}_z}\|_{L^2(\mathbb{S})} = 0$, it follows that $\lim_{z \xrightarrow{\Omega} \partial\Omega} |\mathcal{A}_{p_x} v^{\infty}_{\widetilde{g}_z}(x)| = 0$. Hence by (3.20) and the triangle inequality,

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} \left| v^s_{\widetilde{g}_z}(x) - (\mathcal{A}_{p_x} v^{\infty}_{\widetilde{g}_z})(x) \right| = \lim_{z \xrightarrow{\Omega} \partial\Omega} |v^s_{\widetilde{g}_z}(x)| < C\delta' \tag{3.21}$$

for arbitrary $\delta' > 0$.

As in the proof of Corollary 3.2, the corresponding incident field behavior follows from the continuity of the total field and the fact that Ω has Dirichlet boundary conditions; that is, $|v^i_{\widetilde{g}_z}(x)| \leq \epsilon''$ for arbitrary $\epsilon'' > 0$ and x near enough to $\partial\Omega$. In summary, the scattered and incident fields corresponding to the density $\widetilde{g}_z$ have the behavior

$$\lim_{z \xrightarrow{\Omega} \partial\Omega} v^s_{\widetilde{g}_z}(x) = 0 \quad \text{for } x \in \Omega^o \qquad \text{and} \qquad \lim_{x \xrightarrow{\Omega^o} \partial\Omega} \lim_{z \xrightarrow{\Omega} \partial\Omega} v^i_{\widetilde{g}_z}(x) = 0;$$

hence given any $\gamma > 0$, there are $\delta > 0$ and $\rho > 0$ such that $\left|\sum_n \widetilde{a}_n \left\langle \xi_n(\cdot), u^i(x, -\cdot)\right\rangle\right| < \gamma$ whenever dist $(z, \partial\Omega) < \delta$ $(z \in \Omega)$ and dist $(x, \Omega) < \rho$ $(x \in \Omega^o)$. But by our construction of $\widetilde{a}_n$ the summands are all nonnegative real numbers; that is,

$$\sum_n \left|\widetilde{a}_n \left\langle \xi_n(\cdot), u^i(x, -\cdot)\right\rangle\right| < \gamma \quad \text{for dist } (x, \Omega) < \rho. \tag{3.22}$$

Finally, recalling that $u^i(x, -\widehat{\eta}) = \frac{1}{\beta}\Phi^\infty(\widehat{\eta}; x)$, after normalization of the coefficients $\widetilde{a}_n$ the result (3.14) follows. □

Remark 3.6. Inequality (3.17) alone could be used for imaging with the MUSIC methodology; however, the contrast of the resulting images is not strong enough for adequate results. In other words, x need not be very close to Ω in order to satisfy (3.17), and the resulting image does not have a sharp cutoff near the boundary.

4. Practical implementation.

4.1. The MUSIC algorithm. Theorem 3.1 only states that *there exists* a density $\widehat{g}_z$ that can be used to construct nonscattering incident fields; it does not, however, suggest *how* one might calculate such a density. Arens [3] has shown that for a sound-soft scatterer as we have here a regularization strategy such as Tikhonov regularization or spectral cutoff gives rise to a density with the desired properties. In other words, if at a point $z \in \Omega$ we solve (3.1a) using Tikhonov or spectral cutoff regularization, then the corresponding density $\widehat{g}_{\alpha,z}$ can be used to construct an incident field satisfying Corollary 3.2 in the limit as the regularization parameter $\alpha \to 0$. There are two reasons why this is impractical: first, we do not know where the scatterer lies, and second, we do not know about the behavior of $\widehat{g}_z$ for points $z \in \Omega^o$. Note, however, that the location of point z in the computation of the density $\widehat{g}_z$ is arbitrary, so long as it is not in the exterior of Ω. This suggests that the orthogonality of $\widehat{g}_z$ with the fundamental solution far fields $\Phi^\infty(\cdot; x)$ expressed in (3.14) is a phenomenon more intimately tied to the spectrum of the far field operator $\mathcal{F}$ than to the particular density $\widehat{g}_z$. Indeed, as (3.1b) shows, the desired density is in the "noise space" of $\mathcal{F}$. Denote the noise subspace of $\mathcal{F}$ by $\mathcal{N}_\gamma$ corresponding to the span of the singular functions ξ_n with singular values $|\sigma_n| < \gamma$ for $n > N_\gamma$. In the numerical experiments detailed below we take $\widehat{g}$ to be simply a linear combination of the elements $\xi_n \in \mathcal{N}_\gamma$ for a large enough cutoff.

In the conventional MUSIC application one usually works with the MUSIC γ-pseudospectrum defined by

$$\mathcal{P}(x) := \frac{1}{\sum_{n>N_\gamma} \left|a_n \langle \xi_n(\cdot), \Phi^\infty(\cdot; x)\rangle_{L^2(\mathbb{S})}\right|} \geq \frac{1}{\gamma}. \tag{4.1}$$

This is what is usually imaged as a function of x. Note that for $x \in \Omega^o$, we have $\mathcal{P}(x) \to \infty$ as $x \to \partial\Omega$, yielding the image of the support of the obstacle as the points where $\mathcal{P}(x)$ is large.

Our focus thus far has been on finding the location and shape of the scatterer, but the fact that the constructed incident field is arbitrarily small at the boundary of the scatterer opens the door to the construction of fields that *avoid* certain obstacles that one might like to protect, while targeting others. In other words, the constructed incident field $v^i_{\widehat{g}_z}$ effectively *does not scatter.* In order to illustrate this point, in our numerical experiments, instead of the usual MUSIC implementation, we show the *inverse* of the γ-pseudospectrum.

4.2. Resolution analysis. The analysis above is in the continuum. In any practical application one will sample the far field at a finite number of discrete points $\widehat{x}_i$ for a finite number of incident fields with direction $\widehat{\eta}_j$; that is, the far field operator $\mathcal{F}$ given by (2.13) is replaced with the discrete multistatic response matrix $F \in \mathbb{C}^{M\times N}$. In this section we apply sampling criteria derived from the physical optics approximation in order to estimate how many incident fields and far field measurements one needs in order to achieve a specified spatial resolution. Other approaches are presented in [2, 16].

The criteria we develop are based on the physical optics approximation which is valid for very large wavenumbers k. The technique discussed above is not dependent on the wavenumber. Indeed, it works especially well at small wavenumbers in the *resonance region* for the scatterer, that is, where the wavelength is on the order of the scatterer. Our estimates for sampling rates are overestimates in the sense that the spatial resolution predicted from a particular sampling rate in the far field is not as fine as what is actually achieved. The analysis of this section thus provides lower bounds on the predicted spatial resolution from a given sampling rate.

To begin, we recall the physical optics, or Kirchhoff approximation. Our treatment is standard (see [8, 19]), with the exception that our derivation also holds in $\mathbb{R}^2$. The calculation follows [20] where $\mathbb{R}^2$ and $\mathbb{R}^3$ are considered, but is short enough to include here. For very large wavenumbers, that is, very small wavelengths relative to the curvature of the obstacle, the face upon which the incident field impinges is nearly planar. As such, we can then approximate the normal derivative of the scattered field by the normal derivative of the incident field. Define Ω_+ to be the illuminated side of the scattering domain $\Omega_+ := \{x \in \partial\Omega \mid \langle \nu(x), \widehat{\eta}\rangle < 0\}$. The shadow of the scattering domain, Ω_-, is defined as $\Omega_- := \partial\Omega \setminus \overline{\Omega_+}$. The physical optics approximation for the scattered field is written

$$\frac{\partial u^s(x,\widehat{\eta})}{\partial\nu(x)} \approx \begin{cases} \frac{\partial u^i(x,\widehat{\eta})}{\partial\nu(x)}, & x \in \Omega_+, \\ -\frac{\partial u^i(x,\widehat{\eta})}{\partial\nu(x)}, & x \in \Omega_-, \end{cases} \qquad k \gg 0. \tag{4.2}$$

This leads to the physically intuitive approximation that the normal derivative of the total field is twice the normal derivative of the incident field on the illuminated side and zero on the shadow of the scatterer.

Together with the representation for the scattered field [8, Theorem 3.12],

$$u^s(x,\widehat{\eta}) = -\int_{\partial\Omega} \Phi(x,y)\frac{\partial u(y,\widehat{\eta})}{\partial\nu(y)}ds(y) \tag{4.3}$$

for $x \in \Omega^o$, the Kirchhoff approximation yields

$$u^s(x,\widehat{\eta}) \approx -2\int_{\partial\Omega_+} \Phi(x,y)\frac{\partial}{\partial\nu(y)}u^i(y,\widehat{\eta})\; ds(y), \tag{4.4}$$

and

$$u^\infty(\widehat{x},\widehat{\eta}) \approx -2\beta\int_{\partial\Omega_+} e^{-iky\cdot\widehat{x}}\frac{\partial}{\partial\nu(y)}u^i(y,\widehat{\eta})\; ds(y) = -2ik\beta\int_{\partial\Omega_+} e^{iky\cdot(\widehat{\eta}-\widehat{x})}\widehat{\eta}\cdot\nu(y)\; ds(y). \tag{4.5}$$

Similarly, on the shadow region we have

$$u^\infty(\widehat{x},-\widehat{\eta}) \approx 2\beta\int_{\partial\Omega_-} e^{-iy\cdot\widehat{x}}\frac{\partial}{\partial\nu(y)}u^i(y,-\widehat{\eta})\; ds(y) = 2ik\beta\int_{\partial\Omega_-} e^{-iky\cdot(\widehat{\eta}+\widehat{x})}\widehat{\eta}\cdot\nu(y)\; ds(y). \tag{4.6}$$

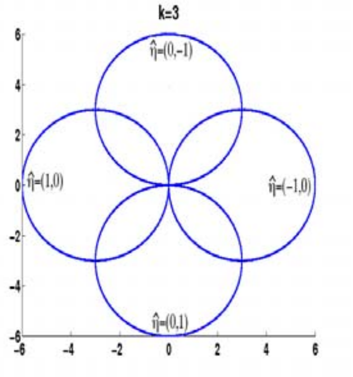


FIG. 1. *Sampling in the Fourier domain of the scatterer* Ω *corresponding to the geometry of the far field pattern in* $\mathbb{R}^2$*. The far field data are depicted here with wavenumber* $k = 3$ *at four incident fields with directions* $\widehat{\eta} = (-1,0), (0,-1), (1,0)$*, and* $(0,1)$*. For each fixed* $\widehat{\eta}$ *the far field samples are depicted here as a continuum on a full aperture* $\mathbb{S}$*.*

The divergence theorem together with (4.5)–(4.6) yields

$$\begin{aligned} u^\infty(\widehat{x},\widehat{\eta}) + \overline{u^\infty(-\widehat{x},-\widehat{\eta})} &\approx -2ik\beta \int_{\partial\Omega} e^{iky\cdot(\widehat{\eta}-\widehat{x})}\widehat{\eta}\cdot\nu(y)ds(y) + R(\widehat{x},\widehat{\eta}) \\ &= 2\beta k^2(1-\widehat{\eta}\cdot\widehat{x}) \int_{\Omega} e^{ikz\cdot(\widehat{\eta}-\widehat{x})}\,dz + R(\widehat{x},\widehat{\eta}) \\ &= 2(2\pi)^{m/2}\beta k^2(1-\widehat{\eta}\cdot\widehat{x})\widehat{\mathcal{X}_\Omega}(k(\widehat{x}-\widehat{\eta})) + R(\widehat{x},\widehat{\eta}), \end{aligned} \tag{4.7}$$

where $\widehat{\mathcal{X}_\Omega}$ is the Fourier transform on $\mathbb{R}^m (m = 2 \text{ or } 3)$ of the indicator function of the obstacle and

$$R(\widehat{x},\widehat{\eta}) = 4ik\mathrm{Im}(\beta) \int_{\partial\Omega_-} e^{iky\cdot(\widehat{\eta}-\widehat{x})}\widehat{\eta}\cdot\nu(y)ds(y). \tag{4.8}$$

In most discussions of the physical optics approximation (see, for example, [19]) the setting is $\mathbb{R}^3$, and here the remainder term does not appear since by (2.9) $\mathrm{Im}(\beta) = \mathrm{Im}(1/(4\pi)) = 0$. In $\mathbb{R}^2$, however, β is complex-valued, which gives rise to the unusual remainder term in the calculations above.

The connection to the Fourier transform above allows us to estimate the sampling requirements for the scatterer via the Whittaker–Shannon sampling theorem. Our data, the far field pattern, are in the so-called Fourier or frequency domain of the scatterer. These data lie on circles in the Fourier domain of the scatterer centered at the point $-k\widehat{\eta}$, where $\widehat{\eta}$ is the direction of the incident field. This is depicted in Figure 1.

For our purposes it is not necessary to carry out a detailed sampling calculation for the geometry shown in Figure 1—an estimate based on sampling on a rectangular grid suffices. Our discussion of the sampling theory is terse; interested readers are referred to [14] for more details. We consider a cubic lattice of samples of some smooth function φ on $\mathbb{R}^m$ defined by

$$\varphi_s := \mathrm{comb}\left(\frac{x}{\Delta x}\right)\varphi(x), \tag{4.9}$$

where Δx is distance between the samples and comb (x) is the m-dimensional "comb" function

$$\text{comb }(x) := \sum_{|n|=-\infty}^{\infty} \delta(x-n)$$

for n a multi-index depending on the dimension $m = 2$ or 3. By the convolution theorem, the Fourier spectrum of the sampled function is

$$\widehat{\varphi}_s(\xi) = (\Delta x)^m \,\text{comb }(\Delta x \xi) * \widehat{\varphi}(\xi),$$

where ξ is the Fourier dual variable to x and $*$ denotes convolution. It can be shown [14, eq. (2–53)] that the sampled Fourier spectrum has the explicit representation

$$\widehat{\varphi}_s(\xi) = \sum_{|n|=-\infty}^{\infty} \widehat{\varphi}\left(\xi - \frac{n}{\Delta x}\right), \tag{4.10}$$

where, again, n is a multi-index. If we assume that φ is bandlimited, then $\widehat{\varphi}$ has compact support. Suppose $\widehat{\varphi}$ is supported on the cube $\mathcal{R}$. If the sample spacing $1/\Delta x$ is large enough that for all $\xi \in \mathcal{R}$

$$\widehat{\varphi}\left(\xi - \frac{n}{\Delta x}\right) = \widehat{\varphi}_s\left(\xi\right),$$

then by (4.10) the sampled Fourier spectrum is just a periodic extension of the true Fourier spectrum; hence we can reconstruct φ exactly from the spectrum of the sampled function. If r is the length of the *smallest* cube that supports the spectrum of φ, then the sampled spectrum will exactly represent the true spectrum as long as

$$\Delta x \leq \frac{1}{2r}.$$

When equality holds, the sampling is said to be at the Nyquist frequency. At the Nyquist frequency, we have the Whittaker–Shannon sampling theorem

$$\varphi(x) = \sum_{|n|=-\infty}^{\infty} \varphi\left(\frac{n}{2r}\right) \text{sinc }\left(2r\left(x - \frac{n}{2r}\right)\right), \tag{4.11}$$

where sinc is the m-dimensional sinc function.

Let us suppose that the smallest feature of our scatterer is $1/M$ of the size of the illuminating wavelength. By the Whittaker–Shannon sampling theorem, a sampling rate of at least $1/(2M)$ in the physical domain represents a highest frequency component of M in each direction and thus $2M$ sample points on a Cartesian grid along each dimension in the Fourier domain. In Figure 1 we see that the "frequency domain" is covered by circles of radius k centered at $-k\widehat{\eta}$. The gaps in the frequency domain are determined by the smallest sampling rate with respect to either the direction of incidence $\widehat{\eta}$ or the far field samples $\widehat{x}$. Suppose that the far field is sampled at infinitesimal intervals and the directions of incidence are sampled at $4N$ points evenly distributed on $[-\pi, \pi]$. The largest gap in the frequency domain for this sampling geometry is bounded above by $2\pi k\sqrt{2}/N$. In order to achieve the same resolution as could be achieved by sampling on the Cartesian grid, N must be chosen so that

$$\frac{2\pi\sqrt{2}}{N} \leq \frac{1}{M}. \tag{4.12}$$

This yields a conservative lower bound on the sampling frequency N of the far field pattern and incident field directions needed to achieve a desired spatial resolution $1/M$ relative to the wavelength of the illuminating field. Since this analysis is based on the physical optics approximation, we expect these sampling requirements in the far field to be greater than what will actually be needed to resolve the scatterers. This is illustrated in section 5.

5. Examples.

5.1. An infinite cylinder. As a first example we consider scattering from an infinite cylinder over which the field satisfies homogeneous Dirichlet conditions. While this example is didactic it has the advantage that the fields have explicit formulations. Taking advantage of radial symmetry, we parameterize directions on the unit sphere $\mathbb{S}$ by the angles α, where α_i is the direction of the incident field and α_0 is the observation point on the far field sphere. The incident and scattered fields can be represented in series of Bessel and Hankel functions, respectively. Let b be the cylinder radius, and let J_n and H_n^+ denote Bessel and Hankel functions of the first kind, respectively; then

$$u^\infty(\alpha_0, \alpha_i) = -\sum_{n=-\infty}^{\infty} \frac{J_n(kb)}{H_n^+(kb)} e^{in(\alpha_i - \alpha_0)}. \tag{5.1}$$

It is easy to verify that the singular system $\{\psi_n, \xi_n, \sigma_n\}$ is, in this case, given by

$$\begin{aligned} \psi_n(\alpha_0) &= \frac{1}{\sqrt{2\pi}} e^{\pm in\alpha_0}, \quad \xi_n(\alpha_i) = \frac{e^{i\phi_n}}{\sqrt{2\pi}} e^{\pm in\alpha_i}, \\ \sigma_n &= \left| \frac{J_n(kb)}{H_j(kb)} \right|, \end{aligned} \tag{5.2}$$

where $\phi_n = \text{Arg}\,[J_n(kb)/H_n(kb)]$ and the plus sign gives one of the two singular vectors and the minus sign the second for each singular value σ_n.

For the density

$$g = \sum_n^\infty a_n \xi_n(\widehat{\eta})$$

we construct the incident Herglotz wave function

$$\begin{aligned} v_g^i(x) &:= \int_{\mathbb{S}} g(-\widehat{\eta}) u^i(x, -\widehat{\eta})\ ds(\widehat{\eta}) = \sum_n^\infty a_n v_n(x), \qquad \text{where} \\ v_n(x) &= \int_{\mathbb{S}} \xi_n(-\widehat{\eta}) u^i(x, -\widehat{\eta})\ ds(\widehat{\eta}). \end{aligned} \tag{5.3}$$

By (3.12), the corresponding far field is given by

$$v_g^\infty(\widehat{x}) := \int_{\mathbb{S}} g(-\widehat{\eta}) u^\infty(\widehat{x}, -\widehat{\eta})\ ds(\widehat{\eta}) = \sum_n^\infty a_n (\mathcal{F}\xi_n)(\widehat{x}) = \sum_n^\infty a_n \sigma_n \psi_n(\widehat{x}). \tag{5.4}$$

For this simple geometry, the incident and scattered fields have explicit formulations. The scattered field corresponding to v_g^i has the representation

$$v_g^s = \sum_n a_n \sigma_n v_n^s(x), \qquad \text{where} \qquad v_n^s(x) = \int_{\mathbb{S}} \xi_n(-\widehat{\eta}) u^s(x, -\widehat{\eta}) ds(\widehat{\eta}) \tag{5.5}$$

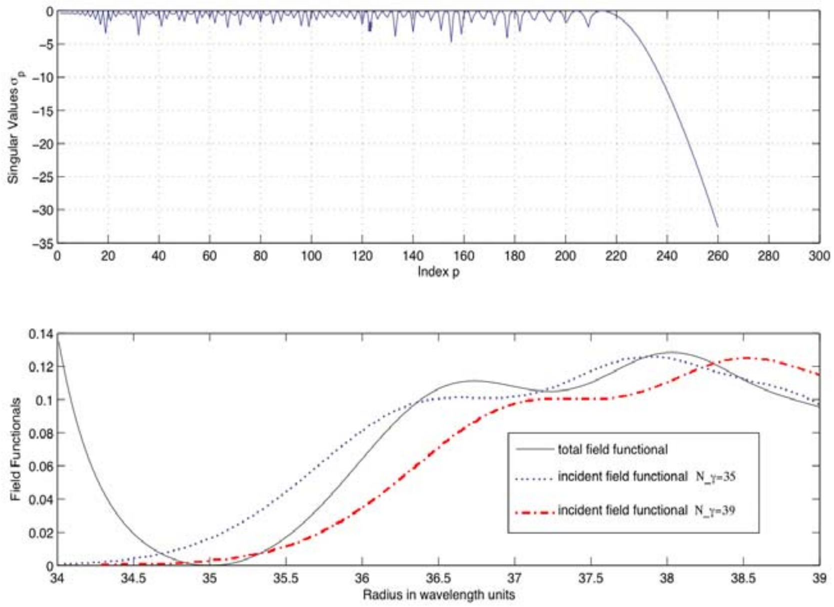


FIG. 2. *Top: Plot of the singular values $\sigma_n = |J_n(kb)/H_n(kb)|$ with $k = 1$ for a cylinder having radius $b = 35$. Bottom: Plots of the left-hand side of* (5.8) *and the right-hand side of this equation for $N_\gamma = 35$, for $\overline{N} = N_\gamma + 20$, and for $b = 35$. Also shown is a plot of the right-hand side of this equation for the case where $N_\gamma = 39$.*

and u^s is the scattered field corresponding to an incident plane wave. In cylindrical polar coordinates, $x = (r, \theta)$, this simplifies to

$$v_n^i(r,\theta) = \sqrt{2\pi} i^n J_n(kr) e^{\pm in\theta}, \tag{5.6}$$

$$v_n^s(r,\theta) = \sqrt{2\pi} e^{i\phi_n} i^n H_n^+(kr) e^{\pm in\theta}. \tag{5.7}$$

For details see [22].

For a Dirichlet obstacle, the total field is zero on the boundary, and, by Theorem 3.5, the incident field constructed from the finite collection of singular functions from N_γ to $\overline{N}$ is approximately zero:

$$\begin{aligned} &\lim_{r\to b} \sum_{n=N_\gamma}^{\overline{N}} \left| J_n(kr) - \frac{J_n(kb)}{H_n(kb)} H_n(kr) \right|^2 \\ &\approx \lim_{r\to b} \sum_{n=N_\gamma}^{\overline{N}} |J_n(kr)|^2 \approx 0 \end{aligned} \tag{5.8}$$

with $\overline{N} > N_\gamma$ and N_γ such that $\sigma_n < \gamma$, for all $n > N_\gamma$.

We present a plot of the singular values $\sigma_n = |J_n(kb)/H_n(kb)|$ using unit wavelength ($k = 2\pi$) and cylinder radius $b = 35\lambda = 35$ in the top of Figure 2. It is clear from this figure that the cutoff $N_\gamma \approx [kb] \approx 220$, where $[x]$ indicates the nearest integer approximation of x. In the bottom of the figure we show plots of the sums on

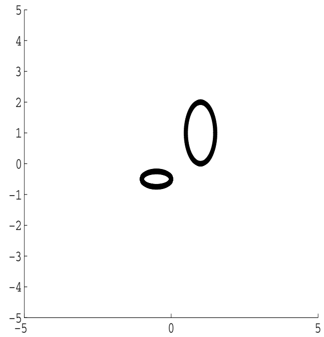


FIG. 3. *Sound-soft obstacles to be recovered.*

the left- and right-hand sides of (5.8) for the cases where $N_\gamma = [kb]$, $N_\gamma = [kb] + 4$, and $\overline{N} = N_\gamma + 20$. The boundary of the cylinder is identified by where the incident field amplitude falls below a chosen cutoff. If this cutoff is chosen to be .02, then one would estimate the radius of the cylinder to be about 35 for the case $N_\gamma = 35$, while one would estimate the radius to be 36 for the case $N_\gamma = 39$. We obtained similar results for other choices of the cylinder radius b. We observed in our experiments that the sharpness of the zero of the constructed incident field at the boundary depends on the cutoff N_γ. The closer N_γ is to the optimal cutoff, kb, the higher the contrast.

5.2. Two ellipses. Our second example is of two ellipses in $\mathbb{R}^2$ shown in Figure 3. We use potential theoretic techniques to calculate the far field pattern for an incident plane wave. We introduce the acoustic single- and double-layer operators given, respectively, by

$$(S\varphi)(x) := 2\int_{\partial\Omega} \varphi(y)\Phi(x,y)\; ds(y), \quad x \in \partial\Omega,$$

$$(K\varphi)(x) := 2\int_{\partial\Omega} \varphi(y)\frac{\partial\Phi(x,y)}{\partial\nu(y)}\; ds(y), \quad x \in \partial\Omega, \tag{5.9}$$

where Φ is the two-dimensional outgoing free-space fundamental solution to the Helmholtz equation, a zeroth-order Hankel function of the first kind. It can be shown [8] that, if the potential φ satisfies the integral equation

$$(I + K - \mathrm{i}S)\varphi(\cdot;\widehat{\eta}) = -u^i(\cdot;\widehat{\eta}), \tag{5.10}$$

then the scattered and far fields are given by

$$u^s(x,\widehat{\eta}) = \int_{\partial\Omega} \left(\frac{\partial\Phi(x,y)}{\partial\nu(y)} - \mathrm{i}\Phi(x,y)\right)\varphi(y;\widehat{\eta})\; ds(y), \quad x \in \Omega^o, \tag{5.11a}$$

$$u^\infty(\widehat{x};\widehat{\eta}) = \beta\int_{\partial\Omega} \left(\frac{\partial e^{-\mathrm{i}\kappa\widehat{x}\cdot y}}{\partial\nu(y)} - \mathrm{i}e^{-\mathrm{i}\kappa\widehat{x}\cdot y}\right)\varphi(y;\widehat{\eta})\; ds(y), \quad \widehat{x} \in \mathbb{S}. \tag{5.11b}$$

We do not use a sophisticated quadrature rule to resolve the point source on the boundary. This introduces a numerical error of about 10% which has the advantage of introducing noise into our calculations, albeit systematic noise.

For $x \in \mathbb{R}^2$ fixed, we calculate the nonscattering incident field v_g^i shown in Figures 4–6 as

$$v_g^i(x) = \sum_{n=N_\gamma}^{\overline{N}} |v_n^i(x)|, \tag{5.12}$$

where v_n^i is given by (5.3). The corresponding scattered field is $v_g^s = \sum_{n=N_\gamma}^{\overline{N}} \sigma_n v_n^s(x)$ for v_n^s given by (5.5) with u^s given by (5.11a).

To illustrate the resolution limits we sample the far field at $\overline{N} = 16$, 32, 64, and 128 points uniformly distributed over $\mathbb{S}$ with 16, 32, 64, and 128 incident field directions, respectively, also uniformly distributed over $\mathbb{S}$. The wavenumber is $k = 3$, and the smaller ellipse has minor axis of radius 0.25. By the resolution analysis of section 4.2, our smallest physical feature is 1/6 the illuminating wavelength, which according to (4.12) suggests that we need to sample the far field at $N \geq 2\pi\sqrt{2}(6) \approx 53$ points with more than 53 incident directions. This is more than is actually required, as an examination of the singular values of the far field operator shows. In Figure 4 we show the magnitude of the constructed incident field for $\overline{N} = 32$ through 128 with $N_\gamma = \overline{N} - 12$. This value of N_γ was chosen based on the decay of the singular values shown in the left column of Figure 4. The singular values decay rapidly after the 7th singular value as predicted by the eventual exponential decay of the singular values of the far field operator. They flatten out, however, beyond the 20th singular value because of the error, or noise, in our calculation of the far field pattern. In other words, the 20th and higher singular vectors of the far field operator appear to be in the noise subspace. For a sampling rate of 32, we have $N_\gamma = \overline{N} - 12 = 20$, and our constructed incident field then consists of all available singular vectors in the noise subspace. For higher sampling rates of 64 and 128 there is not a significant difference in resolution when only the last 12 singular vectors are used.

The case $\overline{N} = 16$ shown in Figure 5 illustrates the reduction in resolution that results from constructing the incident field from singular vectors that are not in the noise subspace of $\mathcal{F}$. In Figure 5(b) we used the 12 smallest singular vectors to construct the incident field. As Figure 5(a) shows, most of these are still well within the signal subspace of the far field operator. With only the last 4 singular vectors we are able to achieve remarkably good results, as demonstrated in Figure 5(c).

To illustrate the relative robustness of the method with respect to the choice of the cutoff N_γ, so long as it is above the critical cutoff, in Figure 6 we show the constructed incident field with $N_\gamma = 25, 78$, and 124. The incident fields are not normalized in order to gauge the relative contrast between the images.

To verify that the constructed scattered field, v_g^s, is indeed small outside the scatterer, we show in Figure 7 the computed scattered field using (5.5). The constructed scattered field is $O(10^{-13})$ around the scatterer and decays to zero rapidly away from the scatterer. The corresponding incident field, in contrast, is at least $O(10^{-4})$ on the exterior of the scatterer. This demonstrates Corollary 3.2.

6. Conclusion. Our main results, Theorem 3.1, Corollary 3.2, and Theorem 3.5, show that there is a density $\widehat{g}$ that approaches, nontrivially, the null space of the far field operator corresponding to some fixed, smooth scatterers. A superposition of plane waves weighted by such a density is a nonscattering incident field for these scatterers. The density can be constructed from the singular functions of the far field operator and the nonscattering phenomenon understood as the orthogonality of the singular functions to the far field pattern of a point source with sources located on

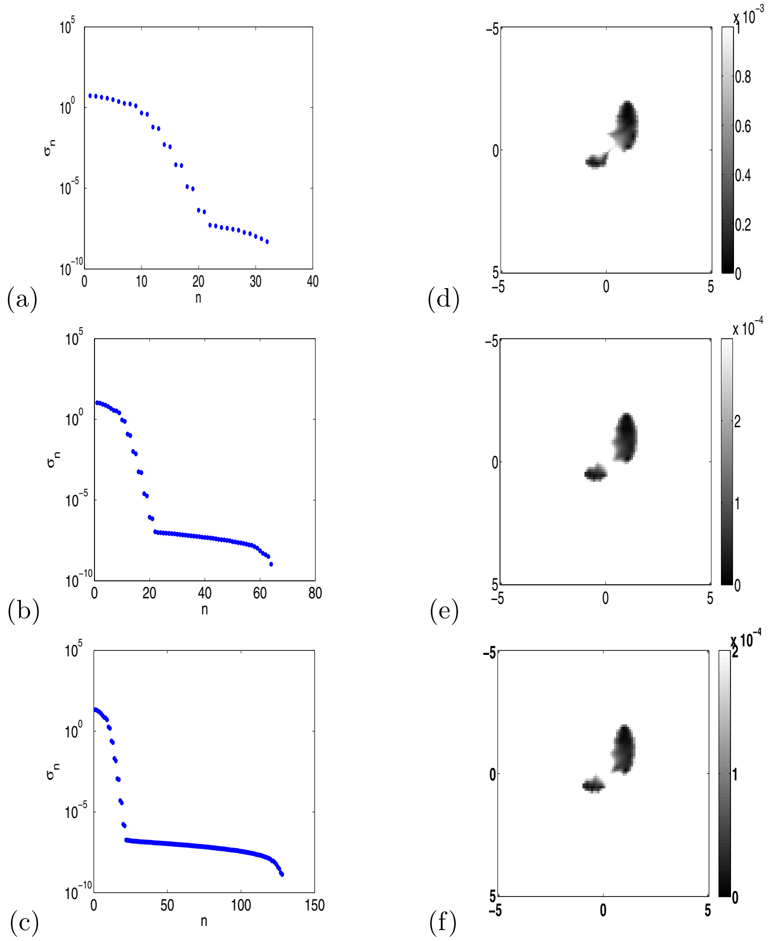


FIG. 4. (a)–(c) *Decay of the singular values of the multistatic response matrix for the far field sampled at* 32 (a), 64 (b), *and* 128 (c) *points for* 32, 64, *and* 128 *incident field directions, respectively, evenly distributed on* $\mathbb{S}$. *The far field pattern is calculated by* (5.11b) *to only about* 10% *accuracy, which introduces noise into the experiment reflected in the lower plateau of the singular values.* (d)–(f) *The magnitude of the corresponding incident field* $|v^i|$ *calculated by* (5.12) *and* (5.3) *for the far field sampled at* 32 (d), 64 (e), *and* 128 (f) *points for* 32, 64, *and* 128 *incident field directions, respectively, evenly distributed on* $\mathbb{S}$. *The cutoff for each of these examples is* $N_\gamma = \overline{N} - 12$*, where* $\overline{N} = 32, 64$*, and* 128*, respectively.*

the boundary of the scatterer. Our statement of Theorem 3.1 also raises unanswered questions about the rate of blowup of the densities in the linear sampling method.

The point source method of Potthast [24, 25] rests on the approximation of the scattered field u^s by computing the correct density for the construction of a *backpropagation* operator (2.16). As already noted, constructing such a density is a nontrivial task since this requires some knowledge of the boundary of the scatterer which we assume is unknown. The linear sampling method approaches the problem of finding the shape and location of the scatterer by looking for points where the fundamental solution far field pattern is not in the range of the far field operator, but still, one must solve an ill-posed linear integral equation at each point in some computational domain. One of the disadvantages of the linear sampling methodology, however, is that, since it is not constructive, it provides very little information about numerical

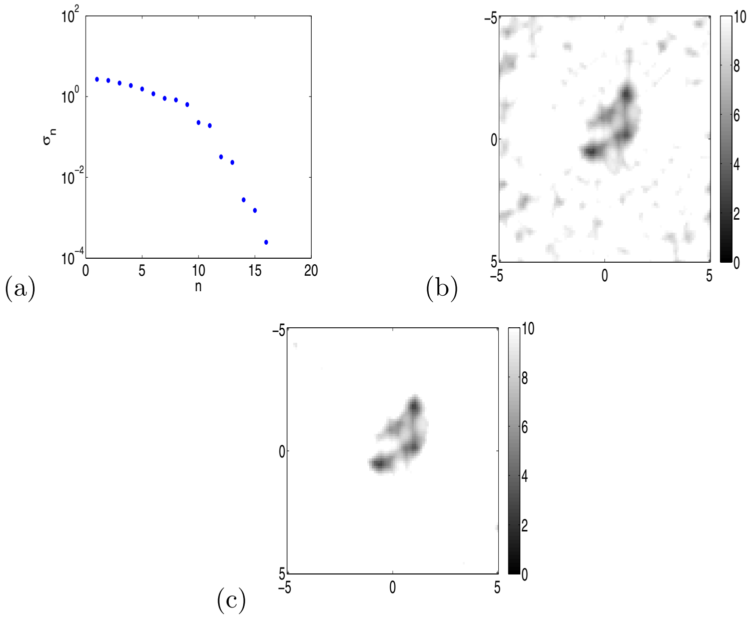


Fig. 5. (a) *Decay of the singular values of the multistatic response matrix for the far field sampled at* 16 *points for* 16 *incident field directions evenly distributed on* $\mathbb{S}$. (b) *The magnitude of the incident field* $|v^i|$ *calculated by* (5.12) *and* (5.3) *with cutoff* $N_\gamma = 4$, $\overline{N} = 16$. (c) *Incident field with* $N_\gamma = 12$ *and* $\overline{N} = 16$.

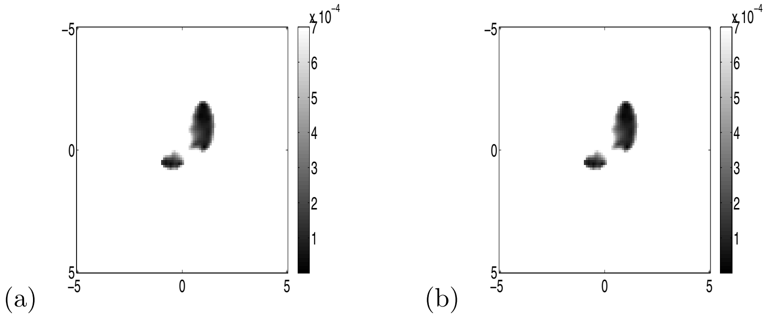


Fig. 6. *The magnitude of the incident field* $|v^i|$ *calculated by* (5.12) *and* (5.3) *for the far field sampled at* 128 *points with* 128 *incident field directions evenly distributed on* $\mathbb{S}$. *For each of these examples* $\overline{N} = 128$ *with cutoff* $N_\gamma = 104$ (a) *and* $N_\gamma = 20$ (b).

algorithms. Any numerical implementation will involve some sort of regularization strategy. The actual behavior of regularized solutions, or indeed any indication that a particular regularization strategy will deliver the desired behavior, remains open with the exception of the analysis of [3].

Our numerical experiments indicate, however, that particular details about implementing the linear sampling or point source methods are somewhat beside the point: it is not necessary to create an approximate domain as with the point source method, nor is it necessary to solve many ill-posed linear integral equations as in the linear sampling method. We need only work with incident plane waves and the known singular functions of the far field operator. This remains to be proved. We believe that

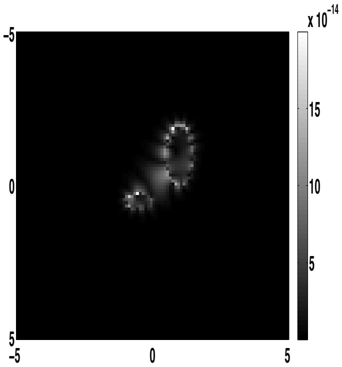


FIG. 7. *The magnitude of the scattered field calculated via* (5.5) *and* (4.3) *for the far field sampled at* 128 *points with* 128 *incident field directions evenly distributed on* $\mathbb{S}$*. Here* $\overline{N} = 128$*, and* $N_\gamma = 116$.

the answer lies with a closer examination of the connections between linear sampling and the factorization method as detailed in [3]. This is the subject of future research.

These results have intriguing implications for inverse scattering and signal design. The method works very much like the linear sampling method for inverse scattering in that the proposed incident field is constructed from the measured far field data and the scatterer is identified by those points in the domain where the incident field (and scattered field) is small. For signal design the method opens the door to the possibility of constructing signals that avoid certain known obstacles while irradiating others. Our application of the linear sampling method to the MUSIC algorithm is novel and clarifies the connections between many different inverse scattering approaches.

Acknowledgments. The authors gratefully acknowledge the support of the Institute for Mathematics and its Applications at the University of Minnesota where this work was started and the Mathematisches Forschungsinstitut Oberwolfach where the work was finished. We also thank David Colton and Fioralba Cakoni for many helpful discussions.

REFERENCES

[1] H. AMMARI, E. IAKOVLEVA, AND D. LESSELIER, *Two numerical methods for recovering small inclusions from the scattering amplitude at a fixed frequency*, SIAM J. Sci. Comput., 27 (2005), pp. 130–158.

[2] R. ARAMINI, M. BRIGNONE, AND M. PIANA, *The linear sampling method without sampling*, Inverse Problems, 22 (2006), pp. 2237–2254.

[3] T. ARENS, *Why linear sampling works*, Inverse Problems, 20 (2004), pp. 163–173.

[4] F. CAKONI AND D. COLTON, *Qualitative Methods in Inverse Scattering Theory*, Springer-Verlag, Berlin, 2006.

[5] M. CHENEY, *The linear sampling method and the MUSIC algorithm*, Inverse Problems, 17 (2001), pp. 591–595.

[6] D. COLTON, H. HADDAR, AND M. PIANA, *The linear sampling method in inverse electromagnetic scattering theory*, Inverse Problems, 19 (2003), pp. S105–S137.

[7] D. COLTON AND A. KIRSCH, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.

[8] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[9] D. COLTON AND R. KRESS, *On the denseness of Herglotz wave functions and electromagnetic Herglotz pairs in Sobolev spaces*, Math. Methods Appl. Sci., 24 (2001), pp. 1289–1303.

[10] D. Colton and D. Sleeman, *An approximation property of importance in inverse scattering theory*, Proc. Edinb. Math. Soc. (2), 44 (2001), pp. 449–454.

[11] A. J. Devaney, *Time reversal imaging of obscured targets from multistatic data*, IEEE Trans. Antennas and Propagation, 53 (2005), pp. 1600–1610.

[12] A. J. Devaney and E. Marengo, *Nonradiating sources with connections to the adjoint problem*, Phys. Rev. E, 70 (2004), 037601.

[13] A. J. Devaney, E. Marengo, and F. Gruber, *Time-reversal-based imaging and inverse scattering of multiply scattering point targets*, J. Acoust. Soc. Amer., 118 (2005), pp. 3129–3138.

[14] J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, New York, 1996.

[15] C. Hazard and K. Ramdani, *Selective acoustic focusing using time-harmonic reversal mirrors*, SIAM J. Appl. Math., 64 (2004), pp. 1057–1076.

[16] S. Hou, K. Solna, and H. Zhao, *A direct imaging algorithm for extended targets*, Inverse Problems, 22 (2006), pp. 1151–1178.

[17] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[18] A. Kirsch, *The MUSIC algorithm and the factorization method in inverse scattering theory for inhomogeneous media*, Inverse Problems, 18 (2001), pp. 1025–1040.

[19] R. Leis, *Initial-boundary value and scattering problems in mathematical physics*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, Springer-Verlag, Berlin, 1988, pp. 23–60.

[20] D. R. Luke, *Multifrequency inverse obstacle scattering: The point source method and generalized filtered backprojection*, Math. Comput. Simulation, 66 (2004), pp. 297–314.

[21] D. R. Luke, *Image synthesis for inverse obstacle scattering using the eigenfunction expansion theorem*, Computing, 75 (2005), pp. 181–196.

[22] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge University Press, Cambridge, UK, 1995.

[23] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[24] R. Potthast, *A fast new method to solve inverse scattering problems*, Inverse Problems, 12 (1996), pp. 731–742.

[25] R. Potthast, *A point-source method method for inverse acoustic and electromagnetic obstacle scattering problems*, IMA J. Appl. Math., 61 (1998), pp. 119–140.

[26] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Chapman & Hall, London, UK, 2001.

[27] C. Prada and M. Fink, *Eigenmodes of the time-reversal operator: A solution to selective focusing in multiple-target media*, Wave Motion, 20 (1994), pp. 151–163.

SIAM J. APPL. MATH.
Vol. 68, No. 1, pp. 292–310

CONVECTIVE STABILIZATION OF A LAPLACIAN MOVING BOUNDARY PROBLEM WITH KINETIC UNDERCOOLING[*]

UTE EBERT[†], BERNARD MEULENBROEK[‡], AND LOTHAR SCHÄFER[§]

Abstract. We study the shape stability of disks moving in an external Laplacian field in two dimensions. The problem is motivated by the motion of ionization fronts in streamer-type electric breakdown. It is mathematically equivalent to the motion of a small bubble in a Hele–Shaw cell with a regularization of kinetic undercooling type, namely, a mixed Dirichlet–Neumann boundary condition for the Laplacian field on the moving boundary. Using conformal mapping techniques, linear stability analysis of the uniformly translating disk is recast into a single PDE which is exactly solvable for certain values of the regularization parameter. We concentrate on the physically most interesting exactly solvable and nontrivial case. We show that the circular solutions are linearly stable against smooth initial perturbations. In the transformation of the PDE to its normal hyperbolic form, a semigroup of automorphisms of the unit disk plays a central role. It mediates the convection of perturbations to the back of the circle where they decay. Exponential convergence to the unperturbed circle occurs along a unique slow manifold as time $t \to \infty$. Smooth temporal eigenfunctions cannot be constructed, but excluding the far back part of the circle, a discrete set of eigenfunctions does span the function space of perturbations. We believe that the observed behavior of a convectively stabilized circle for a certain value of the regularization parameter is generic for other shapes and parameter values. Our analytical results are illustrated by figures of some typical solutions.

Key words. moving boundaries, kinetic undercooling, Laplacian growth, streamer discharges, convective stabilization

AMS subject classifications. 37L15, 37L25, 76D27, 80A22, 78A20

DOI. 10.1137/070683908

1. Introduction.

1.1. Problem formulation in physical and mathematical context.

The mathematical model considered in this paper is motivated by the physics of electric breakdown of simple gases like nitrogen or argon [1, 2, 3, 4, 5]. During the initial "streamer" phase of spark formation, a weakly ionized region extends in a strong externally applied electric field. As the ionized cloud is electrically conducting, it screens the electric field from its interior by forming a thin surface charge layer. This charged layer moves by electron drift within the local electric field and creates additional ionization, i.e., additional electron-ion pairs, by collisions of fast electrons with neutral molecules. We here approximate the ionized and hence conducting bulk of the streamer as equipotential. In the nonionized and hence electrically neutral region outside the streamer, the electric field obeys the Laplace equation. The thin surface charge layer can be approximated as an interface which moves according to

[*] Received by the editors June 19, 2006; accepted for publication (in revised form) July 16, 2007; published electronically November 7, 2007. Both the exact solution (4.4) and the slow manifold (4.15) were briefly presented in *Physical Review Letters* [14].

http://www.siam.org/journals/siap/68-1/68390.html

[†] CWI, MAS, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands, and Department of Applied Physics, Technical University Eindhoven, 5600 MB Eindhoven, The Netherlands (ebert@cwi.nl).

[‡] CWI, MAS, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. Current address: Department of Applied Mathematics, Delft University of Technology, 2600 AA Delft, The Netherlands (B.J.Meulenbroek@tudelft.nl). This author's research was supported by a Ph.D. position from CWI Amsterdam.

[§] Department of Theoretical Physics, Universität Duisburg–Essen, Lotharstr. 1, 47048 Duisburg, Germany (lsphy@theo-phys.uni-essen.de).

the electric field extrapolated from the neutral region onto the interface. We therefore are concerned with a typical moving boundary problem.

Such moving boundary problems occur in various branches of physics, chemistry, or biology. The most extensively studied examples are viscous fingering observed in two-fluid flows [6] or the Stefan problem of solidification from an undercooled melt [7]. Other physical phenomena like the motion of voids in current carrying metal films [8] lead to similar mathematical models [9].

We here discuss the streamer model in two spatial dimensions, where in the simplest "unregularized" version the basic equations coincide with those describing the motion of a small bubble in a liquid streaming through a Hele–Shaw cell [10, 11, 12, 13], which is a special case of two-fluid flow. The unregularized streamer model has been discussed in [4, 14]. Restriction to two dimensions in space allows us to use standard conformal mapping techniques [6, 15] to reduce the moving boundary problem to the analysis of the time dependence of the conformal map that maps the unit disk to the exterior of the streamer.

It is well known that unregularized moving boundary problems of this type are mathematically ill posed [15], in the sense that the moving interface generically develops cusps within finite time which leads to a breakdown of the model. To suppress such unphysical behavior, the models are regularized by imposing nontrivial boundary conditions on the interface. For viscous fingering typically some curvature correction to the interfacial energy is considered. For the streamer problem a mixed Dirichlet–Neumann boundary condition can be derived [14, 16] by analyzing the variation of the electric potential across the screening layer. Such a boundary condition is well known from the Stefan problem, where it is termed "kinetic undercooling." It rarely has been considered for Hele–Shaw-type problems. There are strong hints [15, 17, 18, 19] but no clear proof that it suppresses cusp formation. In particular, it has been shown that an initially smooth interface stays smooth for some finite time interval.

Here we consider the linear stability of uniformly translating circles in a Laplacian potential φ that approaches a constant slope $\varphi \propto x$ far from the circle; this means that the electric field $\mathbf{E} = -\nabla\varphi$ is constant far from the circle. Though this field breaks radial symmetry, uniformly translating circles are exact solutions of the regularized problem [14]. However, perturbations of these circles do not simply grow or decay locally as on a planar front or on circles in a radially symmetric force field [17, 18], but are also convected along the boundary; this convection turns out to be a determining part of the dynamics. Though physical streamers are elongated objects frequently connected to an electrode, the front part of a streamer is well approximated by a circular shape. Since it is this part that determines the dynamics, our analysis should be relevant also for more realistic shapes like fingers where no closed analytical solutions of the regularized uniformly translating shape are known [19]. In what follows we will use the term "streamer" to denote the translating circles, being aware that this is a slight abuse of the term.

1.2. Overview of content and structure of the paper. Regularization of the streamer model introduces some parameter ϵ that measures the effective width of the interface relative to the typical size of the ionized region. The regularized problem allows for a class of solutions of the form of uniformly translating circles, and linear stability analysis of these solutions can be reduced to solving a single PDE. For the special case $\epsilon = 1$, the general solution of this PDE can be found analytically, as we briefly discussed in [14]. The present paper is restricted to this special case as well.

The main results of the letter [14] are the following: The dynamics of infinitesimal perturbations is governed by a subgroup of the automorphisms of the unit disk. Generically, these automorphisms convect the perturbations to the back of the moving body. Initially, perturbations might grow, but they decay exponentially for $t \to \infty$. Furthermore, this final convergence back to the unperturbed circle follows some universal slow manifold.

The present paper contains a detailed derivation, discussion, and extension of the results presented in [14]. Furthermore, the analyticity and completeness of temporal eigenfunctions and the Fourier decomposition of perturbations are discussed, limit cases of the dynamics are worked out analytically, and results are demonstrated in a set of figures.

In detail, the time evolution determined by a PDE is often analyzed in terms of temporal eigenfunctions. For the present problem in a space of functions representing smooth initial perturbations of the moving circle, no such eigenfunctions exist. They can be constructed only if we allow for singularities on the boundary. We find here that a subset of these functions with time dependence $e^{-n\tau}$, $n \in \mathbf{N}_0$, is intimately related to the asymptotic convergence of the perturbations. These functions show singularities only at the backside of the circle, and the front part of any smooth perturbation can be expanded in this set of functions. The spatial domain of convergence of this expansion increases with time and, asymptotically for $t \to \infty$, it covers almost the whole streamer. In this restricted sense these eigenfunctions form a complete set.

These results dealing with infinitesimal perturbations, of course, do not imply the asymptotic stability of the circular shape against finite perturbations. To solve this problem, the full nonlinear theory must be considered. Nevertheless, a first hint might be gained by considering the evolution of a finite perturbation under the linearized dynamics. Due to the conformal mapping involved, the absence of cusps under this evolution is not a completely trivial question. We show here that for a large range of smooth initial conditions, the shape of the streamer stays smooth under the linearized dynamics.

All the present work deals with the exactly solvable case $\epsilon = 1$, whereas the physically most interesting case is $\epsilon \ll 1$. We believe, however, that the features we could identify explicitly for $\epsilon = 1$ are generic for all $\epsilon > 0$. In particular, the subgroup of automorphisms of the unit circle leads to the basic mechanism of convective stabilization, it is for all $\epsilon > 0$ intimately related to the characteristic curves of the PDE, and it also governs the dynamics in another exactly solvable case, namely, for $\epsilon = \infty$. Furthermore, it can be shown [20] that the temporal eigenvalues $\lambda_n(\epsilon)$ emerging from $\lambda_n(1) = -n$ stay negative for all $\epsilon > 0$, which also indicates that the circle might be asymptotically stable for arbitrary $\epsilon > 0$.

This paper is organized as follows. In section 2 we introduce the model, and the linear stability analysis of translating circles is carried through in section 3. These two sections are extended versions of [14]. Analytical results based on the PDE of linear stability analysis are derived in section 4, in particular, center of mass motion, internal motion, (non)analyticity and completeness of eigenfunctions, intermediate growth and asymptotic decay of perturbations, Fourier representation, and motion of nonanalytical points in the complex plane of the conformal map. These dynamic features are illustrated by explicit examples in section 5. The appendix contains a discussion of the case $\epsilon = \infty$.

2. Physical model and conformal mapping approach.

2.1. The model. We assume the ionized bulk of the streamer to be a compact, simply connected domain $\bar{\mathcal{D}}_i$ of the (x, y)-plane (see Figure 2.1). Outside the streamer,

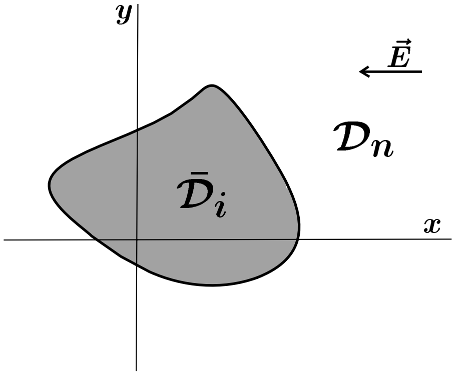


FIG. 2.1. *Geometry of the streamer model; $\vec{E}$ is the constant far field.*

i.e., in the open domain $\mathcal{D}_n$, there are no charges and the electric potential obeys the Laplace equation

$$\Delta\varphi = 0 \quad \text{for } (x, y) \in \mathcal{D}_n. \tag{2.1}$$

The streamer moves in an external electric field that becomes homogeneous far from the ionized body; therefore the electric potential φ at infinity obeys the boundary condition

$$\varphi \to E_0 x + \text{const} \quad \text{for } \sqrt{x^2 + y^2} \to \infty. \tag{2.2}$$

This condition excludes a contribution to φ diverging as $\ln(x^2 + y^2)$, which implies that the total charge due to the sum of all electrons and ions vanishes within $\bar{\mathcal{D}}_i$ and that the far field has the form

$$\vec{E} = -\nabla\varphi \to -E_0\hat{\mathbf{x}},$$

where $\hat{\mathbf{x}}$ is the unit vector in the x-direction. On the surface of the streamer we impose the boundary condition

$$\varphi = \ell\, \hat{\mathbf{n}} \cdot \nabla\varphi, \tag{2.3}$$

where $\hat{\mathbf{n}}$ is the unit vector normal to the surface pointing into $\mathcal{D}_n$. Here as well as in (2.4) below it is understood that the surface is approached from $\mathcal{D}_n$. As mentioned in the introduction, this boundary condition results from the analysis of the variation of the potential across the interface, and the length parameter ℓ can be interpreted as the effective thickness of the screening layer. The case $\ell = 0$ corresponds to the unregularized case with a pure Dirichlet condition on the moving boundary. Dynamics is introduced via the relation

$$v_n = \hat{n} \cdot \nabla\varphi, \tag{2.4}$$

which holds on the boundary and determines its normal velocity v_n. This defines our model. For further discussion of its physical background, we refer to [1, 2, 3, 4, 5, 16]. Now obviously, E_0 can be absorbed into a rescaling of the potential φ and of the time scale inherent in the velocity v_n; therefore we henceforth take $E_0 = 1$. Clearly the model defined here is most similar to a model of the motion of a small bubble in a Hele–Shaw cell [11, 12], except that the boundary condition (2.3) is of the form of a kinetic undercooling condition [17, 18].

2.2. Conformal mapping. A standard approach to such moving boundary problems proceeds by conformal mapping [6, 14]. We identify the (x, y)-plane with the closed complex plane $z = x + iy$, and we define a conformal map $f(\omega, t)$ that maps the unit disk $\mathcal{U}_\omega$ in the ω-plane to $\mathcal{D}_n$ in the z-plane, with $\omega = 0$ being mapped on $z = \infty$:

$$z = f(\omega, t) = \frac{a_{-1}(t)}{\omega} + \hat{f}(\omega, t), \quad a_{-1}(t) > 0. \tag{2.5}$$

Here the function $\hat{f}$ is holomorphic for $\omega \in \mathcal{U}_\omega$, and we assume that the derivatives ∂_ω^n of all orders n exist on the unit circle $\partial\mathcal{U}_\omega$. This restricts our analysis to smooth boundaries of the streamer. (Weaker assumptions on boundary behavior will be discussed briefly in section 4.8.) We recall that the closed physical boundary can now be retrieved as $x_\alpha(t) = \Re f(e^{i\alpha}, t)$ and $y_\alpha(t) = \Im f(e^{i\alpha}, t))$, where the interface parametrization with the real variable $\alpha \in [0, 2\pi[$ is fixed by the conformal map.

By virtue of (2.1), the potential φ restricted to $\mathcal{D}_n$ is a harmonic function; therefore it is the real part of some analytic function $\tilde{\Phi}(z, t)$, which under the conformal map (2.5) transforms into

$$\Phi(\omega, t) = \tilde{\Phi}\left(f(\omega, t)\right) = \frac{a_{-1}(t)}{\omega} + \hat{\Phi}(\omega, t). \tag{2.6}$$

Here the holomorphic function $\hat{\Phi}$ obeys the same conditions as $\hat{f}$ above. The pole results from the boundary condition (2.2) with $E_0 = 1$ and (2.5).

Conditions (2.3) and (2.4) take the form

$$|\omega\partial_\omega f|\, \Re[\Phi] = -\ell\, \Re\left[\omega\partial_\omega\Phi\right] \quad \text{for } \omega \in \partial\mathcal{U}_\omega, \tag{2.7}$$

$$\Re\left[\frac{\partial_t f}{\omega\partial_\omega f}\right] = \frac{\Re\left[\omega\partial_\omega\Phi\right]}{|\omega\partial_\omega f|^2} \quad \text{for } \omega \in \partial\mathcal{U}_\omega. \tag{2.8}$$

Equations (2.5)–(2.8) form the starting point of our analysis.

3. Linear stability analysis of translating circles.

3.1. Uniformly translating circles. A simple solution of (2.7), (2.8) takes the form

$$\left\{ \begin{aligned} f^{(0)}(\omega, t) &= \frac{R}{\omega} + \frac{2R}{R + \ell}\, t, \\ \Phi^{(0)}(\omega, t) &= R\left[\frac{1}{\omega} - \frac{R - \ell}{R + \ell}\,\omega\right]. \end{aligned} \right. \tag{3.1}$$

In physical coordinates x and y, it describes circles of radius $R > 0$ centered at $x(t) = v_0 t$ and moving with velocity $v_0 = 2R/(R + \ell)$ in direction $\hat{\mathbf{x}}$. Thus the point $\omega = 1$ maps to a point at the front, and the point $\omega = -1$ maps to a point at the back of the streamer. These points will play a crucial role in our analysis.

We note that the one-parameter family (3.1) of solutions parametrized by R, which is found in the regularized model, is a subset of the two-parameter family found in the unregularized case $\ell = 0$. As is well known, for $\ell = 0$ all ellipses with one axis parallel to $\hat{\mathbf{x}}$ are uniformly translating solutions [10].

3.2. Derivation of the operator $\mathcal{L}_\epsilon$ for linear stability analysis. We now derive the equation governing the evolution of infinitesimal perturbations of the circles

(3.1). In general, the parameter R can become time dependent. We use the ansatz

$$\left\{\begin{aligned} f(\omega,t) &= \frac{R(t)}{\omega} + x(t) + \eta\,\beta(\omega,t)\,, \\ \Phi(\omega,t) &= R(t)\left[\frac{1}{\omega} - \frac{R(t)-\ell}{R(t)+\ell}\,\omega + \eta\,\chi(\omega,t)\right], \\ \partial_t x(t) &= \frac{2R(t)}{R(t)+\ell}, \quad R(t)>0, \end{aligned}\right. \tag{3.2}$$

where β and χ are holomorphic functions of ω and where η is a small parameter. However, working to first order in η it is found that R stays constant. This results from the fact that the dynamics embodied in (2.8) strictly conserves the area $|\bar{\mathcal{D}}_i|$ of the streamer, which in this context is equivalent to the temporal conservation of the zero order Richardson moment [13, 15, 21], but integrated over the complement of $\mathcal{D}_n$. In terms of the mapping f, the conserved area $\left|\bar{\mathcal{D}}_i\right|$ can be written as

$$\begin{aligned} \left|\bar{\mathcal{D}}_i\right| &= \left|\int_0^{2\pi} d\alpha\left(\Re\left[f(e^{i\alpha},t)\right] - x(t)\right)\partial_\alpha \Im\left[f(e^{i\alpha},t)\right]\right| \\ &= \pi R^2(t) - \eta^2\int_0^{2\pi} d\alpha\,\Re\left[\beta(e^{i\alpha},t)\right]\partial_\alpha\Im\left[\beta(e^{i\alpha},t)\right]. \end{aligned} \tag{3.3}$$

Now introducing the time independent length R_0 through $\left|\bar{\mathcal{D}}_i\right| = \pi R_0^2$, we find $R(t) = R_0 + \mathcal{O}(\eta^2)$, which proves that R is time independent within linear perturbation theory. In what follows we will use R_0 as our length scale, introducing

$$\epsilon = \frac{\ell}{R_0} \quad \text{and} \quad \tau = \frac{2}{1+\epsilon}\frac{t}{R_0}, \tag{3.4}$$

and rescaling f and Φ by factors $1/R_0$. We note that within a dimensionless time interval τ of order unity, the streamer moves a distance of the order of its size.

With the thus simplified ansatz (3.2), equations (2.7) and (2.8) evaluated to first order in η take the form

$$\left\{\begin{aligned} &\Re\left[\omega(\partial_\omega - \partial_\tau)\beta - \frac{1+\epsilon}{2}\omega\partial_\omega\chi\right] = 0, \\ &\Re\left[\epsilon(\omega^2+1)\omega\partial_\omega\beta - (1+\epsilon)(1+\epsilon\omega\partial_\omega)\chi\right] = 0, \end{aligned}\right. \qquad \text{for } \omega \in \partial\mathcal{U}_\omega. \tag{3.5}$$

Since β and χ are holomorphic for $\omega \in \mathcal{U}_\omega$, these equations imply

$$\left\{\begin{aligned} &\omega(\partial_\omega - \partial_\tau)\beta - \frac{1+\epsilon}{2}\omega\partial_\omega\chi = 0, \\ &\epsilon(\omega^2+1)\omega\partial_\omega\beta - (1+\epsilon)(1+\epsilon\omega\partial_\omega)\chi = ia(t), \end{aligned}\right. \qquad \text{for } \omega \in \mathcal{U}_\omega, \tag{3.6}$$

where $a(t)$ is some real function of time. χ is eliminated by substituting the expressions for $\partial_\omega\chi$ and $\partial_\omega^2\chi$ from the first equation and its derivative into the second equation differentiated with respect to ω. This yields

$$\mathcal{L}_\epsilon\beta = 0, \tag{3.7}$$

where $\mathcal{L}_\epsilon$ is the operator

$$\mathcal{L}_\epsilon = \frac{\epsilon}{2}\,\partial_\omega\,(\omega^2-1)\omega\,\partial_\omega + \epsilon\,\omega\partial_\omega\partial_\tau + (1+\epsilon)\,\partial_\tau - \partial_\omega. \tag{3.8}$$

3.3. Normal form of $\mathcal{L}_\epsilon$ and induced automorphisms of the unit disk. It is instructive to transform $\mathcal{L}_\epsilon$ to the normal form of a hyperbolic differential operator. We introduce

$$T = \tanh \frac{\tau}{2}, \tag{3.9}$$

mapping the time interval $\tau \in [0, \infty[$ to $T \in [0, 1[$, and

$$\zeta = \frac{\omega + T}{1 + \omega T}, \tag{3.10}$$

to find

$$\mathcal{L}_\epsilon = \epsilon h(\zeta, T)\partial_T \partial_\zeta + \frac{\partial h(\zeta, T)}{\partial T} \partial_\zeta + (1 + \epsilon)\partial_T, \tag{3.11}$$

$$h(\zeta, T) = \frac{\omega}{\partial_\zeta \omega} = \frac{(\zeta - T)(1 - T\zeta)}{1 - T^2}. \tag{3.12}$$

This identifies the manifolds $T = \text{const}$ or $\zeta = \text{const}$ as the characteristic manifolds of our problem for all $\epsilon \neq 0$.

As function of the "time-like" parameter T, $0 \leq T < 1$, the transformation $\zeta = \zeta(\omega, T)$ in (3.10) represents a semigroup of automorphisms of the unit disk, with fixed points

$$\zeta = \omega = \pm 1.$$

For $T \to 1$, corresponding to $\tau \to \infty$, all points $\omega \neq -1$ are mapped into $\zeta = +1$, so that the large time behavior of any perturbation is governed by this attractive fixed point.

3.4. Analytical solutions of (3.7) for special values of ϵ. The general solution of (3.7) can be found analytically for the special values $\epsilon = 0$, $\epsilon = \pm 1$, and $\epsilon = \infty$. In the unregularized case $\epsilon = 0$, evidently any function

$$\beta(\omega, \tau) = \tilde{\beta}(\omega + \tau)$$

is a solution, and any singularity of $\tilde{\beta}$ found in the strip

$$0 < \Re[\omega] < \infty, \quad -1 \leq \Im[\omega] \leq 1,$$

will lead to a breakdown of perturbation theory within finite time. This is the fingerprint of the ill-posedness of the problem for $\epsilon = 0$.

For $\epsilon = -1$, $\beta(\omega, \tau)$ generically for all $\tau > 0$ has a logarithmic singularity at $\omega = -T(\tau)$. We recall that negative values of $\epsilon = \ell / R_0$ imply negative thickness of the screening layer and thus are of no physical interest.

The case $\epsilon = +1$ is discussed in detail in the remainder of the paper. Though a regularization length ℓ identical to the object size R_0 is somewhat artificial, it is accessible to rigorous analytical treatment and, as explained in section 1.2, we expect it to reveal generic features of the behavior for all $\epsilon > 0$.

This is supported by the results for $\epsilon = \infty$ which show essentially the same features as the results for $\epsilon = 1$ below. Though the limit $\epsilon \to \infty$ is physically absurd when applied to streamers, it is worth studying with respect to the properties of the operator $\mathcal{L}_\epsilon$, and we present a short discussion in the appendix.

4. Strong screening: Analytical results for $\epsilon = 1$.

4.1. Analytical solution of the general initial value problem. With the form (3.11) of $\mathcal{L}_\epsilon$, the PDE (3.7) for $\epsilon = 1$ reduces to

$$\partial_T \left(2 + h(\zeta, T)\partial_\zeta\right) \beta = 0, \tag{4.1}$$

showing that the function

$$G(\zeta) = \left(2 + h(\zeta, T)\partial_\zeta\right) \beta \tag{4.2}$$

is independent of T. To determine β, we use (3.12), $h(\zeta, T) = \omega/\partial_\zeta\omega$, to find

$$\left(2 + \omega\partial_\omega\right) \beta(\omega, \tau) = G(\zeta), \quad \zeta = \zeta(\omega, T(\tau)). \tag{4.3}$$

The solution regular at $\omega = 0$ takes the form

$$\beta(\omega, \tau) = \int_0^\omega \frac{x\,dx}{\omega^2}\, G\left(\frac{x + T(\tau)}{1 + xT(\tau)}\right). \tag{4.4}$$

A second independent solution is singular in $\omega = 0$:

$$\beta_{\text{sing}}(\omega, \tau) \equiv \frac{1}{\omega^2}. \tag{4.5}$$

The function G in the regular solution (4.4) is determined by the initial condition $\beta(\omega, 0)$ through

$$G(\omega) = \left(2 + \omega\partial_\omega\right) \beta(\omega, 0). \tag{4.6}$$

It thus is holomorphic for ω in the unit disk $\mathcal{U}_\omega$, and all derivatives exist on $\partial\mathcal{U}_\omega$, since we assume the initial surface to be smooth. Equation (4.4) then shows that $\beta(\omega, \tau)$ inherits these properties for all $\tau < \infty$.

4.2. Automorphism of the unit disk and a bound on the perturbation. It is now clear that the automorphisms $\zeta(\omega, T)$ of $\mathcal{U}_\omega$ from (3.12) contain the basic dynamics, and, as shown in the appendix, this also holds for $\epsilon = \infty$. This is to be contrasted to the unregularized case $\epsilon = 0$, where the dynamics amounts to a translation of the unit disk. With the present dynamics, in the course of time larger and larger parts $\mathcal{U}(\delta)$ of the unit disk $\mathcal{U}_\omega$ are mapped to an arbitrarily small neighborhood $|\zeta - 1| < \delta$ of the attractive fixed point $\zeta = 1$. According to (4.4) and (4.6), the initial condition in the neighborhood $|\omega - 1| < \delta$ then determines the evolution of $\beta(\omega, \tau)$ in all $\mathcal{U}(\delta)$. As a consequence, any pronounced structure found initially near ω_0, $|\omega_0 - 1| > \delta$, is convected towards $\omega = -1$. Quantitatively this behavior is embodied in (4.17) below, and explicit examples will be presented in section 5; see, in particular, Figure 5.4(b).

For the further discussion we normalize $G(\omega)$ so that

$$\max_{|\omega|=1} |G(\omega)| = 1. \tag{4.7}$$

Equations (4.4), (4.7) yield a bound on $\beta(\omega, \tau)$:

$$|\beta(\omega, \tau)| \leq \frac{1}{2}, \quad |\omega| \leq 1, \quad 0 \leq T \leq 1. \tag{4.8}$$

Thus the perturbation can shift the position of the streamer by at most $\eta/2$, and therefore it cannot affect the asymptotic velocity of the propagation.

4.3. Center of mass motion for $0 \le \tau < \infty$. In precise terms the position of the streamer can be defined as the center of mass

$$z_{\mathrm{cm}} = x_{\mathrm{cm}} + iy_{\mathrm{cm}} = \frac{1}{|\bar{\mathcal{D}}_i|} \int_{\mathcal{D}_i} dx\, dy\, (x + iy), \tag{4.9}$$

where the integral is related to the first order Richardson moment. Evaluating (4.9) and (4.4), we find to first order in η

$$z_{\mathrm{cm}} = \tau + \eta\, \beta(0, \tau), \tag{4.10}$$

$$\beta(0, \tau) = \frac{G(T(\tau))}{2}. \tag{4.11}$$

Here τ is the uniform translation of the unperturbed circle. The additional center of mass motion (4.11) for all times is explicitly given by the initial condition $\beta(\omega, 0)$ through (4.6) and the transformed time variable $T(\tau)$ from (3.9); for $\tau \to \infty$, it approaches $\beta(0, \tau) \to G(1)/2$.

4.4. Internal motion: Convergence along a universal slow manifold for $\tau \to \infty$. We now concentrate on the perturbation of the circular shape, given by

$$\tilde{\beta}(\omega, \tau) = \beta(\omega, \tau) - \beta(0, \tau). \tag{4.12}$$

The explicit expression

$$\tilde{\beta}(\omega, \tau) = \int_0^1 d\rho\, \rho \left[G\left(\frac{\rho\omega + T}{1 + \rho\omega T} \right) - G(T) \right] \tag{4.13}$$

yields

$$\lim_{\tau \to \infty} \tilde{\beta}(\omega, \tau) = 0 \tag{4.14}$$

for arbitrary G, i.e., for arbitrary initial condition (4.6). Thus the shape perturbation converges to zero as $\tau \to \infty$, and the circular shape is linearly stable.

We note that this holds despite the fact that the limits $\omega \to -1$ and $\tau \to \infty$ (i.e., $T \to 1$) do not commute:

$$\lim_{T \to 1} \lim_{\omega \to -1} G(\zeta(\omega, T)) = G(-1),$$
$$\lim_{\omega \to -1} \lim_{T \to 1} G(\zeta(\omega, T)) = G(+1).$$

This peculiar behavior near the backside of the streamer, at $\omega = -1$, shows up only in the rate of convergence.

Investigating the rate of convergence for $\tau \to \infty$, we first exclude a neighborhood of $\omega = -1$ and expand G in the integral (4.13) as

$$G\left(\frac{\rho\omega + T}{1 + \rho\omega T} \right) = G(T) + (1 - T^2)\, \frac{\rho\omega}{1 + \rho\omega T}\, G'(T) + \mathcal{O}(1 - T^2)^2,$$
$$\text{where } G'(\omega) = \partial_\omega G(\omega).$$

With

$$1 - T^2 = 4e^{-\tau} + \mathcal{O}(e^{-2\tau}),$$

the integral yields

$$\frac{\tilde{\beta}(\omega,\tau)}{G'(1)} = \frac{4}{\omega^2}\left[\ln(1+\omega) - \omega + \frac{\omega^2}{2}\right]e^{-\tau} + \mathcal{O}(e^{-2\tau}), \tag{4.15}$$

valid for

$$|1+\omega| \gg |\omega| e^{-\tau}.$$

Thus outside the immediate neighborhood of $\omega = -1$, the shape for all smooth initial conditions with $G'(1) \neq 0$ converges exponentially in time as $e^{-\tau}$ along a universal path in function space, given in (4.15). For $G'(1) = 0$ the first nonvanishing term in the expansion of G dominates the convergence.

To analyze the neighborhood of $\omega = -1$ we take the limit $\tau \to \infty$, with

$$s = (1+\omega)e^{\tau} \tag{4.16}$$

fixed. We find

$$\begin{aligned}\frac{\tilde{\beta}(\omega,\tau)}{G'(1)} &= 4\left(\ln(2+s) - \tau\right)e^{-\tau} \\ &\quad + \Bigg\{ 2G'(1) + 4\ln\left(\frac{2+s}{4}\right)\left(G'\left(\frac{s-2}{s+2}\right) - G'(1)\right) \\ &\quad + (2+s)\left(G(1) - G\left(\frac{s-2}{s+2}\right)\right) - 4\int_0^{4/(2+s)} dy \ln y \, G''(1-y)\Bigg\} \frac{e^{-\tau}}{G'(1)} \\ &\quad + \mathcal{O}\left(\tau e^{-2\tau}\right).\end{aligned} \tag{4.17}$$

In terms of ω, the first contribution on the right-hand side takes the form

$$4\left(\ln(2+s) - \tau\right)e^{-\tau} = 4e^{-\tau}\ln\left(2e^{-\tau} + 1 + \omega\right),$$

which shows that a logarithmic cut of $\tilde{\beta}(\omega,\tau)$ reaches $\omega = -1$ for $\tau \to \infty$, but with a prefactor vanishing exponentially in that limit. We thus have found a weak anomaly of the asymptotic relaxation near $\omega = -1$: In a spatial neighborhood of order $e^{-\tau}$ the exponential relaxation is modified by a factor τ. Furthermore, as mentioned above, all the initial structure of $\beta(\omega, 0)$ is compressed into that region. This is obvious from the occurrence of $G\left(\frac{s-2}{s+2}\right)$ etc. in (4.17).

To summarize, we have found that the shape of the interface for $\tau \to \infty$ converges to the circle along a universal slow manifold (4.15), except for a weak anomaly (4.17) at the backside at $\omega = -1$.

4.5. (Non)analyticity of temporal eigenfunctions. In many cases, a full dynamical solution for arbitrary initial values cannot be found, and rather temporal eigenfunctions are searched for. However, in the present problem, functions $\beta(\omega,\tau)$ resulting from smooth initial conditions cannot exhibit exponential behavior in time for all τ, $0 \leq \tau < \infty$. This is seen easily by introducing

$$G(x) = \hat{G}\left(\frac{x-1}{x+1}\right), \tag{4.18}$$

writing $G(\zeta)$ in the equivialent form

$$G\left(\frac{\omega+T}{1+\omega T}\right)=\hat{G}\left(\frac{\omega-1}{\omega+1}\,e^{-\tau}\right), \tag{4.19}$$

and substituting this form into (4.4). Postulating strict exponential time behavior $\beta\sim e^{-\lambda\tau}$, one finds

$$\beta(\omega,\tau)\propto e^{\lambda\tau}\,\beta_\lambda(\omega),\quad \beta_\lambda(\omega)=\int_0^1 d\rho\,\rho\left(\frac{\omega\rho-1}{\omega\rho+1}\right)^\lambda. \tag{4.20}$$

Any eigenfunction $\beta_\lambda(\omega,0)$ with $\lambda\neq 0$ clearly is singular at $\omega=+1$, at $\omega=-1$, or at both points. It therefore conflicts with smooth initial conditions. On the other hand, omitting a neighborhood of $\omega=-1$, eigenfunctions exist for all $-\lambda\in\mathbf{N}_0$.

4.6. Completeness of the eigenfunctions near $\omega = 1$. In some neighborhood of $\omega=1$, we can even show that any regular solution $\beta(\omega,\tau)$ can be expanded in terms of the "eigenfunctions" $\beta_{-n}(\omega)$, $n\in\mathbf{N}_0$. This results from the Taylor expansion

$$\hat{G}(y)=\sum_{n=0}^{\infty}\hat{g}_n y^n, \tag{4.21}$$

which by assumption converges in a disk of radius $\hat{r}>0$. Rewriting (4.4) as

$$\begin{aligned}\beta(\omega,\tau)&=\int_0^1\frac{x\,dx}{\omega^2}\,G\left(\frac{x+T}{1+xT}\right)-\int_\omega^1\frac{x\,dx}{\omega^2}\,G\left(\frac{x+T}{1+xT}\right)\\&=\frac{M(T)}{\omega^2}-\sum_{n=0}^{\infty}\hat{g}_n\,e^{-n\tau}\int_\omega^1\frac{x\,dx}{\omega^2}\left(\frac{1-x}{1+x}\right)^n\end{aligned} \tag{4.22}$$

and $\beta_{-n}(\omega)$ in a similar form as

$$\beta_{-n}(\omega)=\frac{M_n}{\omega^2}-\int_\omega^1\frac{x\,dx}{\omega^2}\left(\frac{1-x}{1+x}\right)^n, \tag{4.23}$$

we find

$$\beta(\omega,\tau)=\frac{M(T)}{\omega^2}+\sum_{n=0}^{\infty}\hat{g}_n\left[\beta_{-n}(\omega)-\frac{M_n}{\omega^2}\right]e^{-n\tau}. \tag{4.24}$$

Provided that $e^{-\tau}<\hat{r}$, we can separate the sum into the contribution $\propto 1/\omega^2$ and the rest. Since both $\beta(\omega,\tau)$ and $\beta_{-n}(\omega)$ are regular at $\omega=0$, the contributions $\propto 1/\omega^2$ have to cancel, which yields the final result

$$\beta(\omega,\tau)=\sum_{n=0}^{\infty}\hat{g}_n\,\beta_{-n}(\omega)\,e^{-n\tau}. \tag{4.25}$$

This result is valid for $e^{-\tau}<\hat{r}$ in the disk

$$\left|\frac{1-\omega}{1+\omega}\right|e^{-\tau}<\hat{r}.$$

It generalizes the asymptotic result (4.15). Indeed, the universal shape relaxation found in (4.15) together with the center of mass relaxation (4.11) precisely follows the slowest eigenfunction from (4.21) with $\lambda=-n=-1$. Furthermore this result shows that the range of validity of the expansion (4.25) increases with τ and asymptotically covers the whole complex plane except for the special point $\omega=-1$.

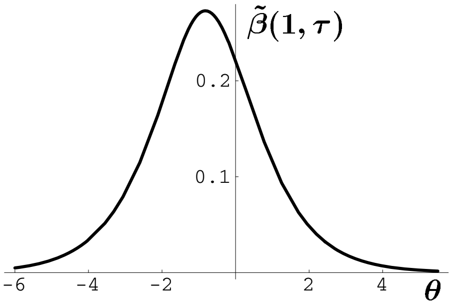


FIG. 4.1. *$\tilde{\beta}(e^{i\alpha},\tau)$ from (4.27) for $\alpha = 0$ as a function of subtracted time $\theta = \tau - \ln 2k$.*

4.7. Intermediate temporal growth and coupling of Fourier modes. Having found that the space of regular functions does not allow for strictly exponential time behavior, we now consider the typical time variation of smooth perturbations. Before the exponential relaxation sets in, such perturbations typically will grow, and this growth can be quite dramatic. As an illustration we consider a perturbation defined by

$$G(\omega) = \omega^k, \quad k \gg 1,$$

corresponding to initial conditions

$$\beta(\omega, 0) = \frac{\omega^k}{k+2}. \tag{4.26}$$

For $T = 1 - e^{-\theta}/k$, corresponding to times $\tau = \theta + \ln(2k) + \mathcal{O}(1/k)$, we can write

$$G\left(\frac{\omega + T}{1 + \omega T}\right) = \left(\frac{1 - \frac{e^{-\theta}}{1+\omega}\frac{1}{k}}{1 - \frac{\omega e^{-\theta}}{1+\omega}\frac{1}{k}}\right)^k = \exp\left[-e^{-\theta}\frac{1-\omega}{1+\omega}\right]\left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right),$$

where we again exclude some neighborhood of $\omega = -1$. Substituting this expression into (4.13), we find on the unit circle $\omega = e^{i\alpha}$

$$\begin{aligned}&\tilde{\beta}(e^{i\alpha}, \tau)\\ &= \int_0^1 d\rho\, \rho \exp\left[-e^{-\theta}\frac{1 - \rho^2 - 2i\rho\sin\alpha}{1 + \rho^2 + 2\rho\cos\alpha}\right] - \frac{1}{2}\exp\left[-e^{-\theta}\right] + \mathcal{O}\left(\frac{1}{k}\right).\end{aligned} \tag{4.27}$$

Figure 4.1 shows this function, evaluated at $\alpha = 0$ ($\omega = 1$). The behavior is quite peculiar. Up to times of order $\ln k$ the perturbation stays of order $1/k \ll 1$, then it increases roughly exponentially up to values of order 1, and finally it decreases again exponentially, approaching the slow manifold (4.15). Thus for very large k the initial perturbation $\beta(\omega, 0) \sim 1/k$ in some time interval can be amplified by a factor of order k, and (4.27) shows that the leading behavior in that time interval is independent of k.

Closer analysis shows that in terms of a formal Fourier expansion

$$\tilde{\beta}(e^{i\alpha}, \tau) = \sum_{n=1}^{\infty} a_n(\tau) e^{in\alpha}, \tag{4.28}$$

the amplification is carried by the low modes, $n = \mathcal{O}(1)$. As will be illustrated by an explicit example below (cf. Figure 5.2(b)), in such a mode representation the time evolution feeds the strength of the perturbation successively into lower and lower modes. This is equivalent to the observation that the automorphism $e^{i\alpha} \to \zeta(e^{i\alpha}, T)$ drives all the perturbative structure towards $\alpha = \pi$ and smooths the remainder of the interface. Note, however, that, starting with a perturbation $\sim \omega^k$, in the course of time modes $n > k$ are also (weakly) populated to build up a complicated structure near $\omega = -1$. We recall that for the unregularized model $\epsilon = 0$, the time evolution of a perturbation $\propto \omega^k$ populates only modes $k \leq n$ [4].

4.8. Motion of the zeros of $\partial_\omega f$ and cusps. So far we have shown that the propagating circle is linearly stable; i.e., we implicitly considered perturbations of infinitesimal strength η. The full nonlinear evolution of a finite perturbation is beyond the scope of this paper. Still, it clearly is a question of practical interest, whether a finite perturbation evolving under the linearized dynamics for all times satisfies the assumptions underlying the conformal mapping approach. For the mapping to stay conformal, all the zeros of $\partial_\omega f(\omega, \tau)$ must stay outside the unit circle. Thus, here we analyze the roots of the equation

$$0 = \partial_\omega f(\omega, \tau) = -\frac{1}{\omega^2} + \eta\, \partial_\omega \beta(\omega, \tau). \tag{4.29}$$

Using (4.3), (4.4), we can rewrite this equation as

$$2\eta \int_0^1 d\rho\, \rho \left[G\left(\frac{\omega + T}{1 + \omega T} \right) - G\left(\frac{\rho\omega + T}{1 + \rho\omega T} \right) \right] = \frac{1}{\omega}. \tag{4.30}$$

With our normalization (4.7) of G, for all ω in the closed unit disk the left-hand side of this equation is bounded by $2|\eta|$. We conclude that the bound

$$|\eta| < \frac{1}{2} \tag{4.31}$$

guarantees that within the framework of first order perturbation theory the mapping stays conformal for all times. We now will show that in general this bound cannot be improved.

For $\tau \to \infty$, zeros of $\partial_\omega f(\omega, \tau)$ reach $\omega = -1$, which is a consequence of the fact that in this limit an infinitesimally small neighborhood of $\omega = -1$ under the mapping $\omega \to \zeta$ is mapped essentially on the whole complex plane. We now analyze this limit for the simple example $G(\omega) = \omega$. Substituting this form into the asymptotic behavior (4.17) and using the definition (4.16) of s, we find

$$\partial_\omega \beta = \frac{4}{2 + s} + \mathcal{O}(\tau e^{-\tau}).$$

Equation (4.29) reduces to $s = 4\eta - 2$, showing that a zero ω_0 of $\partial_\omega f(\omega, \tau)$ approaches $\omega = -1$ as

$$\omega_0 = -1 + (4\eta - 2) e^{-\tau}.$$

For ω_0 to come from outside the unit circle we clearly must have

$$\Re[\eta] < \frac{1}{2}. \tag{4.32}$$

To get some feeling for the estimate (4.31), we note that for $G(\omega) = \omega^k$ the map initially (for $\tau = 0$) is conformal provided that $|\eta| < 1 + 2/k$. We conclude that under the linearized dynamics a large part of smooth initial conditions relaxes to the circle.

Throughout this section we have assumed the initial boundary to be smooth, so that all derivatives $\partial_\omega^n G(\omega)$ exist on the boundary $|\omega| = 1$. Inspecting the results, it is obvious that this assumption can be considerably relaxed, since only those derivatives which show up explicitly have to exist. Thus, for exponential relaxation (4.15) outside the neighborhood of $\omega = -1$ to prevail, the existence of $\partial_\omega G(\omega)$ is sufficient, which amounts to the condition that the curvature of the initial boundary is well defined. For the circle to be linearly stable, as in (4.14), it is sufficient that $G(e^{i\alpha})$ is bounded and continuous, which implies that the boundary has a well-defined slope.

If the initial boundary shows a cusp, the time evolution sensitively depends on the details. If the cusp is found in the forward direction, so that $G(\omega)$ diverges for $\omega \to 1$, the streamer will be strongly accelerated. In a related model [12], such an effect has been pointed out before. Furthermore, the shape will not relax to a circle, and the conformal map will presumably break down at finite time. If the cusp does not affect the analyticity of $G(\omega)$ near $\omega = 1$, it is convected towards the back and broadened, whereas the front of the streamer approaches the circular shape. Still, however, conformality of the map may break down at finite time.

5. Explicit examples for $\epsilon = 1$.

We here illustrate the general results by some examples.

5.1. The evolution of Fourier perturbations.

We first consider perturbations of the form

$$\beta^{[k]}(\omega, 0) = \frac{\omega^k}{k+2}, \quad \text{i.e., } G(\omega) = \omega^k. \tag{5.1}$$

The integral (4.4) is easily evaluated to yield

$$\begin{aligned}\beta^{[k]}(\omega,\tau) = \frac{1}{2\omega^2 T^2}\Bigg\{ & T^k + \left((T\omega)^2 - 1\right)\zeta^k \\ & + k\left(1 - T^2\right)\left[T^k - (\omega T + 1)\zeta^k + \frac{1 + k + T^2(1-k)}{T^k}\right. \\ & \left.\cdot \left(\ln(1+\omega T) - \sum_{\nu=1}^{k-1} \frac{T^\nu}{\nu}(\zeta^\nu - T^\nu)\right)\right]\Bigg\},\end{aligned} \tag{5.2}$$

where $T = T(\tau)$ and $\zeta = \zeta(\omega, T(\tau))$ are given by (3.9) or (3.10), respectively. In Figure 5.1 we have plotted snapshots of the resulting motion of the interface, determined as

$$z = x + iy = \frac{1}{\omega} + \tau + \eta\,\beta^{[k]}(\omega,\tau), \quad \omega = e^{i\alpha}, \quad 0 \le \alpha \le 2\pi. \tag{5.3}$$

The direction of motion, i.e., the positive x-direction, is downwards. Together with the moving interface, we show the unperturbed circular streamer at different times as gray disks with the center moving according to

$$z_{\mathrm{cm}}(\tau) = \tau + \frac{\eta}{2}\,G(T(\tau)) = \tau + \frac{\eta}{2}\tanh^k \frac{\tau}{2}, \tag{5.4}$$

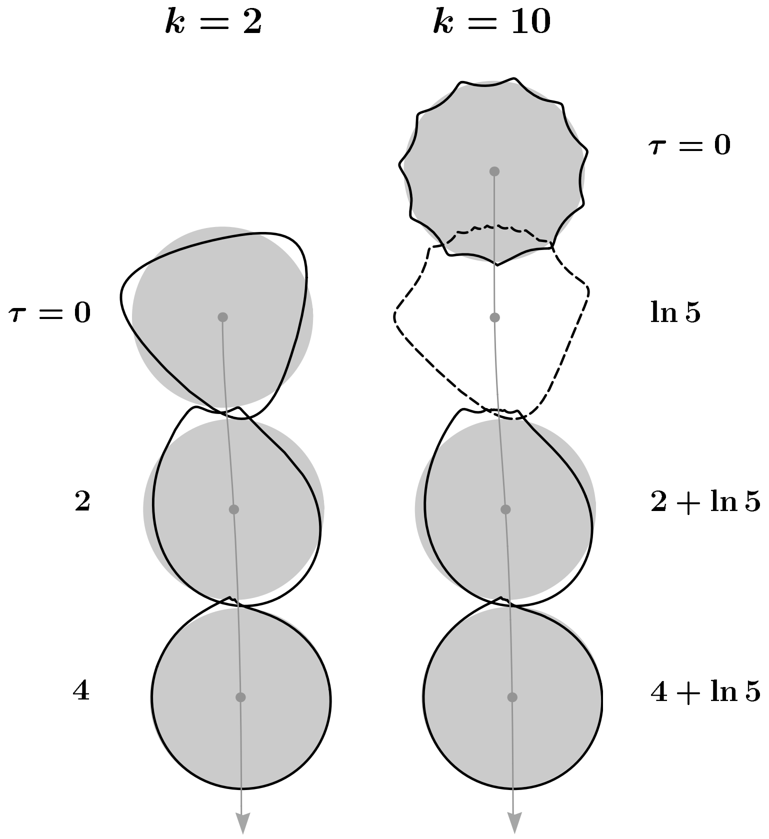


FIG. 5.1. *Snapshots of the evolution of the streamer for $k = 2$ (left column) and $k = 10$ (right column) at the indicated instants of time. The solid lines represent the perturbed interfaces. The gray disks move with the center of mass velocity* (5.4) *of the perturbed circles. One gray disk has been omitted for clarity. See the text for further discussion.*

as predicted for the center of mass motion for the perturbed streamer in (4.10).

In Figure 5.1 we perturbed the circle by $\eta\, \beta^{[k]}$, $k = 2$ or $k = 10$, using the same parameter $\eta = 0.6e^{i\pi/4}$ in both cases. The starting position for $k = 10$ is shifted relative to that for $k = 2$ by a distance corresponding to $\Delta\tau = \ln 5$. As discussed below (4.27), for $1 \ll k_1 < k_2$ we expect

$$\beta^{[k_1]}(\omega, \tau) \approx \beta^{[k_2]}(\omega, \tau + \ln(k_2/k_1)).$$

Figure 5.1 illustrates that such a "universality" for the gross structure holds down to very small k. (Of course the choice of differing values of η would distort the figures and mask this feature.) Basically during time evolution the initial maximum closest to the forward direction is smeared out and builds up the asymptotic circle, whereas all other structures are compressed at the backside. For $k = 10$ the complicated structure at the back is magnified in Figure 5.2(a). Figure 5.2(b) shows the time dependence of the coefficients a_n of the low modes $e^{in\alpha}$ in the expansion (4.28), again for $k = 10$. It illustrates how the strength of the perturbation cascades downwards in n and increases in time, until it is completely absorbed into the lowest mode, i.e., the overall shift of the circle. We should recall, however, that modes $n > k$ are also weakly populated to build up the structure at the back.

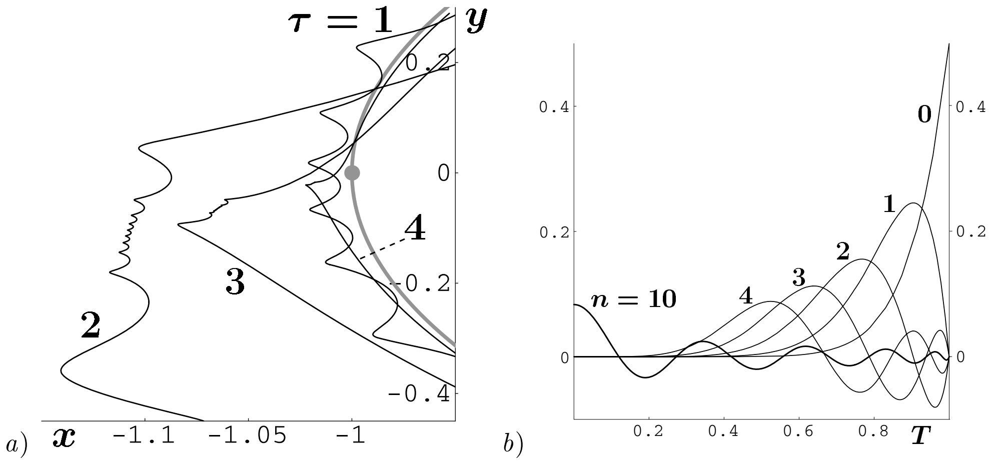


FIG. 5.2. (a) *Magnified plot of the backside of the streamer for $k = 10$, $\eta = 0.6\, e^{i\pi/4}$ (as in the right column of Figure* 5.1*) for the τ values given. The overall motion is subtracted. We observe the compression of the fine structure and the intermediate growth of the perturbation. Asymptotically for $\tau \to \infty$, the structure converges to the gray circle. In the comoving frame, the gray dot marks $x + iy = -1$, which is the point to which the structure finally is contracted. Note that the scale of x is stretched compared to that of y, and that the figure is turned relative to Figure* 5.1. (b) *The amplitudes a_n as in* (4.28) *as a function of T for $k = 10$; the values of n are given.*

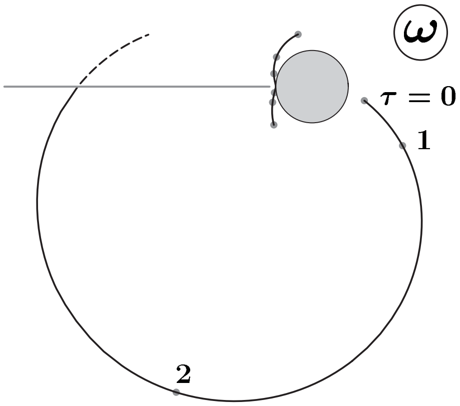


FIG. 5.3. *Motion of the zeros of $\partial_\omega f$ in the ω-plane for $k = 2$ and $\eta = 0.6\, e^{i\pi/4}$ (as in the left column of Figure* 5.1*). The dots give the position for $\tau = 0, 1, 2$. The horizontal line is the cut for $\tau = 2.51$, where one zero enters the second sheet (broken curve). The unit disk is also shown.*

For $k = 2$, Figure 5.3 shows the motion of the zeros of $\partial_\omega f(\omega, \tau)$ in the complex ω-plane, as discussed in section 4.8. It corresponds to the $k = 2$ part of Figure 5.1. Two zeros, which initially are close to the backside of the unit circle, approach $\omega = -1$ for $\tau \to \infty$. They clearly are associated with the two maxima that in the comoving frame are convected towards $z = x + iy = -1$. The third zero, originally found close to $\omega = +1$, after a large excursion leaves the physical sheet at time $\tau \simeq 2.51$. The logarithmic cut is on the negative axes, with the branchpoint $\omega_{\rm bp} = -1/T(\tau)$ reaching $\omega = -1$ for $\tau \to \infty$.

5.2. The evolution of localized perturbations. We finally consider some more localized perturbation, defined by

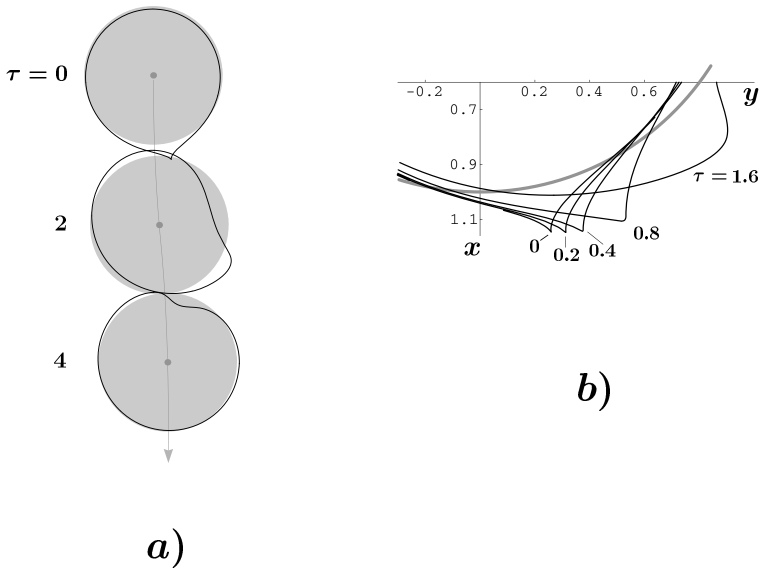


FIG. 5.4. (a) *Time evolution of a localized perturbation as described in the text.* (b) *Evolution of the initial peak for shorter times as indicated. The overall motion of the streamer is subtracted. A part of the asymptotic circle is shown in gray.*

$$G(\omega) = \frac{(1-\gamma)e^{i\alpha_0}}{\omega - \gamma e^{i\alpha_0}}, \quad \gamma > 1, \tag{5.5}$$

corresponding to an initial perturbation

$$\eta\,\beta(\omega,0) = \eta\,\frac{(1-\gamma)\gamma}{\omega^2}\,e^{2i\alpha_0}\left[\ln\left(1 - \frac{\omega}{\gamma}\,e^{-i\alpha_0}\right) - \frac{\omega}{\gamma}\,e^{-i\alpha_0}\right]. \tag{5.6}$$

The result for $\beta(\omega,\tau)$ reads

$$\beta(\omega,\tau) = \frac{(\gamma-1)e^{i\alpha_0}}{\gamma e^{-i\alpha_0} - T(\tau)}\left\{\frac{T(\tau)}{2b(\tau)} - \left(1 - \frac{T(\tau)}{b(\tau)}\right)\frac{\ln(1+b(\tau)\omega) - b(\tau)\omega}{(b(\tau)\omega)^2}\right\}, \tag{5.7}$$

where

$$b(\tau) = \frac{1 - T(\tau)\gamma e^{i\alpha_0}}{T(\tau) - \gamma e^{i\alpha_0}}. \tag{5.8}$$

We note that $b(\tau) \to 1$ for $T(\tau) \to 1$, so that in the large time limit the logarithmic cut reaches $\omega = -1$. As discussed in the context of (4.17), this is a generic feature of the present problem. Our choice of parameters ($\gamma = 1.1$, $\alpha_0 = -\pi/12$, $\eta = 1.5$) almost produces a cusp in the initial condition: The only zero of $\partial_\omega f(\omega, 0)$ is found at $\omega_0 = 1.001\exp(-.243i)$. This zero, however, is driven away from the unit circle and leaves the physical sheet. Another zero, which entered the physical sheet somewhat earlier, for $\tau \to \infty$ reaches $\omega = -1$. Figure 5.4(a) shows snapshots of the time evolution of the perturbed interface in a representation like Figure 5.1. It illustrates how the peak is rapidly smeared out and the interface becomes smooth. Figure 5.4(b) follows the evolution of the peak for short times and shows how it is convected and broadened.

We finally note that, in the special case where the initial peak strictly points in the forward direction ($\alpha_0 = 0$), convection cannot take place. The peak simply is broadened and vanishes, whereas some new peak shows up at the back for intermediate times.

Appendix A. The limit $\epsilon \to \infty$.

For $\epsilon \to \infty$, the PDE (3.7) with the form (3.11) of $\mathcal{L}_\epsilon$ reduces to

$$\Big(h(\zeta, T)\,\partial_\zeta + 1\Big)\,\partial_T\,\hat{\beta}(\zeta, T) = 0, \quad \text{where } \hat{\beta}(\zeta, T) \equiv \beta(\omega, \tau). \tag{A.1}$$

Equation (A.1) allows for a large set of solutions obeying the same initial condition

$$\beta(\omega, 0) = \beta_0(\omega), \tag{A.2}$$

but imposing regularity on the unit disk $\mathcal{U}_\omega$, we single out the simple form

$$\beta(\omega, \tau) = \beta_0(\zeta). \tag{A.3}$$

Thus for $\epsilon = \infty$, the dynamics is simply given by the automorphisms $\omega \longrightarrow \zeta(\omega, T)$. This implies that $\beta(\omega, \tau)$ is bounded uniformly in τ as

$$\big|\beta(\omega, \tau)\big| \le \max_{\omega \in \partial\mathcal{U}_\omega} \big|\beta_0(\omega)\big|, \tag{A.4}$$

so that in contrast to the case $\epsilon = 1$, there is no intermediate growth of the perturbations.

The shift of the center of mass is given by (cf. (4.10))

$$\beta(0, \tau) = \beta_0(T(\tau)) = \beta_0(1) - 2\,\beta_0'(1)\,e^{-\tau} + \mathcal{O}\left(e^{-2\tau}\right), \tag{A.5}$$

and except for the point $\omega = -1$, the shape again converges exponentially in time to the circle along the universal slow manifold

$$\beta(\omega, \tau) - \beta(0, \tau) = \beta_0'(1)\,\frac{4\,\omega}{1 + \omega}\,e^{-\tau} + \mathcal{O}\left(e^{-2\tau}\right); \tag{A.6}$$

cf. (4.15) for $\epsilon = 1$. Again the neighborhood of $\omega = 1$ for time $\tau = 0$, more precisely $\beta_0(1)$ and $\beta_0'(1)$, determines the long time convergence. Since by assumption $\beta_0(\omega)$ is analytical at $\omega = 1$, evidently an eigenfunction expansion in the sense of subsection 4.5 exists.

The only major difference compared to the case $\epsilon = 1$ concerns the point $\omega = -1$. Clearly,

$$\beta(-1, \tau) \equiv \beta_0(-1) \tag{A.7}$$

independently of τ, and indeed for $\tau \to \infty$ the conformality of the mapping breaks down in the neighborhood of $\omega = -1$ since $\partial_\omega \beta(\omega, \tau)\big|_{\omega=-1}$ diverges.

Acknowledgments. We acknowledge helpful and motivating discussions with F. Brau, A. Doelman, J. Hulshof, H. Levine, L. P. Kadanoff, S. Tanveer, and S. Thomae.

REFERENCES

[1] U. Ebert, W. van Saarloos, and C. Caroli, *Streamer propagation as a pattern formation problem: Planar fronts*, Phys. Rev. Lett., 77 (1996), pp. 4178–4181.

[2] M. Arrayás, U. Ebert, and W. Hundsdorfer, *Spontaneous branching of anode-directed streamers between planar electrodes*, Phys. Rev. Lett., 88 (2002), 174502.

[3] M. Arrayás and U. Ebert, *Stability of negative ionization fronts: Regularization by electric screening?*, Phys. Rev. E (3), 69 (2004), 036214.

[4] B. Meulenbroek, A. Rocco, and U. Ebert, *Streamer branching rationalized by conformal mapping techniques*, Phys. Rev. E (3), 69 (2004), 067402.

[5] U. Ebert, C. Montijn, T. M. P. Briels, W. Hundsdorfer, B. Meulenbroek, A. Rocco, and E. M. van Veldhuizen, *The multiscale nature of streamers*, Plasma Sources Sci. Technol., 15 (2006), pp. S118–S129.

[6] D. Bensimon, L. P. Kadanoff, S. Liang, B. I. Shraiman, and C. Tang, *Viscous flows in two dimensions*, Rev. Modern Phys., 58 (1986), pp. 977–999.

[7] L. I. Rubinstein, *The Stefan Problem*, Transl. Math. Monogr. 27, AMS, Providence, RI, 1971.

[8] P. S. Ho, *Motion of inclusion induced by a direct current and a temperature gradient*, J. Appl. Phys., 41 (1970), pp. 64–68.

[9] M. Mahadevan and R. M. Bradley, *Stability of a circular void in a passivated, current-carrying metal film*, J. Appl. Phys., 79 (1996), pp. 6840–6847.

[10] G. Taylor and P. G. Saffman, *A note on the motion of bubbles in a Hele-Shaw cell and porous medium*, Quart. J. Mech. Appl. Math., 12 (1959), pp. 265–279.

[11] S. Tanveer and P. G. Saffman, *Stability of bubbles in a Hele-Shaw cell*, Phys. Fluids, 30 (1987), pp. 2624–2635.

[12] D. C. Hong and F. Family, *Bubbles in the Hele-Shaw cell: Pattern selection and tip perturbations*, Phys. Rev. A (3), 38 (1988), pp. 5253–5259.

[13] V. M. Entov, P. I. Etingof, and D. Ya. Kleinbock, *Hele-Shaw flows with a free boundary produced by multipoles*, European J. Appl. Math., 4 (1993), pp. 97–120.

[14] B. Meulenbroek, U. Ebert, and L. Schäfer, *Regularization of moving boundaries in a Laplacian field by a mixed Dirichlet-Neumann boundary condition: Exact results*, Phys. Rev. Lett., 95 (2005), 195004.

[15] S. D. Howison, *Complex variable methods in Hele-Shaw moving boundary problems*, European J. Appl. Math., 3 (1992), pp. 209–224.

[16] F. Brau, A. Luque, B. Meulenbroek, U. Ebert, and L. Schäfer, *Construction and test of a moving boundary model for negative streamer discharges*, Phys. Rev. E (3), submitted; available online from http://arxiv.org/abs/0707.1402.

[17] Yu. E. Hohlov and M. Reissig, *On classical solvability for the Hele-Shaw moving boundary problems with kinetic undercooling regularization*, European J. Appl. Math., 6 (1995), pp. 421–439.

[18] M. Reissig, S. V. Rogosin, and F. Hübner, *Analytical and numerical treatment of a complex model for Hele-Shaw moving boundary value problems with kinetic undercooling regularization*, European J. Appl. Math., 10 (1999), pp. 561–579.

[19] S. J. Chapman and J. R. King, *The selection of Saffman-Taylor fingers by kinetic undercooling*, J. Engrg. Math., 46 (2003), pp. 1–32.

[20] S. Tanveer, F. Brau, U. Ebert, and L. Schäfer, in preparation.

[21] S. Richardson, *Hele Shaw flows with a free boundary produced by the injection of fluid into a narrow channel*, J. Fluid Mech., 56 (1972), pp. 609–618.

SIAM J. APPL. MATH.
Vol. 68, No. 2, pp. 311–333

INVERSE SCATTERING IN MULTIMODE STRUCTURES[*]

OLE HENRIK WAAGAARD[†] AND JOHANNES SKAAR[‡]

Abstract. We consider the inverse scattering problem associated with any number of interacting modes in one-dimensional structures. The coupling between the modes is contradirectional in addition to codirectional and may be distributed continuously or in discrete points. The local coupling as a function of position is obtained from reflection data using a layer-stripping-type method, and the separate identification of the contradirectional and codirectional coupling is obtained using matrix factorization. Ambiguities are discussed in detail, and different a priori information that can resolve the ambiguities is suggested. The method is exemplified by applications to multimode optical waveguides with quasi-periodical perturbations.

Key words. inverse scattering, layer-stripping, multimode structures

AMS subject classifications. 15A06, 15A23, 15A90, 78A45, 78A50

DOI. 10.1137/060672224

1. Introduction.

In waveguides that support several modes, scattering, or coupling between the different modes, may appear due to different kinds of perturbations. Possible perturbations are reflectors, gratings, bends, tapering, and other kinds of geometrical or material modulation along the waveguide. The coupling may be both codirectional (coupling between modes that propagate in the same direction) or contradirectional (coupling between modes that propagate in opposite directions). The direct scattering problem of computing the scattered field when the probing waves and the scattering structure are known has been extensively discussed in the literature [24, 38, 21]. The inverse scattering problem associated with two interacting modes is also well understood and has been treated in several contexts since the pioneering work by Gel'fand and Levitan [13], Boutet de Monvel and Marchenko [3], and Kreĭn [22]. In geophysics the so-called dynamic deconvolution or layer-stripping (layer-peeling) methods emerged for the identification of layered-earth models from acoustic scattering data [28, 31, 2, 6, 5]. More recently the inverse scattering methods have been applied to the design and characterization of optical devices involving two interacting modes. Both contradirectional coupling and codirectional coupling have been treated. Optical components based on contradirectional coupling include thin-film filters and fiber Bragg gratings [40, 10, 37, 30, 36, 34], while codirectional coupling is present in, e.g., grating-assisted codirectional couplers and long-period gratings [19, 39, 9, 42, 4]. While the inverse scattering problem associated with two interacting modes is well known, the inverse scattering problem of several, possibly nondegenerate modes (i.e., with different propagation constants) seems unsolved so far. Some work has been done in the case of four degenerate modes, that is, two polarization modes in each direction [35, 41], and several degenerate modes with only contradirectional coupling [1].

On the other hand, several methods for the inverse scattering of acoustic or electromagnetic waves in two or three dimensions have been reported. In particular,

[*] Received by the editors October 13, 2006; accepted for publication (in revised form) July 16, 2007; published electronically November 9, 2007.

http://www.siam.org/journals/siap/68-2/67222.html

[†] Optoplan AS, NO-7448 Trondheim, Norway (ole.henrik.waagaard@optoplan.com).

[‡] Department of Electronics and Telecommunications, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway (johannes.skaar@iet.ntnu.no).

Yagle et al. have developed layer-stripping methods for the multidimensional case [45, 43, 44]. By Fourier transforming the problem with respect to the transversal coordinates, the multidimensional problem may be regarded as one-dimensional with several interacting modes.

In this paper we will extend these lines of thought to cover the general inverse scattering problem associated with any number of interacting modes in one-dimensional, reciprocal structures. In the model (section 2) both codirectional and contradirectional coupling may be present simultaneously. We limit ourselves to the case where the known probing waves and the scattered waves propagate in opposite directions. In other words the scattered wave is considered as a reflection from the unknown structure. A layer-stripping inverse scattering algorithm is presented in section 3. Ambiguities related to the simultaneous presence of co- and contradirectional coupling are discussed in detail. Possible a priori information that can resolve these ambiguities will be suggested. The formalism is particularly useful for the quasi-periodical case (section 4), since only the slowly varying envelope needs to be represented rather than the structure itself, yielding an efficient algorithm. In section 5, the method is applied for the numerical reconstruction of a quasi-periodical waveguide structure. Sections 4 and 5 are exemplified by a multimode fiber Bragg grating: an optical fiber with quasi-periodic refractive index perturbation along the fiber axis, giving rise to both co- and contradirectional coupling. Finally, analogies to the multidimensional case are discussed in section 6.

2. Continuous and discrete coupling model. Consider a structure with P modes propagating in each direction along the x-axis. We visualize the x-axis as being directed to the right and say that the $+x$-direction is the forward direction. The propagation constant of the pth mode is $\pm\beta_p$; i.e., the x-dependence of the complex field associated with mode p is described by the factor $\exp(\pm i\beta_p x)$, where the upper (lower) sign applies to forward (backward) propagating modes. Note that the propagation constants of different modes may or may not be different. The propagation constants are related to frequency through the dispersion relation of the structure. The propagation constants may be expressed as $\beta_p = n_p\omega/c$, where ω is the angular frequency, c is some fixed reference velocity (common for all modes), and n_p accounts for the actual phase velocity. (However, in some cases it may rather be convenient to express the propagation constants in the form $n_p\omega/c - \pi/\Lambda$, where Λ is a constant; see section 4.) For electromagnetic waves, it is natural to set c equal to the vacuum velocity, and consequently we will refer to n_p as the effective index associated with mode p. In principle, the effective indices may be complex and dependent on frequency, meaning that modal loss and dispersion are permitted in the model. However, the dispersion must be limited by relativistic causality in the sense that any signal carried by the modes travels no faster than the vacuum light velocity. Also, the modal field profiles are assumed to have uniform phases such that they can be written real.

Coupling may occur due to a continuous or discrete scattering structure. In the first case, the field is assumed to be governed by the coupled-mode equation

$$\frac{\mathrm{d}\mathbf{E}}{\mathrm{d}x} = i\mathbf{C}\mathbf{E}, \tag{2.1}$$

where $\mathbf{E}$ is a column vector containing the $2P$ mode amplitudes. In the absence of the scattering structure ($\mathbf{C}_{\boldsymbol{\sigma}} = \mathbf{C}_{\boldsymbol{\kappa}} = 0$; see below), the first P elements are the mode amplitudes of the forward propagating modes (propagating in the $+x$-direction) and the last P elements are those of the backward propagating modes. The coupling matrix $\mathbf{C}$ can be decomposed into three contributions:

$$\mathbf{C} = \mathbf{D} + \mathbf{C}_{\boldsymbol{\sigma}} + \mathbf{C}_{\boldsymbol{\kappa}}. \tag{2.2}$$

The contributions can be expressed as 2×2 block matrices consisting of $P \times P$ blocks:

$$\mathbf{D} = \begin{bmatrix} \boldsymbol{\beta} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\beta} \end{bmatrix}, \tag{2.3a}$$

$$\mathbf{C}_{\boldsymbol{\kappa}} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\kappa} \\ -\boldsymbol{\kappa}^* & \mathbf{0} \end{bmatrix}, \tag{2.3b}$$

$$\mathbf{C}_{\boldsymbol{\sigma}} = \begin{bmatrix} \boldsymbol{\sigma} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\sigma}^* \end{bmatrix}, \tag{2.3c}$$

where * denotes complex conjugate. The first term $\mathbf{D}$ describes the frequency dependence due to the propagation of the different modes ("self-coupling") and is independent on x; and $\boldsymbol{\beta} = \mathrm{diag}\{\beta_1, \beta_2, \ldots, \beta_P\}$. Only this term is permitted to be lossy in the model ($\boldsymbol{\beta}$ may be complex). In practice, we should require $\mathrm{Im}\,\beta_p L \lesssim 1$, where L is the total length of the structure; otherwise the field at the far end of the structure may be close to zero (i.e., the mode will be bound at the left interface to the structure, and very little reflection will originate from the far end). The second term $\mathbf{C}_{\boldsymbol{\kappa}}$ describes the coupling between counterpropagating modes, whereas the last term $\mathbf{C}_{\boldsymbol{\sigma}}$ accounts for the coupling between copropagating modes. The coupling coefficients $\boldsymbol{\kappa}$ and $\boldsymbol{\sigma}$ are dependent on x but assumed independent on frequency. As will become clear shortly, the above forms of $\mathbf{C}_{\boldsymbol{\kappa}}$ and $\mathbf{C}_{\boldsymbol{\sigma}}$ are consequences of reciprocity and losslessness. It should be noted that in structures such as long-period gratings, where the coupling is purely codirectional, the coupling is described by $\boldsymbol{\kappa} = \mathbf{0}$ and a $\boldsymbol{\sigma}$ with nonzero off-diagonal elements. The conventional way of describing such structures would be to consider only the upper-left $P \times P$ block of $\mathbf{C}$. The layer-stripping method in section 3 cannot be used to reconstruct such structures since the reflection response is zero.

The coupling region in the waveguide is discretized into N layers, each of thickness $\Delta x = L/N$. If N is sufficiently large so that the matrices in (2.3) can be treated as constants in each layer, we can solve (2.1):

$$\mathbf{E}(x_j + \Delta x) = \exp{(i\mathbf{C}\Delta x)}\mathbf{E}(x_j), \quad x_j = j\Delta x. \tag{2.4}$$

This transfer matrix relation can be used to propagate the fields through the piecewise uniform structure. With the help of the connection between the transfer matrix and the scattering matrix (Appendix B) we can find the reflection and transmission response from the total transfer matrix, obtained as a product of the transfer matrices $\exp{(i\mathbf{C}\Delta x)}$ of each layer (direct scattering).

While direct scattering is achieved straightforwardly using the piecewise-uniform discretization, for inverse scattering it is convenient to push the discretization further, in order to identify the different contributions to the transfer matrix $\exp{(i\mathbf{C}\Delta x)}$. To first order in Δx, we have $\exp(i\mathbf{C}\Delta x) = \exp(i\mathbf{D}\Delta x)\exp(i\mathbf{C}_{\boldsymbol{\kappa}}\Delta x)\exp(i\mathbf{C}_{\boldsymbol{\sigma}}\Delta x)$. For a continuous structure of finite thickness, the bandwidth where the reflection spectrum is significantly different from zero is finite. Thus we need only be concerned with frequencies satisfying $|\omega| \leq \omega_{\mathrm{b}}$ for some positive constant ω_{b}. Note that this model may give entirely incorrect results for $|\omega| > \omega_{\mathrm{b}}$. For instance, if $P = 1$, the reflection spectrum calculated with the discrete model will be periodic with period

$\pi c/(n_1\Delta x)$, while the spectrum associated with a continuous structure tends to zero for large frequencies. For inverse scattering, the reflection spectrum and therefore ω_{b} are known. Therefore, provided Δx is chosen sufficiently small we can approximate each layer by a cascade of three sections: a section with codirectional coupling, a section with contradirectional coupling, and a time-delay section. The physical implication of this factorization is that the mode coupling appears in a discrete point within the layer rather than distributed along the whole layer. The contradirectional section may therefore be pictured as a discrete reflector. The transfer matrix of the jth layer becomes

$$\mathbf{T}_j = \mathbf{T}_{\boldsymbol{Z}}\mathbf{T}_{\boldsymbol{\rho}_j}\mathbf{T}_{\boldsymbol{\Phi}_j}, \tag{2.5}$$

where

(2.6a)

$$\mathbf{T}_{\boldsymbol{Z}} \equiv \exp(i\mathbf{D}\Delta x) = \begin{bmatrix} \boldsymbol{Z}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z} \end{bmatrix}, \quad \boldsymbol{Z}^{-1} = \exp(i\boldsymbol{\beta}\Delta x),$$

(2.6b)

$$\mathbf{T}_{\boldsymbol{\rho}_j} \equiv \exp(i\mathbf{C}_{\boldsymbol{\kappa}}\Delta x) = \begin{bmatrix} \boldsymbol{t}_j^{-1*} & -\boldsymbol{t}_j^{-1*}\boldsymbol{\rho}_j^* \\ -\boldsymbol{t}_j^{-1}\boldsymbol{\rho}_j & \boldsymbol{t}_j^{-1} \end{bmatrix}, \quad \begin{array}{l} \boldsymbol{\rho}_j = i\tanh[(\boldsymbol{\kappa}^*\boldsymbol{\kappa})^{1/2}\Delta x](\boldsymbol{\kappa}^*\boldsymbol{\kappa})^{-1/2}\boldsymbol{\kappa}^*, \\ \boldsymbol{t}_j = \cosh[(\boldsymbol{\kappa}^*\boldsymbol{\kappa})^{1/2}\Delta x]^{-1}, \end{array}$$

(2.6c)

$$\mathbf{T}_{\boldsymbol{\Phi}_j} \equiv \exp(i\mathbf{C}_{\boldsymbol{\sigma}}\Delta x) = \begin{bmatrix} \boldsymbol{\Phi}_j & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_j^* \end{bmatrix}, \quad \boldsymbol{\Phi}_j = \exp(i\boldsymbol{\sigma}\Delta x).$$

The form of the matrix in (2.6b) may, for example, be verified by evaluating the power series expansion of the matrix exponential. In principle, it suffices to express (2.6) to first order in Δx; however, the exact form is kept to emphasize the properties of each of the three sections, to ensure that each section is lossless regardless of the value of Δx, and to retain the correspondence to the discrete case (below).

We are now in the position that we can argue for the forms of the coupling matrices (2.3). Note that while we have permitted loss in the propagation section $\boldsymbol{Z}^{-1}$, the coupling sections are assumed lossless. Since the coupling sections also are assumed to be reciprocal, their transfer matrices satisfy (B.10) and (B.11) (Appendix B). Allowing a more general $\mathbf{C}_{\boldsymbol{\kappa}}$ by substituting $\boldsymbol{\kappa}^* \to \boldsymbol{\kappa}_{21}$ into the $(2,1)$ block and expanding $\exp(i\mathbf{C}_{\boldsymbol{\kappa}}\Delta x)$ to first order in Δx, the lossless and reciprocity conditions give $\boldsymbol{\kappa}_{12} = -\boldsymbol{\kappa}^*$ and dictate $\boldsymbol{\kappa}$ to be symmetric. Similarly, we can derive the form of $\mathbf{C}_{\boldsymbol{\sigma}}$ and establish that $\boldsymbol{\Phi}_j$ must be unitary; i.e., $\boldsymbol{\sigma}$ is hermitian.

From the discussion above, each layer is characterized by a unitary codirectional coupling matrix $\boldsymbol{\Phi}_j$ and a discrete reflector. Let superscript T denote transpose and let $\|\cdot\|$ be the usual matrix 2-norm. The discrete reflector satisfies $\boldsymbol{\rho}_j = \boldsymbol{\rho}_j^{\mathrm{T}}$ and $\|\boldsymbol{\rho}_j\| < 1$ and has an associated, positive definite transmission matrix $\boldsymbol{t}_j$ with $\boldsymbol{t}_j^2 = \boldsymbol{I} - \boldsymbol{\rho}_j\boldsymbol{\rho}_j^*$.

So far we have considered a continuous scattering structure and discretized it into a cascade of codirectional coupling, reflection, and pure propagation. Obviously, we can also describe discrete coupling directly. The most general, lossless, reciprocal coupling element can be described as a discrete reflector sandwiched between two codirectional coupling sections (Appendix B). Compared to our discrete model above, there is an extra codirectional coupling section on the right-hand side of the reflector. In the special case where all modes have equal effective index, $\boldsymbol{Z}^{-1} \propto \boldsymbol{I}$, this coupling section commutes with the delay section, and as a result it can be absorbed into the next, adjacent layer on the right-hand side. However, in the general case this extra coupling section does not commute with the delay section and cannot be ignored. For inverse scattering, this coupling section should therefore not be present since it would not be possible to determine the transmission through the layer uniquely from the reflection which prevents unique reconstruction, at least when only using the reflection response as the starting point. Under this assumption, $\boldsymbol{t}_j$ is positive semidefinite and uniquely determined by $\boldsymbol{t}_j^2 = \boldsymbol{I} - \boldsymbol{\rho}_j \boldsymbol{\rho}_j^*$. We restrict ourselves to reflectors that satisfy $\|\boldsymbol{\rho}_j\| < 1$; otherwise the reflector will mask the later part of the structure such that the inverse scattering procedure will not be possible. Also, with two or more layers with $\|\boldsymbol{\rho}_j\| = 1$, the structure may behave as an ideal resonator with bound modes.

Writing out the transfer matrix (2.5) of each layer, we obtain

$$\mathbf{T}_j = \begin{bmatrix} \boldsymbol{Z}^{-1}\boldsymbol{t}_j^{-1*}\boldsymbol{\Phi}_j & -\boldsymbol{Z}^{-1}\boldsymbol{t}_j^{-1*}\boldsymbol{\rho}_j^*\boldsymbol{\Phi}_j^* \\ -\boldsymbol{Z}\boldsymbol{t}_j^{-1}\boldsymbol{\rho}_j\boldsymbol{\Phi}_j & \boldsymbol{Z}\boldsymbol{t}_j^{-1}\boldsymbol{\Phi}_j^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}^{-1}\boldsymbol{K}_j & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}\boldsymbol{K}_j^* \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\Upsilon}_j^* \\ -\boldsymbol{\Upsilon}_j & \boldsymbol{I} \end{bmatrix}, \tag{2.7}$$

where $\boldsymbol{\Upsilon}_j = \boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\rho}_j\boldsymbol{\Phi}_j$ and $\boldsymbol{K}_j = \boldsymbol{t}_j^{-1*}\boldsymbol{\Phi}_j$. The transfer matrix can be converted into a scattering matrix (Appendix B):

$$\mathbf{S}_j = \begin{bmatrix} \boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\rho}_j\boldsymbol{\Phi}_j & \boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{t}_j\boldsymbol{Z}^{-1} \\ \boldsymbol{Z}^{-1}\boldsymbol{t}_j^*\boldsymbol{\Phi}_j & -\boldsymbol{Z}^{-1}\boldsymbol{t}_j^{-1*}\boldsymbol{\rho}_j^*\boldsymbol{t}_j\boldsymbol{Z}^{-1} \end{bmatrix}. \tag{2.8}$$

Thus, $\boldsymbol{\Upsilon}_j$ represents the reflection response from the left of layer j.

The combined transfer matrix describing the total structure with N layers is given by

$$\mathbf{T} = \mathbf{T}_{N-1}\mathbf{T}_{N-2}\cdots\mathbf{T}_1\mathbf{T}_0. \tag{2.9}$$

From this matrix we can determine the reflection and transmission response using (B.3). For example, the reflection response from the left is

$$\boldsymbol{R}(\omega) \equiv \boldsymbol{S}_{11} = -\boldsymbol{T}_{22}^{-1}\boldsymbol{T}_{21}, \tag{2.10}$$

where $\boldsymbol{T}_{kl}$ are the $P \times P$ blocks in $\mathbf{T}$. Assuming $\|\boldsymbol{\rho}_j\| < 1$ for all j, it can be proven by induction that $\boldsymbol{T}_{22}$ is invertible on and above the real frequency axis in the complex ω-plane for any number of layers. Physically this is obvious since $\boldsymbol{T}_{22}^{-1}$ is the transmission response from the right, and therefore it must exist and be causal and stable.

Reciprocity (B.4a) gives $\boldsymbol{R}(\omega) = \boldsymbol{R}(\omega)^{\mathrm{T}}$. Using $\|\boldsymbol{\rho}_j\| < 1$ for all j, it can be shown by induction that $\|\boldsymbol{R}(\omega)\| < 1$ for a passive structure (a passive structure is characterized by $\operatorname{Im}\beta_p \geq 0$ for all p). By causality the reflection response can be written in the form

$$\boldsymbol{R}(\omega) = \int_0^\infty \boldsymbol{h}(t)\exp(i\omega t)\mathrm{d}t, \tag{2.11}$$

where $\boldsymbol{h}(t)$ is called the time-domain impulse response.

When the modes are nondispersive, i.e., $\boldsymbol{\beta}$ is linearly related to frequency, $\boldsymbol{h}(t)$ equals a train of nonequally spaced, weighted delta pulses:

$$\boldsymbol{h}(t) = \sum_{k=0}^{\infty} \boldsymbol{h}^k \delta(t - t^k). \tag{2.12}$$

Here $\boldsymbol{h}^k$ and t^k are the weight and arrival time of the kth pulse, respectively. Substituting (2.12) into (2.11) gives

$$\boldsymbol{R}(\omega) = \sum_{k=0}^{\infty} \boldsymbol{h}^k \exp(i\omega t^k). \tag{2.13}$$

The weights $\boldsymbol{h}^k$ can, in principle, be calculated from $\boldsymbol{R}(\omega)$ using an inverse transform of the form

$$\boldsymbol{h}^k = \lim_{\omega_{\max}\to\infty} \frac{1}{2\omega_{\max}} \int_{-\omega_{\max}}^{\omega_{\max}} \boldsymbol{R}(\omega) \exp(-i\omega t^k) \mathrm{d}\omega. \tag{2.14}$$

The arrival times are determined by the delay from one layer to the next of each mode. Let Δt_p be the delay of mode p through a single layer. A delta pulse at $t = 0$ is incident to the structure on the left-hand side. Consider the reflection from the different layers, as seen from the left-hand side of the structure. From layer 0, the arrival times in all modes will be zero. An impulse in mode p reflected from layer 1 into mode q will arrive at $\Delta t_p + \Delta t_q$. Thus, considering layer 1, the arrival times are any combinations of two unit delays Δt_p. Considering layer 2, the arrival times are any combinations of four unit delays, and so forth.

When the modes are dispersive, the impulse response is no longer a train of delta functions. Nevertheless, for $t = 0$ it can still be written as $\boldsymbol{h}^0\delta(t)$, and the weight $\boldsymbol{h}^0$ can be found from (2.14).

Equation (2.13) clearly demonstrates that, in principle, for a discrete structure the reflection response $\boldsymbol{R}(\omega)$ does not approach zero for large frequencies. Only in the special case where the modal effective indices are rational numbers with common denominator is the reflection spectrum periodic. Fortunately, in practice, it is not necessary to represent the entire bandwidth to enable inverse scattering for a discrete structure. As shown in the next section, what is needed in the layer-stripping algorithm is the zeroth point of the impulse response at time $t = 0$. Since the next nonzero value is for $t = 2\min_p \Delta t_p$,[1] the zeroth point is computed accurately, provided the represented bandwidth $\omega_{\max}$ satisfies $\omega_{\max} \gg 1/\min_p \Delta t_p$. Then, if the true reflection spectrum is multiplied by a smooth window function $W(\omega)$ that goes to zero at $\omega = \pm\omega_{\max}$, the inverse Fourier transform evaluated around zero is approximately $w(t)\boldsymbol{h}^0$, where $w(t)$ is the inverse Fourier transform of $W(\omega)$. Since $w(0)\boldsymbol{h}^0 \approx \frac{1}{2\pi}\int_{-\omega_{\max}}^{\omega_{\max}} W(\omega)\boldsymbol{R}(\omega)\mathrm{d}\omega$, we can find $\boldsymbol{h}^0$ from a measurement of $\boldsymbol{R}(\omega)$ in the bandwidth $(-\omega_{\max}, \omega_{\max})$:

$$\boldsymbol{h}^0 \approx \frac{\int_{-\omega_{\max}}^{\omega_{\max}} W(\omega)\boldsymbol{R}(\omega)\mathrm{d}\omega}{\int_{-\omega_{\max}}^{\omega_{\max}} W(\omega)\mathrm{d}\omega}. \tag{2.15}$$

In many practical cases, the structure to be reconstructed is quasi-sinusoidal. More generally, the structure is often quasi-periodic, and, e.g., the first "Fourier component" is to be reconstructed. In such cases, one can define modal field envelopes which vary slowly with respect to x (compared to a wavelength). Similarly, one can extract slowly varying coupling coefficient envelopes. As a result, all quantities in (2.1) vary slowly with x. The relevant bandwidth in (2.15) will then be centered about a chosen "design frequency" rather than zero. The main advantage of this

[1] For simplicity it is assumed to be a nondispersive structure.

procedure is that it leads to considerably fewer requirements on the spatial resolution and, as a result, efficient inverse scattering. This modification to the model is detailed in section 4.

3. Layer-stripping method. The inverse scattering problem can now be stated as follows: Given a structure consisting of N layers, each layer consists of three sections (sublayers), the first ($\mathbf{\Phi}_j$) responsible for coupling between copropagating modes, the second ($\boldsymbol{\rho}_j$) responsible for coupling between counterpropagating modes, and the third a pure propagating section ($\boldsymbol{Z}^{-1}$). The propagation constants of the involved modes are known and specified in terms of $\boldsymbol{Z}^{-1}$.[2] From a set of excitation-response pairs (that is, $\boldsymbol{R}(\omega)$), we want to reconstruct $\boldsymbol{\rho}_j$ and $\mathbf{\Phi}_j$ for all j.

The structure itself and the medium to the right are assumed to be at rest at time $t = 0$. For incident waves from the left, the reflection response from the structure is described by the matrix $\boldsymbol{R}(\omega)$ of dimension $P \times P$. This matrix can be viewed as the operator which takes the excitation field vector to the reflected field vector. Its columns can be interpreted as the responses for orthonormal excitation basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_P$, respectively. Here $\boldsymbol{e}_p$ has only one nonzero element (equal to unity) at position p. Similarly, we can define the forward ($\boldsymbol{u}_j(\omega)$) and backward ($\boldsymbol{v}_j(\omega)$) propagating field matrices as $P \times P$ matrices where the columns are the fields for orthonormal excitations $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_P$. A subscript j is specified to emphasize that $\boldsymbol{u}_j(\omega)$ and $\boldsymbol{v}_j(\omega)$ are the fields at the beginning (left-hand side) of layer j. The field matrices of layer $j + 1$ are related to the field matrices of layer j by

$$\begin{bmatrix} \boldsymbol{u}_{j+1}(\omega) \\ \boldsymbol{v}_{j+1}(\omega) \end{bmatrix} = \mathbf{T}_j \begin{bmatrix} \boldsymbol{u}_j(\omega) \\ \boldsymbol{v}_j(\omega) \end{bmatrix}, \tag{3.1}$$

where $\mathbf{T}_j$ is given by (2.7).

The layer-stripping algorithm is based on the simple fact that the leading edge of the impulse response is independent on later parts of the structure due to causality. Hence, one can identify the first layer of the structure and subsequently remove its effect using the associated transfer matrix.

For layer 0, we initialize $\boldsymbol{u}_0(\omega) = \boldsymbol{I}$ and $\boldsymbol{v}_0(\omega) = \boldsymbol{R}(\omega)$. We define a local reflection spectrum $\boldsymbol{R}_j(\omega) = \boldsymbol{v}_j(\omega)\boldsymbol{u}_j(\omega)^{-1}$ and the associated impulse response $\boldsymbol{h}_j(t)$ as the response of the structure after removing the first $j-1$ layers. Similarly to the impulse response of the entire structure, $\boldsymbol{h}_j(t)$ contains an isolated delta function at $t = 0$. Due to causality, this pulse is equal to the reflection from the zeroth layer alone. Denoting the weight of this pulse $\boldsymbol{h}_j^0$, we find from (2.8) that

$$\boldsymbol{h}_j^0 = \mathbf{\Upsilon}_j \equiv \mathbf{\Phi}_j^{\mathrm{T}} \boldsymbol{\rho}_j \mathbf{\Phi}_j. \tag{3.2}$$

Note that $\boldsymbol{R}_j(\omega)$ is symmetric for all ω as a result of reciprocity; thus $\boldsymbol{h}_j^0$ is symmetric as well. Writing out (3.1) and (2.7) and substituting $\boldsymbol{v}_j(\omega) = \boldsymbol{R}_j(\omega)\boldsymbol{u}_j(\omega)$, we obtain

$$\boldsymbol{u}_{j+1}(\omega) = \boldsymbol{Z}^{-1}\boldsymbol{K}_j \left[\boldsymbol{I} - \mathbf{\Upsilon}_j^* \boldsymbol{R}_j(\omega)\right] \boldsymbol{u}_j(\omega), \tag{3.3a}$$

$$\boldsymbol{v}_{j+1}(\omega) = \boldsymbol{Z}\boldsymbol{K}_j^* \left[\boldsymbol{R}_j(\omega) - \mathbf{\Upsilon}_j\right] \boldsymbol{u}_j(\omega), \tag{3.3b}$$

[2] The effective indices may contain small, real, unknown parts Δn_p, i.e., $n_p = n_{p,\mathrm{known}} + \Delta n_p$, where $n_{p,\mathrm{known}}$ are known. Provided Δn_p is sufficiently small, the variation of the associated phase factor $\exp(i\omega\Delta n_p \Delta x/c)$ may be small over the relevant bandwidth. In such cases the unknown parts can be absorbed into the $\mathbf{\Phi}_j$'s.

and therefore

$$\boldsymbol{R}_{j+1}(\omega) = \boldsymbol{Z}\boldsymbol{K}_j^* \left[\boldsymbol{R}_j(\omega) - \boldsymbol{\Upsilon}_j\right] \left[\boldsymbol{I} - \boldsymbol{\Upsilon}_j^* \boldsymbol{R}_j(\omega)\right]^{-1} \boldsymbol{K}_j^{-1} \boldsymbol{Z}. \tag{3.4}$$

Provided $\boldsymbol{\Upsilon}_j$ and $\boldsymbol{K}_j$ are known, (3.4) shows that the local reflection spectrum of layer $j+1$ can be calculated directly from the local reflection spectrum of layer j without calculating the fields $\boldsymbol{u}_{j+1}$ and $\boldsymbol{v}_{j+1}$. Note the similarity to the Schur formula used in scalar layer-stripping [6].

To characterize layer j completely and to identify $\boldsymbol{K}_j$, we must determine $\boldsymbol{\rho}_j$ and $\boldsymbol{\Phi}_j$. By counting the available degrees of freedom (in $\boldsymbol{\Upsilon}_j$), we immediately find that this cannot be done uniquely. It is therefore necessary to use a priori information on $\boldsymbol{\rho}_j$ and/or $\boldsymbol{\Phi}_j$. The available information may vary from situation to situation. Here we will consider the following situations, where $\boldsymbol{\rho}_j$ and $\boldsymbol{\Phi}_j$ can be found using the methods in sections A.1 and A.2.

(a) $\boldsymbol{\Phi}_j = \boldsymbol{I}$. In this case there is no codirectional coupling. The identification of the layer is now particularly simple, as $\boldsymbol{\rho}_j = \boldsymbol{\Upsilon}_j$ uniquely. Note that while there is no codirectional coupling, $\boldsymbol{\rho}_j$ describes reflection from all modes into all modes. Thus the different modes may still interact.

(b) $\boldsymbol{\rho}_j$ is diagonal and nonnegative. Now $\boldsymbol{\rho}_j$ is a simple partial reflector which reflects only light into the same mode as the incident field (no reflection into other modes). The coupling between different modes is instead described by $\boldsymbol{\Phi}_j$. Since $\boldsymbol{\Upsilon}_j = \boldsymbol{\Phi}_j^{\mathrm{T}} \boldsymbol{\rho}_j \boldsymbol{\Phi}_j$, $\boldsymbol{\rho}_j$ is found uniquely as the singular value matrix associated with $\boldsymbol{\Upsilon}_j$ up to reordering of the singular values. Once the order of the singular values has been established, the unitary $\boldsymbol{\Phi}_j$ is found uniquely up to the sign of its rows, provided all singular values are distinct and nonzero (see section A.1). When one or more singular values of $\boldsymbol{\Upsilon}_j$ are zero, the corresponding row(s) of $\boldsymbol{\Phi}_j$ cannot be determined uniquely. More precisely, $\boldsymbol{\Phi}_j$ is determined up to a premultiplicative unitary matrix $\boldsymbol{J}$ operating on the associated mode(s). Physically, this is obvious since when a singular value is zero, the associated mode is not reflected from the layer. When two or more nonvanishing singular values are equal, $\boldsymbol{\Phi}_j$ is determined up to a premultiplicative, real unitary $\boldsymbol{J}$ operating on the associated modes. Physically, this means that these modes experience the same reflection, and thus an arbitrary (real) "rotation" of the modes is not detected. In such cases, the unitary section $\boldsymbol{\Phi}_j$, as determined by the method in section A.1, does not necessarily correspond to the physical section. This error will propagate to the next layers according to (3.4).

(c) $\boldsymbol{\Phi}_j$ is symmetric and $\boldsymbol{\rho}_j$ is real and positive semidefinite. A special case in which there are only two degenerate modes in each direction is treated in [41]. The reflector matrix $\boldsymbol{\rho}_j$ can be written $\boldsymbol{P}_j^{\mathrm{T}} \boldsymbol{\Sigma}_j \boldsymbol{P}_j$, where $\boldsymbol{P}_j$ is a real, special unitary matrix and $\boldsymbol{\Sigma}_j$ is diagonal and nonnegative. Since $\boldsymbol{\Upsilon}_j = \boldsymbol{\Phi}_j^{\mathrm{T}} \boldsymbol{\rho}_j \boldsymbol{\Phi}_j = \boldsymbol{\Phi}_j^{\mathrm{T}} \boldsymbol{P}_j^{\mathrm{T}} \boldsymbol{\Sigma}_j \boldsymbol{P}_j \boldsymbol{\Phi}_j$, we find $\boldsymbol{\Sigma}_j$ and $\boldsymbol{P}_j \boldsymbol{\Phi}_j$ as in the previous case, with the identical ambiguity issues. The separate identification of $\boldsymbol{P}_j$ and $\boldsymbol{\Phi}_j$ is accomplished using the factorization method in section A.2, with certain ambiguities related to the sign of the eigenvalues of $\boldsymbol{\Phi}_j$.

The ambiguities when determining $\boldsymbol{\Phi}_j$ in situation (b) are in fact very similar to the well-known ambiguities in the scalar case with a single mode in each direction. In the scalar case any π phase-shift sections between the reflectors cannot be identified since the associated round-trip phase accumulated to and from a reflector becomes 2π. In our multimode case, the sign of the rows of the "phase-delay" section ($\boldsymbol{\Phi}_j$)

between two reflectors cannot be identified. Similarly, in the scalar case, any phase-shift section preceding a zero reflector cannot be determined uniquely. Instead it is chosen arbitrarily (e.g., removed) and attributed to the next layer with a nonzero reflector.

When the structure to be reconstructed is a discretized version of a smooth structure, the smoothness can be used to resolve ambiguites. First we consider situation (b). For small Δx, $\mathbf{\Phi}_j$ is close to identity; thus the sign of the rows of $\mathbf{\Phi}_j$ can be determined uniquely. If $\boldsymbol{\rho}_j$ has distinct eigenvalues, valid for all j, the order of the eigenvalues of $\boldsymbol{\rho}_j$ can be determined from the order of the eigenvalues of $\boldsymbol{\rho}_{j-1}$ using the smoothness of $\boldsymbol{\kappa} = \boldsymbol{\kappa}(x)$. If there are equal eigenvalues for a certain reflector $\boldsymbol{\rho}_j$, or if $\boldsymbol{\rho}_j$ is singular, the ambiguities of $\mathbf{\Phi}_j$ are characterized by the premultiplicative $\boldsymbol{J}$ matrix (section A.1). In other words, the chosen $\mathbf{\Phi}_j$ is related to the corresponding true matrix ($\mathbf{\Phi}_{j,\text{true}}$) by $\mathbf{\Phi}_j = \boldsymbol{J}\mathbf{\Phi}_{j,\text{true}}$. By choosing $\boldsymbol{J}$ such that $\|\mathbf{\Phi}_j - \mathbf{\Phi}_{j-1}\|$ is minimum, the resulting $\boldsymbol{J}$ is close to identity (that is, $\|\boldsymbol{J} - \boldsymbol{I}\| \leq 2\|\mathbf{\Phi}_{j,\text{true}} - \mathbf{\Phi}_{j-1}\|$). Since $\boldsymbol{t}_j$ and $\boldsymbol{Z}^{-1}$ are close to identity as well, the order of three sections $\boldsymbol{J}$, $\boldsymbol{t}_j$, and $\boldsymbol{Z}^{-1}$ can be interchanged (see section 2). Thus the error due to wrong choice of $\mathbf{\Phi}_j$ can be absorbed into $\mathbf{\Phi}_{j+1}$. More generally, provided only a few neighboring layers have singular or degenerate $\boldsymbol{\rho}_j$'s, only the corresponding and following $\mathbf{\Phi}_j$ sections may be determined erroneously, and the determination of the later part of the structure is (approximately) unaffected.

In situation (c), the order of eigenvalues of $\boldsymbol{\rho}_j$ can be determined as in situation (b). However, $\boldsymbol{P}_j\mathbf{\Phi}_j$ is not necessarily close to identity. Nevertheless, the sign of its rows can be determined from $\boldsymbol{P}_{j-1}\mathbf{\Phi}_{j-1}$ if $\boldsymbol{\kappa} = \boldsymbol{\kappa}(x)$ is sufficiently smooth. (Recall that $\boldsymbol{P}_j\mathbf{\Phi}_j$ is unitary, which means that in each row there exists at least one element of magnitude $\geq 1/\sqrt{P}$.) Finally, since $\mathbf{\Phi}_j$ is close to identity, its eigenvalues are close to unity. It follows that the factorization of $\boldsymbol{P}_j\mathbf{\Phi}_j$ into $\boldsymbol{P}_j$ and $\mathbf{\Phi}_j$ is unique (section A.2).

From the discussion above, we summarize the layer-stripping algorithm, analogously to the scalar version described in [6, 5], that can be applied to identify a structure supporting multiple modes:

(1) Initialize $j = 0$. Set $\boldsymbol{R}_j(\omega) = \boldsymbol{R}(\omega)$.

(2) Compute the zeroth weight $\boldsymbol{h}_j^0$ of the impulse response. In practice this is achieved by the substitutions $\boldsymbol{h}^0 \to \boldsymbol{h}_j^0$ and $\boldsymbol{R}(\omega) \to \boldsymbol{R}_j(\omega)$ in (2.15).

(3) Use a model-specific factorization of $\boldsymbol{h}_j^0 = \mathbf{\Phi}_j^{\mathrm{T}}\boldsymbol{\rho}_j\mathbf{\Phi}_j$ to find $\mathbf{\Phi}_j$ and $\boldsymbol{\rho}_j$.

(4) Calculate $\boldsymbol{t}_j = (\boldsymbol{I} - \boldsymbol{\rho}_j\boldsymbol{\rho}_j^*)^{1/2}$ such that the associated eigenvalues are positive, and set $\boldsymbol{K}_j = \boldsymbol{t}_j^{-1}\mathbf{\Phi}_j$.

(5) Calculate the next, local reflection response $\boldsymbol{R}_{j+1}(\omega)$ using (3.4).

(6) If $j < N - 1$, increase j and return to (2).

When the scattering structure is continuous, one can use the true reflection spectrum as input to the layer-stripping algorithm, even though the structure is modeled discrete. This can be justified as follows: The layer thickness Δx is chosen small such that the first order approximations of $\exp(i\mathbf{C}_{\boldsymbol{\kappa}}\Delta x)$ and $\exp(i\mathbf{C}_{\boldsymbol{\sigma}}\Delta x)$ are accurate. (Thus an upper bound on $\|\mathbf{C}_{\boldsymbol{\kappa}}\|$ and $\|\mathbf{C}_{\boldsymbol{\sigma}}\|$ should be known a priori.) Let $\omega \leq \omega_{\text{b}}$ be the bandwidth where the true reflection spectrum is significantly different from zero. For sufficiently small Δx, the first order approximation of $\exp(i\mathbf{D}\Delta x)$ is valid, and the true reflection spectrum is approximately equal to that of the corresponding discrete model in the bandwidth $\omega \leq \omega_{\text{b}}$. In the limit $t \to 0^+$, the (p, q) element of the impulse response of the continuous structure can be calculated exactly from (2.1) using the Born approximation, yielding

$$(3.5)\quad h_{pq}(t=0^+) \equiv \frac{1}{2\pi}\lim_{t\to 0^+}\int_{-\infty}^{\infty} R_{pq}(\omega)\exp(-i\omega t)\mathrm{d}\omega = i\kappa_{pq}^*(x=0^+)c/(n_p+n_q).$$

Here $\kappa_{pq}(x=0^+)$ is the (p,q) element of $\boldsymbol{\kappa}(x)$ at $x=0^+$. For practical computations, the integral in (3.5) must be truncated at $\pm\omega_{\mathrm{b}}$; thus, to find the leading edge of $h_{pq}(t)$, one can take $t=0$ in the integral and multiply the result by a factor of two. (Recall that by causality $\lim_{t\to 0^-} h_{pq}(t)=0$.) Once $\boldsymbol{\kappa}$ for the zeroth layer is found, one can propagate the fields using (3.4). Since we have not identified the codirectional coupling $\boldsymbol{\Phi}_0$ of the zeroth layer, $\boldsymbol{\Phi}_0$ is associated with the next layer. Thus, after the zeroth layer has been stripped off, the leading edge of the impulse response of the remaining structure becomes

$$(3.6)\qquad \boldsymbol{\Phi}_0^{\mathrm{T}}\left[i\kappa_{pq}^*(x=\Delta x^+)c/(n_p+n_q)\right]\boldsymbol{\Phi}_0,$$

where the square bracket denotes a matrix formed by the elements inside. The identification of $\boldsymbol{\Phi}_0$ and $\left[i\kappa_{pq}^*/(n_p+n_q)\right]$ can now be accomplished using the factorization methods described above. The algorithm continues in the same way, until finally the bandwidth of the reflection spectrum of the remaining structure exceeds ω_{b}. This remaining part of the structure can be made arbitrarily thin by choosing a sufficiently small Δx.

The difference between the latter “quasi-continuous” formulation and the discrete algorithm is essentially the factor n_p+n_q and the method for evaluating the leading edge or first point of the impulse response. When the effective indices can be approximated by some number n_0 for all p, $n_p\approx n_0$, one can in fact use the discrete algorithm directly: A periodic extension of the true reflection spectrum outside a principal bandwidth $[-\omega_{\max},\omega_{\max}]$ corresponds then to a discrete model with $\Delta x=\pi c/(2n_0\omega_{\max})$. The first point of the impulse response is calculated by (2.15) using a rectangular window function $W(\omega)$. For a broad class of waveguides of practical interest, the effective indices are similar (see section 4). While the phase relation between the modes, as described by $\boldsymbol{Z}^{-1}$, may still result in a nontrivial multimode coupling, the discrete algorithm gives accurate results. The errors due to this periodic spectrum approximation can be corrected to some extent by including the factor $(n_p+n_q)/(2n_0)$ in the elements on the right-hand side of (2.15). This can be justified, e.g., using the Born approximation.

4. Quasi-sinusoidal coupling structures. Continuous coupling in acoustical, radio frequency, or optical waveguides may be obtained by perturbation of the effective indices n_p associated with each mode. This can be achieved by modulation of the wall profile or waveguide medium properties. As a concrete example, we will discuss fiber Bragg gratings [17], which have attracted large interest recently due to their applications in fiber optical communications and sensors. A fiber grating is formed in an optical fiber by modulating the refractive index of the core periodically or quasi-periodically. The main peak of the reflection spectrum appears for the frequency where the reflection from a crest in the index modulation is in phase with the next reflection. Permanent gratings are fabricated by UV-illumination. In fibers doped with certain dopants such as germanium, the UV-illumination will permanently rise the refractive index of the core. Advanced fabrication methods have made it possible to manufacture complex gratings with varying index modulation amplitude and period. The layer-stripping algorithm is the most widely used method for designing the index profile to obtain a given reflection spectrum [10, 37, 36].

In most cases, the fiber grating is formed in a single-mode fiber, and coupling is considered only between the forward-propagating and backward-propagating fundamental mode. The field matrices $\boldsymbol{u}_j(\omega)$ and $\boldsymbol{v}_j(\omega)$ are then scalar functions. However, in some cases it is not sufficient to consider only one forward-propagating mode and one backward-propagating mode. For instance, a single mode fiber is always slightly birefringent, and the photosensitivity can be polarization-dependent [16]. In this case, two forward-propagating and two backward-propagating polarization modes must be considered. An inverse scattering algorithm that takes into account polarization mode coupling is described in [41]. The coupling between the two polarization modes is described by Jones matrices [20]. Both polarization modes have approximately the same effective index, and so $\boldsymbol{Z}^{-1} = \exp(i\beta\Delta x)\boldsymbol{I}$, where the common propagation constant β is scalar.

In a multimode fiber, the modulation of the refractive index may result in coupling between the fundamental mode and other modes. Each mode has a transversal field profile $\Psi_p(r,\phi)$ which is a solution to the scalar wave equation in polar coordinates r and ϕ [38]:[3]

$$\left\{\nabla_{\mathrm{t}}^2 + k^2(\bar{n}^2(r) - n_p^2)\right\}\Psi_p(r,\phi) = 0. \tag{4.1}$$

Here $\bar{n}(r)$ is the unperturbed, refractive index profile of the fiber, which is assumed to be real, ∇_{t} is the transversal nabla operator, and $k = \omega/c$. The field $\Psi_p(r,\phi)$ and its first derivatives are continuous. For bound modes, the fields are real and orthonormal such that $\int_{A_\infty} \Psi_p(r,\phi)\Psi_q(r,\phi)\mathrm{d}A = \delta(p-q)$, where $\delta(p-q)$ denotes the Kronecker delta, and A_∞ is the entire transversal plane. The effective indices n_p are eigenvalue solutions to (4.1). A mode p is bound when $n_{\mathrm{cl}} < n_p \leq n_{\mathrm{co}}$, where n_{co} and n_{cl} are the refractive indices of the fiber core and cladding, respectively. Ignoring radiation modes, which in the vicinity of the core decay rapidly away from the excitation source, the total electric field $E(r,\phi,x)$ can be written as a superposition of forward- and backward-propagating bound modes:

$$E(r,\phi,x) = \sum_{p=1}^{P}(b_p^+(x) + b_p^-(x))\Psi_p(r,\phi). \tag{4.2}$$

Here $b_p^\pm(x)$ contain all x-dependence including the harmonic propagation factor $\exp(\pm i\beta_p x)$, where $\beta_p = kn_p$.

Coupling between the modes originates from longitudinal modulation of the refractive index. Let the refractive index be perturbed quasi-periodically with a spatial period Λ,

$$n(r,\phi,x) = \bar{n}(r) + \Delta n_{\mathrm{ac}}(r,\phi,x)\cos\left(\frac{2\pi}{\Lambda}x + \theta(x)\right) + \Delta n_{\mathrm{dc}}(r,\phi,x), \tag{4.3}$$

where $\Delta n_{\mathrm{ac}}(r,\phi,x)$, $\Delta n_{\mathrm{dc}}(r,\phi,x)$, and $\theta(x)$ are slowly varying with x over a distance Λ. We assume that $\Delta n_{\mathrm{ac}}(r,\phi,x) \ll \bar{n}$, and $\Delta n_{\mathrm{dc}}(r,\phi,x) \ll \bar{n}$, which is the case for practical fiber gratings. The total electric field must satisfy the scalar wave equation for the perturbed fiber, i.e.,

$$\left\{\nabla_{\mathrm{t}}^2 + \frac{\partial^2}{\partial x^2} + k^2 n^2(r,\phi,x)\right\}E(r,\phi,x) = 0. \tag{4.4}$$

[3] To find the exact electromagnetic modes, the vector wave equation must be solved. However, for weakly guiding waveguides (waveguides with small difference between the refractive index of the core and the cladding), the scalar wave equation can be used. This is the case for most conventional fibers.

We now substitute (4.2) into (4.4), take (4.1) into account, and multiply the resulting equation by $\Psi_q(r,\phi)$. By integration over the entire transversal plane and recalling that the modes are orthonormal, the resulting set of second order differential equations can be decomposed into first order coupled mode equations [38],

$$\frac{\mathrm{d}b_p^+(x)}{\mathrm{d}x} - i\beta_p b_p^+(x) = i\sum_{q=1}^{P} \mathcal{C}_{pq}(x)(b_q^+(x)+b_q^-(x)), \tag{4.5a}$$

$$\frac{\mathrm{d}b_p^-(x)}{\mathrm{d}x} + i\beta_p b_p^-(x) = -i\sum_{q=1}^{P} \mathcal{C}_{pq}(x)(b_q^+(x)+b_q^-(x)), \tag{4.5b}$$

where

$$\mathcal{C}_{pq}(x) = \frac{k}{2n_p}\int_{A_\infty} (n^2(r,\phi,x)-\bar{n}^2(r))\Psi_p(r,\phi)\Psi_q(r,\phi)\mathrm{d}A. \tag{4.6}$$

Note that the frequency dependence of (4.6) can be ignored in practice, since the normalized bandwidth of interest is usually much less than unity, and the field profiles and effective indices are approximately constant in this bandwidth. Also note that since the fiber is assumed to be weakly guiding, n_p can be set equal to n_{co}; thus $\mathcal{C}_{pq} = \mathcal{C}_{qp}$.

In the case of a quasi-periodic structure it is natural to write the coupling coefficient as a quasi-Fourier series:

$$\begin{aligned}\mathcal{C}_{pq}(x) = \sigma_{pq}(x) + \kappa_{pq}(x)\exp\left(i\frac{2\pi}{\Lambda}x\right) + \kappa_{pq}^*(x)\exp\left(-i\frac{2\pi}{\Lambda}x\right)\\ + \sum_{|m|\geq 2}\kappa_{pq}^{(m)}(x)\exp\left(i\frac{2\pi m}{\Lambda}x\right),\end{aligned} \tag{4.7}$$

where the "Fourier coefficients" $\kappa_{pq}(x)$, $\sigma_{pq}(x)$, and $\kappa_{pq}^{(m)}(x)$ are slowly varying over a period Λ. For a fiber grating the index modulation $n(r,\phi,x)-\bar{n}(r)$ is given by (4.3) and is small compared to $\bar{n}(r)$, and so the zeroth and first order Fourier components dominate. Note that $\arg\{\kappa_{pq}(x)\} = \theta(x)$.

The field amplitudes $b_p^{\pm}(x)$ vary rapidly; it is therefore convenient to introduce the slowly varying field envelopes $u_p(x)$ and $v_p(x)$ by setting

$$b_p^+(x) = i^{1/2}u_p(x)\exp\left(i\frac{\pi}{\Lambda}x\right)\exp\left(i\frac{\theta(x)}{2}\right), \tag{4.8a}$$

$$b_p^-(x) = i^{-1/2}v_p(x)\exp\left(-i\frac{\pi}{\Lambda}x\right)\exp\left(-i\frac{\theta(x)}{2}\right). \tag{4.8b}$$

Since an identical phase factor is removed from all modes, the reflection response as calculated from b_p^+ and b_q^- will differ only from that calculated from u_p and v_q by a constant phase factor not dependent on p and q. Inserting (4.7) and (4.8) into (4.5) and ignoring rapidly oscillating terms (since they contribute little to $\mathrm{d}u_p/\mathrm{d}x$ and $\mathrm{d}v_p/\mathrm{d}x$), we obtain an alternative set of coupled-mode equations

$$\frac{\mathrm{d}u_p(x)}{\mathrm{d}x} = i\delta_p u_p(x) - \frac{i}{2}\frac{\mathrm{d}\theta(x)}{\mathrm{d}x}u_p(x) + i\sum_{q=1}^{P}\sigma_{pq}(x)u_q(x) + \sum_{q=1}^{P}|\kappa_{pq}(x)|v_q(x), \tag{4.9a}$$

$$\frac{\mathrm{d}v_p(x)}{\mathrm{d}x} = -i\delta_p v_p(x) + \frac{i}{2}\frac{\mathrm{d}\theta(x)}{\mathrm{d}x}v_p(x) - i\sum_{q=1}^{P}\sigma_{pq}(x)v_q(x) + \sum_{q=1}^{P}|\kappa_{pq}(x)|u_q(x), \tag{4.9b}$$

where $\delta_p = \beta_p - \pi/\Lambda = n_p\omega/c - \pi/\Lambda$ is the wavenumber detuning of mode p. Thus, $-i|\kappa_{pq}(x)|$ is the coupling coefficient between modes p and q propagating in opposite directions, while $\sigma_{pq}(x) - \delta(p-q)(\mathrm{d}\theta(x)/\mathrm{d}x)/2$ is the coupling coefficient between modes p and q in the same direction. With $\boldsymbol{E} = [u_1, u_2, \ldots, u_P, v_1, v_2, \ldots, v_P]^{\mathrm{T}}$ we find that (4.9) coincides with (2.1), where $\sigma_{pq}(x) - \delta(p-q)(\mathrm{d}\theta(x)/\mathrm{d}x)/2$ and $-i|\kappa_{pq}(x)|$ are the (p,q) elements of $\boldsymbol{\sigma}$ and $\boldsymbol{\kappa}$, respectively, and δ_p are the diagonal elements of $\boldsymbol{\beta}$. Note that δ_p do not correspond to the actual propagation constants but rather their detuning from π/Λ. Approximating the effective indices by n_{co}, this means that the bandwidth of interest is not centered about zero but rather about the "design frequency" $\omega_0 \equiv \pi c/(n_{\mathrm{co}}\Lambda)$. The frequency interval of integration in (2.15) should be centered about ω_0. As in the scalar case [36], we also note that, in general, the geometrical phase variation $\theta(x)$ cannot be distinguished from the phase variation associated with the dc index term $\Delta n_{\mathrm{dc}}(r,\phi,x)$.

We observe that $\boldsymbol{\sigma}$ is real and symmetric, and $\boldsymbol{\kappa}$ is imaginary and symmetric. Moreover, it is not difficult to realize that $i\boldsymbol{\kappa}$ is positive semidefinite.[4] Thus $\boldsymbol{\Phi}_j$ defined in (2.6) is unitary and symmetric, and $-\boldsymbol{\rho}_j$ is real and positive semidefinite. It follows that we can use the layer-stripping method together with the factorization approach (c), as given in section 3, to identify the coupling sections $\boldsymbol{\rho}_j$ and $\boldsymbol{\Phi}_j$ (and therefore the coupling matrices $\boldsymbol{\kappa}$ and $\boldsymbol{\sigma}$ as a function of position x). Since $(i\boldsymbol{\Phi}_j)^{\mathrm{T}}(-\boldsymbol{\rho}_j)(i\boldsymbol{\Phi}_j) = \boldsymbol{\Phi}_j^{\mathrm{T}}\boldsymbol{\rho}_j\boldsymbol{\Phi}_j$, the factorization approach gives $-\boldsymbol{\rho}_j$ and $i\boldsymbol{\Phi}_j$.

For a fiber grating it is usually reasonable to assume that the ac and dc index modulations can be written in the forms $\Delta n_{\mathrm{ac}}(r,\phi,x) = \Delta n(r,\phi)\Delta n_{\mathrm{ac}}(x)$ and $\Delta n_{\mathrm{dc}}(r,\phi,x) = \Delta n(r,\phi)\Delta n_{\mathrm{dc}}(x)$, respectively. Here $\Delta n(r,\phi)$ accounts for the transversal variation of the index modulation profile, and $\Delta n_{\mathrm{ac}}(x)$ and $\Delta n_{\mathrm{dc}}(x)$ are the ac and dc modulations as a function of x. As before, we assume that the index modulation and $n_{\mathrm{co}} - n_{\mathrm{cl}}$ are small, yielding

$$\boldsymbol{\kappa}(x) = -i\frac{\Delta n_{\mathrm{ac}}(x)}{2}\boldsymbol{\eta}, \tag{4.10a}$$

$$\boldsymbol{\sigma}(x) = \Delta n_{\mathrm{dc}}(x)\boldsymbol{\eta} - \frac{1}{2}\frac{\mathrm{d}\theta(x)}{\mathrm{d}x}\boldsymbol{I}, \tag{4.10b}$$

where $\boldsymbol{\eta}$ is independent on x. The elements of $\boldsymbol{\eta}$ are

$$\eta_{pq} = k\int_{A_\infty} \Delta n(r,\phi)\Psi_p\Psi_q \mathrm{d}A. \tag{4.11}$$

When the mode profiles and $\Delta n(r,\phi)$ are known, this means that the entire coupling matrix $\boldsymbol{\kappa}(x)$ is determined from only a single nonvanishing element. For $\boldsymbol{\sigma}$, two elements are needed (including at least one diagonal element). Note that, in this case, it is indeed possible to distinguish between the dc index modulation $\Delta n_{\mathrm{dc}}(x)$ and the geometrical phase variation $\mathrm{d}\theta(x)/\mathrm{d}x$ using information contained in $\boldsymbol{\sigma}$.

For characterization of multimode gratings, measurements of the reflection from every mode to every mode are required. Executing such a measurement is not trivial. Reference [33] describes using an auxiliary long-period grating (LPG), i.e., a grating with purely codirectional coupling, to characterize another interrogated LPG. With the auxiliary LPG, both modes were excited. Due to the difference in propagation

[4] The real matrix given by the elements $\Psi_p\Psi_q$ is clearly positive semidefinite, since $\sum_{p,q} a_p\Psi_p\Psi_q a_q = (\sum_p \Psi_p a_p)^2 \geq 0$ for any real a_p. For a fiber grating $\Delta n_{\mathrm{ac}}(r,\phi,x) \geq 0$ for all r and ϕ; thus $|\kappa_{pq}(x)|$ adopts the positive semidefinite property from $\Psi_p\Psi_q$.

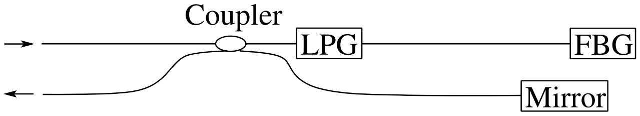


Fig. 4.1. *Measurement setup for characterization of multimode gratings.*

constant, the difference in delay between the modes was large enough to separate the response from the two modes in the time domain. Figure 4.1 shows how this method can be adopted to characterization of multimode fiber Bragg gratings (FBGs) using optical frequency domain reflectometry [12]. Here light is coupled into the fundamental mode of the input fiber and the wavelength of the highly coherent source is swept. The coupler splits the light equally into the two fibers. The LPG couples light from the fundamental modes into the other modes so that the total optical power is distributed between all modes. The light returned by the FBG will again propagate through the LPG, and some light from each mode will be coupled back into the fundamental mode. The mirror reflects only the fundamental mode, and at the coupler the reflected light from the mirror interferes with the light in the fundamental mode out of the LPG. If the fiber between the LPG and FBG is so long that the difference in the delay between the modes is larger than the length of the impulse response of the grating, the individual elements of the reflection matrix will be separable in the time domain, provided that there is no degenerated modes.

5. Numerical example. A potential application of the multimode layer-stripping method is to characterize coupling from the core mode to cladding modes in a single mode fiber. Cladding modes are bound not within the core of the fiber but by the cladding/air boundary [8]. A single mode fiber may support as many as 100 cladding modes. The power in these modes will eventually be lost to the environment. The core-cladding mode coupling can be seen clearly in the transmission spectra of strong gratings. For chirped gratings [26] and chirped, sampled gratings [27], the bandwidth may become larger than the separation in resonant wavelength between the core-core mode coupling and the core-cladding mode coupling. Then the core-cladding mode coupling will interfere with the reflection spectrum associated with the core mode [11]. This unwanted coupling is often handled by writing the grating in fibers with depressed cladding modes [7]. There has also been some attempts to take into account the core-cladding mode coupling in the design of the grating [23, 14]. Here direct scattering is treated with multiple mode coupling, but the inverse scattering has so far been purely single-mode. The layer-stripping algorithm described in section 3 can be used for characterization of such coupling and possibly for design. In contrast to the methods in [23, 14], multiple modes can be taken into account in the inverse scattering part of an iterative design process.

A simpler, but nevertheless interesting, problem is to characterize coupling in an optical fiber with a few bound modes. Here we will present a numerical experiment simulating a grating in a fiber with $n_{\text{co}} = 1.452$, $n_{\text{cl}} = 1.437$, and core radius $r_{\text{co}} = 5\ \mu\text{m}$. By solving the eigenvalue equation for a circular fiber [38], we find that this fiber supports four modes: LP_{01}, LP_{11}, LP_{21}, and LP_{02} at the design wavelength $\lambda_0 = 1.55\ \mu\text{m}$. Here the index l in LP_{lm} means that the transversal field profile can be written in the form $f_{lm}(r)\cos(l\phi)$. In the further discussion, these modes are denoted 1 to 4 in the order indicated above. The eigenvalue equation gives the modal indices $n_1 = 1.449$, $n_2 = 1.444$, $n_3 = 1.439$, and $n_4 = 1.437$. We assume that the refractive index is modulated uniformly in the core of the fiber but not at all in the cladding. This is quite realistic since, during fabrication, the fiber usually is made sensitive to UV exposure only in the core. By evaluating (4.11), we find that there will be no coupling between modes with different azimuthal indices l:

$$\boldsymbol{\eta} = \frac{2\pi}{\lambda_0} \begin{bmatrix} 0.957 & 0 & 0 & -0.116 \\ 0 & 0.874 & 0 & 0 \\ 0 & 0 & 0.707 & 0 \\ -0.116 & 0 & 0 & 0.491 \end{bmatrix}. \tag{5.1}$$

There is no coupling to or from modes 2 and 3; thus the grating profile can be found by applying a scalar layer-stripping method separately to the responses associated with these modes. On the other hand, modes 1 and 4 are coupled, so that the multimode layer-stripping method must be applied when using the associated responses as a starting point.

Defining the nominal mode index $n_0 = (n_1 + n_4)/2$, the grating period is set to $\Lambda = \lambda_0/(2n_0)$. The length of the grating is $L = 20$ mm, and $\Delta n_{\mathrm{ac}}(x)$ has the form of a raised cosine window with maximum value $1 \cdot 10^{-3}$. Furthermore, $\Delta n_{\mathrm{dc}}(x)$ is chosen as a sine-modulated Gaussian window with a full-width-at-half-maximum of 7 mm and a maximum value $5 \cdot 10^{-4}$; the period of the sine modulation is 4 mm. The grating is chirped by varying the grating phase according to

$$\frac{\mathrm{d}\theta}{\mathrm{d}x} = \frac{\pi}{8} \cdot 10^4 \left(x - \frac{L}{2} \right) \mathrm{m}^{-1}. \tag{5.2}$$

The reflection matrix as a function of frequency detuning is generated using the piecewise uniform approximation (section 2) with $\Delta x = 10$ μm, which gives $N = 2000$. Zero detuning is taken to be the frequency $f_0 = c/\lambda_0$. Figure 5.1(a) shows the resulting reflection matrix spectrum. The maximum values are $[|R_{11}|, |R_{22}|, |R_{33}|, |R_{44}|, |R_{14}|]_{\max} = [99.6, 99.6, 97.0, 83.0, 28.3]\%$. Note that the large chirp has resulted in significant spectral overlap between the different elements.

The reflection matrix is applied as input to the layer-stripping method. As the modal indices are similar in magnitude, we use the discrete algorithm directly, and $\boldsymbol{\Upsilon}_j$ is calculated by taking into account the factor $(n_p + n_q)/(2n_0)$ as discussed in section 3. Moreover, $\boldsymbol{\kappa}(x)$ and $\boldsymbol{\sigma}(x)$ are calculated by inverting the expressions for $\boldsymbol{\rho}_j$ and $\boldsymbol{\Phi}_j$ in (2.6b) and (2.6c), respectively. Figure 5.1(b) shows $\Delta n_{\mathrm{ac}}(x)$ along with its reconstructed version. The reconstructed $\Delta n_{\mathrm{ac}}(x)$ is calculated by a least square fit to (4.10a) using the diagonal elements of the reconstructed $\boldsymbol{\kappa}(x)$. We find that the error in reconstructed profile is less that $4 \cdot 10^{-6}$ m^{-1}. Also shown is the ac modulation profile calculated using scalar layer-stripping on R_{11}. Due to the strong coupling between mode 1 and 4, the scalar layer-stripping method does not reconstruct the profile accurately. Figures 5.1(c) and (d) show that it is possible to separate the dc index variations $\Delta n_{\mathrm{dc}}(x)$ from the grating phase gradient $\mathrm{d}\theta(x)/\mathrm{d}x$. The separation is based on a least square fit to (4.10b) using the diagonal elements of $\boldsymbol{\sigma}(x)$. The error in reconstructed $\Delta n_{\mathrm{dc}}(x)$ is less than $6 \cdot 10^{-5}$ m^{-1}, while the error in reconstructed $\mathrm{d}\theta(x)/\mathrm{d}x$ is less than 300 m^{-1}. Errors are mainly due to the finite Δx in addition to the fact that the reflection matrix spectrum of the discretized structure is strictly nonperiodic (see the last paragraph of section 3).

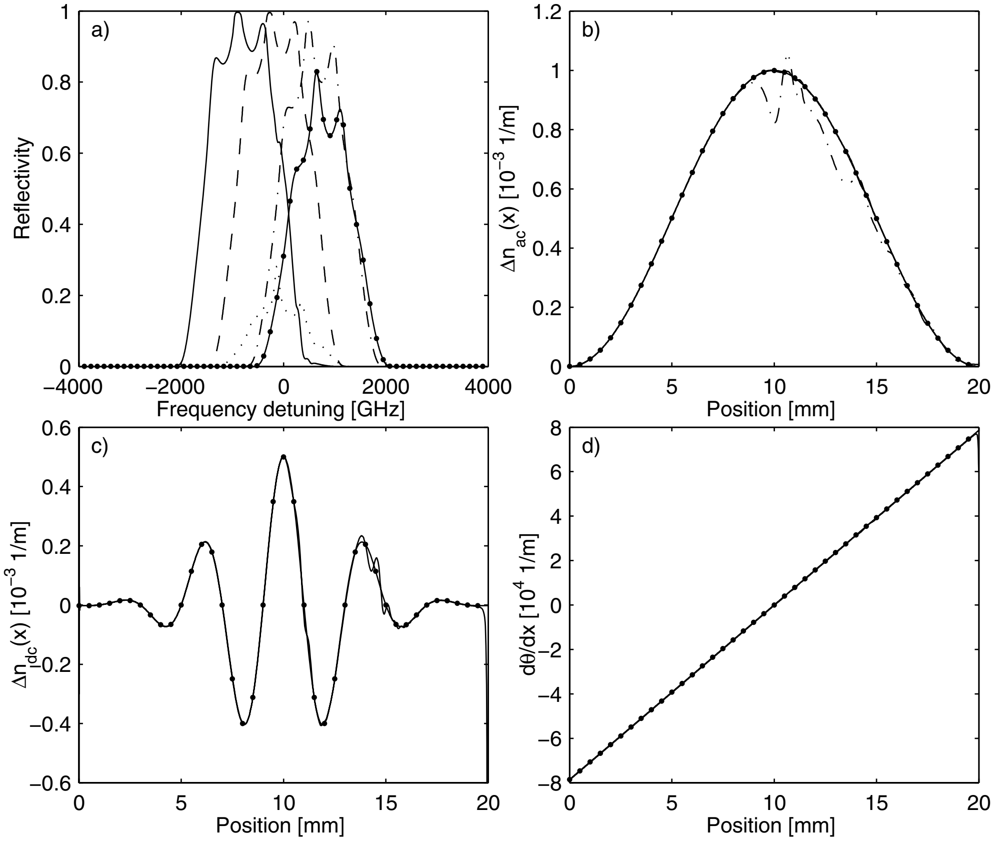


FIG. 5.1. (a) *Magnitude of the reflection spectrum* $|R_{11}|$ *(solid curve),* $|R_{22}|$ *(dashed curve),* $|R_{33}|$ *(dashed-dotted curve),* $|R_{44}|$ *(solid point-marked curve), and* $|R_{14}| = |R_{41}|$ *(dotted curve).* (b) *Reconstructed longitudinal ac modulation* $\Delta n_{\mathrm{ac}}(x)$ *(solid curve), actual ac modulation (solid point-marked curve), and ac modulation calculated using scalar layer-stripping on* R_{11} *(dashed-dotted curve).* (c) *Reconstructed longitudinal dc modulation* $\Delta n_{\mathrm{dc}}(x)$ *(solid curve) and actual dc modulation (solid point-marked curve).* (d) *Reconstructed grating phase gradient* $\mathrm{d}\theta/\mathrm{d}x$ *(solid curve) and actual grating phase gradient (solid point-marked curve).*

6. Analogies to three-dimensional inverse scattering. An important inverse scattering problem is the three-dimensional problem associated with the Schrödinger equation [25]

$$\left\{\nabla^2 + k^2 - V(x,y,z)\right\}\psi(x,y,z;k) = 0, \tag{6.1}$$

where $\psi(x,y,z,k)$ is the wave function and $V(x,y,z)$ is a smooth and nonnegative potential with compact support. In particular, solutions to this problem are applicable to inverse seismic scattering. This problem has been solved using a generalized Marchenko method in [25] and [32], while layer-stripping solutions are suggested in [45] and [43]. Note the close resemblance between (6.1) and (4.4), indicating that a similar method as that in section 4 can be used.

We express the solution as a superposition of the eigenmodes of the Schrödinger equation with $V(x,y,z) = 0$. Writing $\psi(x,y,z;k) = \Psi(y,z;k_y,k_z)\exp(ik_x x)$, these eigenmodes are given by

$$\Psi(y,z;k_y,k_z) = \exp(i(k_y y + k_z z)), \tag{6.2}$$

where k_y and k_z are the wavenumbers in the y-direction and z-direction, respectively, and $k^2 = k_x^2 + k_y^2 + k_z^2$.

In a discrete model, the wavenumbers k_y and k_z can, for example, be discretized in equal intervals Δk, such that $k_y = p\Delta k$ and $k_z = q\Delta k$. In the yz-plane, this means that only a principal range $(-\pi/\Delta k, \pi/\Delta k)$ is considered, and the fields are extended periodically outside this range. The integers p and q are the modal indices satisfying $p^2 + q^2 \leq (k/\Delta k)^2$ for propagating (not evanescent) modes. The modal field profiles are written in normalized form $\Psi_{pq}(y,z) = (\Delta k/2\pi)\Psi(y,z;p\Delta k, q\Delta k)$. The total field $\psi(x,y,z;k)$ is expressed as the superposition

$$\psi(x,y,z;k) = \sum_{p,q} (b_{pq}^{+}(x) + b_{pq}^{-}(x))\Psi_{pq}(y,z), \tag{6.3}$$

where $b_{pq}^{\pm}(x)$ includes all x-dependence of the fields, and ± indicate the sign of k_x, i.e., the propagation direction of the mode.

As in section 4, we insert (6.3) into (6.1), multiply by $\Psi_{pq}^{*}(y,z)$, and integrate over the principal range of the yz-plane. This leads to the coupled mode equations

$$\frac{\mathrm{d}b_{pq}^{+}(x)}{\mathrm{d}x} - ik_{x,pq}b_{pq}^{+}(x) = i\sum_{r,s} \mathcal{C}_{pq,rs}(x)(b_{rs}^{+}(x) + b_{rs}^{-}(x)), \tag{6.4a}$$

$$\frac{\mathrm{d}b_{pq}^{-}(x)}{\mathrm{d}x} + ik_{x,pq}b_{pq}^{-}(x) = -i\sum_{r,s} \mathcal{C}_{pq,rs}(x)(b_{rs}^{+}(x) + b_{rs}^{-}(x)), \tag{6.4b}$$

where the coupling coefficients are given by

$$\begin{aligned} \mathcal{C}_{pq,rs}(x) &= -\frac{1}{2k_x}\int \Psi_{pq}^{*}(y,z)V(x,y,z)\Psi_{rs}(y,z)\mathrm{d}y\mathrm{d}z \\ &= -\frac{1}{2k_x}\left(\frac{\Delta k}{2\pi}\right)^2 \int V(x,y,z)\exp\left[i\Delta k((r-p)y + (s-q)z)\right]\mathrm{d}y\mathrm{d}z, \end{aligned} \tag{6.5}$$

and $k_{x,pq} = [k^2 - (\Delta k)^2(p^2+q^2)]^{1/2}$. We restrict ourselves to the situation where $V(x,y,z)$ is known to be quasi-periodic along the x-direction. Then an expansion of $\mathcal{C}_{pq,rs}(x)$ as in (4.7) together with the transformation (4.8) can be used, resulting in the exact same problem as that described in section 4. Thus the layer-stripping method in section 3 can be applied. The required input data is the reflection into all plane waves upon excitation of the different plane waves onto the plane $x = 0$. The scattering potential $V(x,y,z)$ is found from the inverse of (6.5).

There are two complications. First, in order to use the factorization methods developed in section 3, we must ensure that reciprocity implies symmetric scattering matrices. This is guaranteed when the mode profiles can be written real. Thus we define real mode fields by the transformation

$$\begin{bmatrix} \boldsymbol{\Psi}_{++} \\ \boldsymbol{\Psi}_{-+} \\ \boldsymbol{\Psi}_{+-} \\ \boldsymbol{\Psi}_{--} \end{bmatrix} \rightarrow \mathbf{M} \begin{bmatrix} \boldsymbol{\Psi}_{++} \\ \boldsymbol{\Psi}_{-+} \\ \boldsymbol{\Psi}_{+-} \\ \boldsymbol{\Psi}_{--} \end{bmatrix}, \quad \mathbf{M} = \frac{1}{2}\begin{bmatrix} \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} \\ -i\boldsymbol{I} & +i\boldsymbol{I} & -i\boldsymbol{I} & +i\boldsymbol{I} \\ i\boldsymbol{I} & -i\boldsymbol{I} & +i\boldsymbol{I} & +i\boldsymbol{I}- \\ \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} & -\boldsymbol{I} \end{bmatrix}. \tag{6.6}$$

Here $\boldsymbol{\Psi}_{++}$ denotes a column vector containing the modal field amplitudes Ψ_{pq} with positive p and q, $\boldsymbol{\Psi}_{-+}$ denotes a column vector containing the modal field amplitudes with negative p and positive q, and so forth. The dimension of the identity matrices in the blocks of $\mathbf{M}$ corresponds to the dimension of $\mathbf{\Psi}_{++}$. If $\mathcal{C}$ denotes the matrix formed by the elements $\mathcal{C}_{pq,rs}$, the coupling matrix transforms $\mathcal{C} \to \mathbf{M}^*\mathcal{C}\mathbf{M}^{\mathrm{T}}$. Inspection of (6.5) shows that the transformed $-\mathcal{C}$ is real and positive semidefinite (recall that $V(x,y,z) \geq 0$), thus enabling the factorization method in section 3.

Second, the causality argument of the layer-stripping method works only when the coupling matrix $\mathcal{C}$ is independent on frequency. Equation (6.5) shows that this condition can be justified only when the relevant frequency band is narrow. Therefore the structure must, in addition to being quasi-periodic along the x-direction, vary slowly along the transversal direction. The variation must be sufficiently slow such that the modes with $(p^2+q^2)\Delta k^2 \ll k^2$ contain sufficient information about the transversal dependence, and the other modes may be neglected.

7. Conclusion. A layer-stripping method for the inverse scattering of multimode structures has been proposed. Ambiguities related to factorization of each layer's response into codirectional and contradirectional coupling have been discussed. When there is no codirectional coupling, the ambiguities disappear. Also, when the structure to be reconstructed is smooth, there are important cases with simultaneous co- and contradirectional coupling that can be reconstructed uniquely, provided the reflector eigenvalues are nonzero and nondegenerate. Applications to quasi-periodical structures and analogies to multidimensional inverse scattering have been discussed.

Appendix A. Matrix factorizations.

A.1. Takagi factorization of complex symmetric matrices. Any complex symmtric matrix $\boldsymbol{\Upsilon}$ can be written

$$\boldsymbol{\Upsilon} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{U}, \tag{A.1}$$

where $\boldsymbol{U}$ is unitary and $\boldsymbol{\Sigma}$ is diagonal and nonnegative (see, e.g., [18, Chapter 4.4]). Equation (A.1) is called Takagi factorization.

A constructive proof, suitable for implementation, can be given as follows: Singular value decomposition yields

$$\boldsymbol{\Upsilon} = \boldsymbol{V}_1\boldsymbol{\Sigma}\boldsymbol{V}_2, \tag{A.2}$$

where $\boldsymbol{V}_{1,2}$ are unitary, and $\boldsymbol{\Sigma}$ is diagonal and nonnegative. Using $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}^{\mathrm{T}}$ and $(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\dagger)^{\mathrm{T}} = \boldsymbol{\Upsilon}^\dagger\boldsymbol{\Upsilon}$ we find that $\boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{W}^{\mathrm{T}} = \boldsymbol{\Sigma}\boldsymbol{W}$, where $\boldsymbol{W} \equiv \boldsymbol{V}_2^*\boldsymbol{V}_1$. Thus, provided $\boldsymbol{\Upsilon}$ is nonsingular, $\boldsymbol{W}$ is symmetric. Then $\sqrt{\boldsymbol{W}}$ can be chosen such that it commutes with $\boldsymbol{\Sigma}$ and is symmetric, and we obtain $\boldsymbol{\Upsilon} = \boldsymbol{V}_2^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}_2 = (\sqrt{\boldsymbol{W}}\boldsymbol{V}_2)^{\mathrm{T}}\boldsymbol{\Sigma}\sqrt{\boldsymbol{W}}\boldsymbol{V}_2$, or

$$\boldsymbol{\Upsilon} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{U}, \tag{A.3}$$

where $\boldsymbol{U} \equiv \sqrt{\boldsymbol{W}}\boldsymbol{V}_2$ is unitary and $\boldsymbol{\Sigma}$ is diagonal and positive.

If $\boldsymbol{\Upsilon}$ is singular, we write

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix}, \tag{A.4}$$

where we have arranged $\boldsymbol{\Sigma}$ so that the zero singular values are the last ones, $\boldsymbol{\Sigma}'$ is a diagonal matrix with the nonzero singular values, and $\boldsymbol{W}_{11}$ has the same dimension as $\boldsymbol{\Sigma}'$. We now find $\boldsymbol{\Sigma}'\boldsymbol{W}_{11} = \boldsymbol{W}_{11}\boldsymbol{\Sigma}'$, $\boldsymbol{W}_{12} = \boldsymbol{W}_{21} = \mathbf{0}$, and $\boldsymbol{W}_{11} = \boldsymbol{W}_{11}^{\mathrm{T}}$. The commutation relations do not provide any information on $\boldsymbol{W}_{22}$. Choose $\sqrt{\boldsymbol{W}}$ such that

$$\sqrt{\boldsymbol{W}} = \begin{bmatrix} \sqrt{\boldsymbol{W}_{11}} & \mathbf{0} \\ \mathbf{0} & \sqrt{\boldsymbol{W}_{22}} \end{bmatrix}, \tag{A.5}$$

where $\sqrt{\boldsymbol{W}_{11}}$ is symmetric and $\sqrt{\boldsymbol{W}_{11}}$ and $\boldsymbol{\Sigma}'$ commute. Write $\boldsymbol{\Upsilon} = \boldsymbol{U}_1^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{U}_2$, with

$$\boldsymbol{U}_1 = \sqrt{\boldsymbol{W}}^{\mathrm{T}} \boldsymbol{V}_2 = \begin{bmatrix} \boldsymbol{U}' \\ \boldsymbol{U}_1'' \end{bmatrix}, \tag{A.6}$$

$$\boldsymbol{U}_2 = \sqrt{\boldsymbol{W}} \boldsymbol{V}_2 = \begin{bmatrix} \boldsymbol{U}' \\ \boldsymbol{U}_2'' \end{bmatrix}. \tag{A.7}$$

The matrices $\boldsymbol{U}_1''$ and $\boldsymbol{U}_2''$ are the rows of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ that correspond to the zero singular values, and they do not give any contribution to $\boldsymbol{\Upsilon}$. We may therefore replace the rows $\boldsymbol{U}_1''$ by $\boldsymbol{U}_2''$, which gives $\boldsymbol{U}_1 = \boldsymbol{U}_2 = \boldsymbol{U}$.

The matrix $\boldsymbol{\Sigma}$ is unique up to reordering of the singular values. When the order of the singular values is established, $\boldsymbol{U}$ is unique up to the replacement $\boldsymbol{J}\boldsymbol{U} \to \boldsymbol{U}$, where $\boldsymbol{J}$ is a unitary matrix satisfying $(\boldsymbol{J}\boldsymbol{U})^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{J} \boldsymbol{U} = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{U}$. This leads to $\boldsymbol{J}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{J} = \boldsymbol{\Sigma}$. Assuming the singular values are sorted in, say, descending order, we find that $\boldsymbol{J}$ is a unitary block-diagonal matrix, where each block has a dimension equal to the number of corresponding repeated singular values. For zero singular values, the corresponding block in $\boldsymbol{J}$ is an arbitrary unitary matrix. For repeated nonzero singular values, the corresponding block in $\boldsymbol{J}$ is real. For a distinct, nonzero singular value, the corresponding block of $\boldsymbol{J}$ is either 1 or −1.

A.2. Factorization of a unitary matrix into a symmetric matrix and an orthogonal matrix. A unitary matrix $\boldsymbol{U}$ can be factorized into $\boldsymbol{U} = \boldsymbol{P}\boldsymbol{\Phi}$, where $\boldsymbol{P}$ is a real unitary matrix (orthogonal matrix) and $\boldsymbol{\Phi}$ is a symmetric unitary matrix (see, e.g., [18, Chapter 3.4]). A constructive proof, suitable for implementation, can be given as follows. First we note that the symmetric unitary matrix $\boldsymbol{\Phi}$ can be factorized into $\boldsymbol{\Phi} = \boldsymbol{P}_1 \boldsymbol{D} \boldsymbol{P}_1^{\mathrm{T}}$, where $\boldsymbol{D}$ is a diagonal unitary matrix and $\boldsymbol{P}_1$ is a real unitary matrix (a simple, constructive proof for this particular spectral decomposition is given in [18, Chapter 4.4]). Thus, an equivalent problem is to show that

$$\boldsymbol{U} = \boldsymbol{P}_2 \boldsymbol{D} \boldsymbol{P}_1^{\mathrm{T}}, \tag{A.8}$$

where $\boldsymbol{P}_2 = \boldsymbol{P}\boldsymbol{P}_1$. The decomposition in (A.8) is very similar to singular value decomposition of real matrices, except that $\boldsymbol{D}$ may have complex elements.

The matrix $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ is unitary and symmetric; thus we can write

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{P}_1 \boldsymbol{\Lambda} \boldsymbol{P}_1^{\mathrm{T}}, \tag{A.9}$$

where $\boldsymbol{P}_1$ is a real unitary matrix and $\boldsymbol{\Lambda}$ is a diagonal unitary matrix. Define

$$\boldsymbol{P}_2 = \boldsymbol{U} \boldsymbol{P}_1 \boldsymbol{D}^*, \tag{A.10}$$

where the diagonal matrix $\boldsymbol{D}$ is a solution to $\boldsymbol{D}^2 = \boldsymbol{\Lambda}$. The matrix $\boldsymbol{P}_2$ is unitary since it is produced by multiplication of unitary matrices; thus $\boldsymbol{P}_2^* \boldsymbol{P}_2^{\mathrm{T}} = \boldsymbol{I}$. The matrix is also real since

$$\boldsymbol{P}_2^{\mathrm{T}} \boldsymbol{P}_2 = \boldsymbol{D}^* \boldsymbol{P}_1^{\mathrm{T}} \boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} \boldsymbol{P}_1 \boldsymbol{D}^* = \boldsymbol{D}^* \boldsymbol{P}_1^{\mathrm{T}} \boldsymbol{P}_1 \boldsymbol{D}^2 \boldsymbol{P}_1^{\mathrm{T}} \boldsymbol{P}_1 \boldsymbol{D}^* = \boldsymbol{I}, \tag{A.11}$$

which gives $\boldsymbol{P}_2 = (\boldsymbol{P}_2^*\boldsymbol{P}_2^{\mathrm{T}})\boldsymbol{P}_2 = \boldsymbol{P}_2^*(\boldsymbol{P}_2^{\mathrm{T}}\boldsymbol{P}_2) = \boldsymbol{P}_2^*$.

From (A.10) we therefore conclude that the decomposition in (A.8), with real unitary $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ and diagonal $\boldsymbol{D}$, is always possible. It follows that any unitary matrix can be written $\boldsymbol{U} = \boldsymbol{P}\boldsymbol{\Phi}$, where $\boldsymbol{P}$ is real and unitary, and $\boldsymbol{\Phi}$ is symmetric and unitary. Note that any global phase of $\boldsymbol{P}$ can instead be assigned to $\boldsymbol{\Phi}$, and so without loss of generality we can assume that $\boldsymbol{P}$ is special ($\det \boldsymbol{P} = 1$ and $\det \boldsymbol{\Phi} = \det \boldsymbol{U}$).

Since $\boldsymbol{D}$ is calculated from $\boldsymbol{D}^2 = \boldsymbol{\Lambda}$, the signs of its elements are arbitrary. The ambiguities when determining $\boldsymbol{P}_1$ in (A.9) give rise to ambiguities in $\boldsymbol{P}$ and $\boldsymbol{\Phi}$. The possible $\boldsymbol{P}$ and $\boldsymbol{\Phi}$ can be expressed as $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{P}_1\boldsymbol{J}\boldsymbol{D}^*\boldsymbol{J}^{\mathrm{T}}\boldsymbol{P}_1^{\mathrm{T}}$ and $\boldsymbol{\Phi} = \boldsymbol{P}_1\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}^T\boldsymbol{P}_1^{\mathrm{T}}$ for a real unitary $\boldsymbol{J}$ that commutes with $\boldsymbol{D}^2$. Here $\boldsymbol{P}_1$ is fixed. If the signs of the elements of $\boldsymbol{D}$ are known to be such that any equal elements of $\boldsymbol{D}^2$ correspond to equal elements of $\boldsymbol{D}$, then $\boldsymbol{J}$ commutes with $\boldsymbol{D}$ and can be ignored.

Appendix B. Linear, reciprocal, and lossless components.

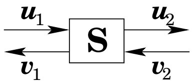


FIG. B.1. *A linear component with P input and P output modes on each side.*

Consider a linear component with P input and P output modes on the left-hand side and P input and P output modes on the right-hand side; see Figure B.1. The component is completely characterized by the $(2P\times 2P)$-dimensional scattering matrix $\mathbf{S}$ which relates the input and output fields:

$$\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{u}_2\end{bmatrix} = \mathbf{S}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{v}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{v}_2\end{bmatrix}. \tag{B.1}$$

The field vectors that propagate to the right and left are denoted $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, and the subscripts 1 and 2 indicate the left- and right-hand sides of the component. The scattering matrix is a block matrix; the blocks $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{22}$ are the reflection from the left- and right-hand sides of the device, respectively, and $\boldsymbol{S}_{21}$ and $\boldsymbol{S}_{12}$ the transmission through the device from the left and right, respectively. These blocks have the dimension $P \times P$.

There exists a similar relation, a transfer matrix relation, that connects the fields on the left-hand side to the fields on the right-hand side:

$$\begin{bmatrix}\boldsymbol{u}_2\\ \boldsymbol{v}_2\end{bmatrix} = \mathbf{T}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{v}_1\end{bmatrix}. \tag{B.2}$$

Comparing (B.1) and (B.2) we find the blocks of $\mathbf{T}$:

$$\mathbf{T} = \begin{bmatrix}\boldsymbol{S}_{21} - \boldsymbol{S}_{22}\boldsymbol{S}_{12}^{-1}\boldsymbol{S}_{11} & \boldsymbol{S}_{22}\boldsymbol{S}_{12}^{-1}\\ -\boldsymbol{S}_{12}^{-1}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}^{-1}\end{bmatrix}. \tag{B.3}$$

To describe a device with a transfer matrix, $\boldsymbol{S}_{12}$ must be invertible; that is, the transmission from the right cannot be zero for any input field vector. Thus, ideal mirrors, for example, cannot be described by a transfer matrix.

Provided the mode profiles can be written real, reciprocity means that the scattering matrix is symmetric [29, 15], i.e.,

$$\boldsymbol{S}_{11} = \boldsymbol{S}_{11}^{\mathrm{T}}, \tag{B.4a}$$

$$\boldsymbol{S}_{22} = \boldsymbol{S}_{22}^{\mathrm{T}}, \tag{B.4b}$$

$$\boldsymbol{S}_{21} = \boldsymbol{S}_{12}^{\mathrm{T}}. \tag{B.4c}$$

Moreover, the lossless condition is expressed as the unitarity condition $\mathbf{S}^{\dagger}\mathbf{S} = \mathbf{I}$:

$$\boldsymbol{S}_{11}^{\dagger}\boldsymbol{S}_{11} + \boldsymbol{S}_{21}^{\dagger}\boldsymbol{S}_{21} = \boldsymbol{I}, \tag{B.5a}$$

$$\boldsymbol{S}_{12}^{\dagger}\boldsymbol{S}_{12} + \boldsymbol{S}_{22}^{\dagger}\boldsymbol{S}_{22} = \boldsymbol{I}, \tag{B.5b}$$

$$\boldsymbol{S}_{12}^{\dagger}\boldsymbol{S}_{11} + \boldsymbol{S}_{22}^{\dagger}\boldsymbol{S}_{21} = \mathbf{0}. \tag{B.5c}$$

With (B.4) in mind, we introduce Takagi factorization of $\boldsymbol{S}_{11}$ and $-\boldsymbol{S}_{22}$ (see section A.1):

$$\boldsymbol{S}_{11} = \boldsymbol{\Phi}_{\mathrm{l}}^{\mathrm{T}}\boldsymbol{\rho}\boldsymbol{\Phi}_{\mathrm{l}}, \tag{B.6a}$$

$$\boldsymbol{S}_{22} = \boldsymbol{\Phi}_{\mathrm{r}}(-\boldsymbol{\rho}')\boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}}, \tag{B.6b}$$

$$\boldsymbol{S}_{21} = \boldsymbol{\Phi}_{\mathrm{r}}\boldsymbol{t}'\boldsymbol{\Phi}_{\mathrm{l}}. \tag{B.6c}$$

Here $\boldsymbol{\Phi}_{\mathrm{l}}$ and $\boldsymbol{\Phi}_{\mathrm{r}}$ are unitary matrices, $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are diagonal and nonnegative, and $\boldsymbol{t}' = \boldsymbol{\Phi}_{\mathrm{r}}^{\dagger}\boldsymbol{S}_{21}\boldsymbol{\Phi}_{\mathrm{l}}^{\dagger}$. By substituting into (B.5) and using (B.4) we obtain

$$\boldsymbol{t}'^{\dagger}\boldsymbol{t}' = \boldsymbol{I} - \boldsymbol{\rho}^{2}, \tag{B.7a}$$

$$\boldsymbol{t}'\boldsymbol{t}'^{\dagger} = \boldsymbol{I} - \boldsymbol{\rho}'^{2}, \tag{B.7b}$$

$$\boldsymbol{\rho}' = \boldsymbol{t}'\boldsymbol{\rho}\boldsymbol{t}'^{*-1}. \tag{B.7c}$$

Introducing the singular value decomposition $\boldsymbol{t}' = \boldsymbol{U}'\boldsymbol{t}\boldsymbol{V}'$, we obtain from (B.7a) that $\boldsymbol{t}^{2} = \boldsymbol{V}'(\boldsymbol{I} - \boldsymbol{\rho}^{2})\boldsymbol{V}'^{\dagger}$, which means $\boldsymbol{t} = \boldsymbol{V}'\sqrt{\boldsymbol{I} - \boldsymbol{\rho}^{2}}\boldsymbol{V}'^{\dagger}$. Backsubstitution shows that $\boldsymbol{t}'$ can be written $\boldsymbol{t}' = \boldsymbol{U}\sqrt{\boldsymbol{I} - \boldsymbol{\rho}^{2}}$ for a unitary $\boldsymbol{U}$; thus (B.7c) reduces to $\boldsymbol{\rho}' = \boldsymbol{U}\boldsymbol{\rho}\boldsymbol{U}^{\mathrm{T}}$. With these properties, it is straightforward to show that (B.6) can be written

$$\boldsymbol{S}_{11} = \boldsymbol{\Phi}_{\mathrm{l}}^{\mathrm{T}}\boldsymbol{\rho}\boldsymbol{\Phi}_{\mathrm{l}}, \tag{B.8a}$$

$$\boldsymbol{S}_{22} = \boldsymbol{\Phi}_{\mathrm{r}}(-\boldsymbol{\rho})\boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}}, \tag{B.8b}$$

$$\boldsymbol{S}_{21} = \boldsymbol{S}_{12}^{\mathrm{T}} = \boldsymbol{\Phi}_{\mathrm{r}}\boldsymbol{t}\boldsymbol{\Phi}_{\mathrm{l}}, \tag{B.8c}$$

where $\boldsymbol{U}$ has been absorbed into $\boldsymbol{\Phi}_{\mathrm{r}}$, $\boldsymbol{\Phi}_{\mathrm{r}}\boldsymbol{U} \to \boldsymbol{\Phi}_{\mathrm{r}}$, and

$$\boldsymbol{t} = \sqrt{\boldsymbol{I} - \boldsymbol{\rho}^{2}}. \tag{B.9}$$

Note that (B.7) implies that $\|\boldsymbol{\rho}\| \leq 1$.

Equations (B.8) and (B.9) can be interpreted as follows: The component can be viewed as a discrete reflector sandwiched between two unitary transmission sections. The discrete reflector provides coupling between equal modes that propagate in opposite directions, and the unitary sections provide coupling between different modes in the same direction. For the discrete reflector, the reflection response from the left and right is $\boldsymbol{\rho}$ and $-\boldsymbol{\rho}$, respectively, and the transmission is $\boldsymbol{t}$. For the two unitary sections, there are no reflections, and the transmission responses from the left are $\boldsymbol{\Phi}_{\mathrm{l}}$ and $\boldsymbol{\Phi}_{\mathrm{r}}$, while the transmission responses from the right are $\boldsymbol{\Phi}_{\mathrm{l}}^{\mathrm{T}}$ and $\boldsymbol{\Phi}_{\mathrm{r}}^{\mathrm{T}}$. Note that this interpretation is consistent with the reciprocity and lossless conditions (B.4) and (B.5) for each of the three sections separately. By inspection, we find that (B.8) is invariant if $\boldsymbol{P}\boldsymbol{\rho}\boldsymbol{P}^{\mathrm{T}} \to \boldsymbol{\rho}$, $\boldsymbol{P}\boldsymbol{t}\boldsymbol{P}^{\mathrm{T}} \to \boldsymbol{t}$, $\boldsymbol{P}\boldsymbol{\Phi}_{\mathrm{l}} \to \boldsymbol{\Phi}_{\mathrm{l}}$, and $\boldsymbol{\Phi}_{\mathrm{r}}\boldsymbol{P}^{\mathrm{T}} \to \boldsymbol{\Phi}_{\mathrm{r}}$, where $\boldsymbol{P}$ is a real unitary matrix. Here $\boldsymbol{P}$ represents an arbitrary rotation of the eigenaxes of the reflector ($\boldsymbol{\rho}$ and $\boldsymbol{t}$ are now real and positive semidefinite).

Using (B.8), the transfer matrix (B.3) can be written

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{A}^* & \boldsymbol{B}^* \\ \boldsymbol{B} & \boldsymbol{A} \end{bmatrix}, \tag{B.10}$$

where the blocks $\boldsymbol{A} = \boldsymbol{\Phi}_r^* \boldsymbol{t}^{-1} \boldsymbol{\Phi}_l^*$ and $\boldsymbol{B} = -\boldsymbol{\Phi}_r^* \boldsymbol{t}^{-1} \boldsymbol{\rho} \boldsymbol{\Phi}_l$ satisfy

$$\boldsymbol{A}^\dagger \boldsymbol{A} - \boldsymbol{B}^{\mathrm{T}} \boldsymbol{B}^* = \boldsymbol{I}, \tag{B.11a}$$

$$\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{A}^{\mathrm{T}} = \mathbf{0}, \tag{B.11b}$$

$$\boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}^* - \boldsymbol{B}^\dagger \boldsymbol{A} = \mathbf{0}. \tag{B.11c}$$

REFERENCES

[1] T. Aktosun, M. Klaus, and C. van der Mee, *Direct and inverse scattering for selfadjoint Hamiltonian systems on the line*, Integral Equations Operator Theory, 38 (2000), pp. 129–171.

[2] V. Bardan, *Comments on dynamic predictive deconvolution*, Geophys. Prosp., 25 (1977), pp. 569–572.

[3] A. Boutet de Monvel and V. Marchenko, *New inverse spectral problem and its application*, in Inverse and Algebraic Quantum Scattering Theory (Lake Balaton, 1996), Lecture Notes in Phys. 488, Springer, Berlin, 1997, pp. 1–12.

[4] J. K. Brenne and J. Skaar, *Design of grating-assisted codirectional couplers with discrete inverse-scattering algorithms*, J. Lightwave Technol., 21 (2003), pp. 254–263.

[5] A. M. Bruckstein and T. Kailath, *Inverse scattering for discrete transmission-line models*, SIAM Rev., 29 (1987), pp. 359–389.

[6] A. M. Bruckstein, B. C. Levy, and T. Kailath, *Differential methods in inverse scattering*, SIAM J. Appl. Math., 45 (1985), pp. 312–335.

[7] L. Dong, L. Reekie, J. L. Cruz, J. E. Caplen, J. P. deSandro, and D. N. Payne, *Optical fibers with depressed claddings for suppression of coupling into cladding modes in fiber Bragg gratings*, IEEE Photonics Technol. Lett., 9 (1997), pp. 64–66.

[8] T. Erdogan, *Cladding-mode resonances in short- and long-period fiber grating filters*, J. Opt. Soc. Amer. A, 14 (1997), pp. 1760–1773.

[9] R. Feced and M. N. Zervas, *Efficient inverse scattering algorithm for the design of grating-assisted codirectional mode couplers*, J. Opt. Soc. Amer. A, 17 (2000), pp. 1573–1582.

[10] R. Feced, M. N. Zervas, and M. A. Muriel, *An efficient inverse scattering algorithm for the design of nonuniform fiber Bragg gratings*, IEEE J. Quantum Electron., 35 (1999), pp. 1105–1115.

[11] V. Finazzi and M. Zervas, *Cladding mode losses in chirped Bragg gratings*, in Bragg Gratings, Photosensitivity, and Poling in Glass Waveguides (BGPP), OSA Technical Digest, Optical Society of America, Washington, DC, 2001, paper BMG16.

[12] M. Foggatt, *Distributed measurement of the complex modulation of a photoinduced Bragg grating in an optical fiber*, Appl. Optics, 35 (1996), pp. 5162–5164.

[13] I. M. Gel'fand and B. M. Levitan, *On the determination of a differential equation from its spectral function*, Amer. Math. Soc. Transl. (2), 1 (1955), pp. 253–304.

[14] F. Ghiringhelli and M. Zervas, *Inverse scattering design of fiber Bragg gratings with cladding mode losses compensation*, in Bragg Gratings, Photosensitivity, and Poling in Glass Waveguides (BGPP), OSA Topic in Optics and Photonics Series 94, Optical Society of America, Washington, DC, 2003, paper TuD2.

[15] H. A. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, Springer, Berlin, 2000.

[16] K. O. Hill, F. Bilodeau, B. Malo, and D. C. Johnson, *Birefringent photosensitivity in monomode optical fibre: Application to external writing of rocking filters*, Electron. Lett., 27 (1991), pp. 1548–1550.

[17] K. O. Hill and G. Meltz, *Fiber Bragg grating technology: Fundamentals and overview*, J. Lightwave Technol., 15 (1997), pp. 1263–1276.

[18] R. A. Horn and C. A. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[19] K. Jinguji and M. Kawachi, *Synthesis of coherent two-port lattice-form optical delay-line circuit*, J. Lightwave Technol., 13 (1995), pp. 73–82.

[20] R. C. Jones, *A new calculus for the treatment of optical systems*, J. Opt. Soc. Amer., 31 (1941), pp. 488–503.

[21] H. Kogelnik, *Theory of Optical Waveguides, Guided-Wave Optoelectronics*, Springer, New York, 1990.

[22] M. Kreĭn, *On a method of effective solution of an inverse boundary problem*, Doklady Akad. Nauk SSSR (N.S.), 94 (1954), pp. 987–990.

[23] H. P. Li, Y. Nakamura, K. Ogusu, Y. L. Sheng, and J. E. Rothenberg, *Influence of cladding-mode coupling losses on the spectrum of a linearly chirped multi-channel fiber Bragg grating*, Optics Express, 13 (2005), pp. 1281–1290.

[24] D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1991.

[25] R. G. Newton, *Inverse scattering.* II. *Three dimensions*, J. Math. Phys., 21 (1980), pp. 1698–1715.

[26] F. Ouellette, *Dispersion cancellation using linearly chirped Bragg grating filters in optical wave-guides*, Optics Lett., 12 (1987), pp. 847–849.

[27] F. Ouellette, P. A. Krug, T. Stephens, G. Dhosi, and B. Eggleton, *Broad-band and WDM dispersion compensation using chirped sampled fiber Bragg gratings*, Electron. Lett., 31 (1995), pp. 899–901.

[28] C. L. Pekeris, *Direct method of interpretation in resistivity prospecting*, Geophysics, 5 (1940), pp. 31–42.

[29] D. M. Pozar, *Microwave Engineering*, Addison–Wesley, Reading, MA, 1993.

[30] Rakesh, *A one-dimensional inverse problem for a hyperbolic system with complex coefficients*, Inverse Problems, 17 (2001), pp. 1401–1417.

[31] E. A. Robinson, *Dynamic predictive deconvolution*, Geophys. Prosp., 23 (1975), pp. 779–797.

[32] J. H. Rose, *The connection between time- and frequency-domain three-dimensional inverse scattering methods*, J. Math. Phys., 25 (1984), pp. 2995–3000.

[33] A. Rosenthal, M. Horowitz, S. Lange, and C. Shäffer, *Experimental reconstruction of a long-period grating from its core-to-core spectrum*, Optics Lett., 30 (2005), pp. 3272–3274.

[34] A. Rosenthal and M. Horowitz, *Inverse scattering algorithm for reconstructing strongly reflecting fiber Bragg gratings*, IEEE J. Quantum Electron., 39 (2003), pp. 1018–1026.

[35] D. Sandel, R. Noé, G. Heise, and B. Borchert, *Optical network analysis and longitudinal structure characterization of fiber Bragg grating*, J. Lightwave Technol., 16 (1998), pp. 2435–2442.

[36] J. Skaar and O. H. Waagaard, *Design and characterization of finite length fiber gratings*, IEEE J. Quantum Electron., 39 (2003), pp. 1238–1245.

[37] J. Skaar, L. Wang, and T. Erdogan, *On the synthesis of fiber Bragg gratings by layer peeling*, IEEE J. Quantum Electron., 37 (2001), pp. 165–173.

[38] A. W. Snyder and J. D. Love, *Optical Waveguide Theory*, Chapman and Hall, New York, 1983.

[39] G.-H. Song, *Toward the ideal codirectional Bragg filter with an acousto-optic-filter design*, J. Lightwave Technol., 13 (1995), pp. 470–480.

[40] G.-H. Song and S.-Y. Shin, *Design of corrugated waveguide filters by the Gel'fand-Levitan-Marchenko inverse-scattering method*, J. Opt. Soc. Amer. A, 2 (1985), pp. 1905–1915.

[41] O. H. Waagaard and J. Skaar, *Synthesis of birefringent reflective gratings*, J. Opt. Soc. Amer. A, 21 (2004), pp. 1207–1220.

[42] L. Wang and T. Erdogan, *Layer peeling algorithm for reconstruction of long-period fibre gratings*, Electron. Lett., 37 (2001), pp. 154–156.

[43] A. E. Yagle, *Differential and integral methods for multidimensional inverse scattering problems*, J. Math. Phys., 27 (1986), pp. 2584–2591.

[44] A. E. Yagle and J. L. Frolik, *On the feasibility of impulse reflection response data for the two-dimensional inverse scattering problem*, IEEE Trans. Antennas and Propagation, 44 (1996), pp. 1551–1564.

[45] A. E. Yagle and B. C. Levy, *Layer-stripping solutions of multidimensional inverse scattering problems*, J. Math. Phys., 27 (1986), pp. 1701–1710.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 334–353

FILTERED BACKPROJECTION INVERSION OF THE CONE BEAM TRANSFORM FOR A GENERAL CLASS OF CURVES[*]

ALEXANDER KATSEVICH[†] AND MIKHAIL KAPRALOV[†]

Abstract. We extend a cone beam transform inversion formula, proposed earlier for helices by one of the authors, to a general class of curves. The inversion formula remains efficient, because filtering is shift-invariant and is performed along a one-parametric family of lines. The conditions that describe the class are very natural. Curves C are smooth, without self-intersections, have positive curvature and torsion, do not bend too much, and do not admit lines which are tangent to C at one point and intersect C at another point. The notions of PI lines and PI segments are generalized, and their properties are studied. The domain U is found, where PI lines are guaranteed to be unique. Results of numerical experiments demonstrate very good image quality.

Key words. shift-invariant filtering, theoretically exact, PI lines

AMS subject classifications. 44A12, 65R10, 92C55

DOI. 10.1137/060673187

1. Introduction.

Image reconstruction from projections is important both in pure mathematics (as a problem of integral geometry) and in applications (as a problem of computed tomography (CT)). Cone beam CT is one of the most common medical imaging modalities. Here one recovers a function $f(x), x \in \mathbb{R}^3$, knowing the integrals of f along lines that intersect a curve C. The curve C is usually called a source trajectory. The ever-increasing needs of medical imaging require the development of inversion algorithms for more and more general source trajectories.

A number of theoretically exact algorithms have been proposed in the past several years. They can be classified into three groups: filtered backprojection (FBP) algorithms, slow-FBP algorithms, and backprojection filtration (BPF) algorithms. Slow-FBP and BPF algorithms are quite flexible, allow some transverse data truncation, and can be used for virtually any complete source trajectory [20, 19, 27, 21, 25, 23, 26]. FBP algorithms are less flexible, but they are by far the fastest and have been developed for a range of source trajectories. They include constant pitch helix [9, 12, 13, 15], dynamic pitch helix [7, 6], circle-and-line [11], circle-and-arc [14, 3], circle-and-helix [2], and saddle [22]. A very nice FBP algorithm was recently proposed by Pack and Noo [20]. It applies to almost any reasonable source trajectory. However, it sometimes leads to excessive detector requirements. The problem is that the algorithm is too general and does not take the geometry of the curve into account. Significant progress has also been achieved in the development of quasi-exact algorithms [1, 16].

With one exception, FBP algorithms have been proposed only for certain types of well-defined trajectories: helices, saddles, etc. There is no FBP algorithm developed specifically for a general class of curves. Ideally, such a class would be described only in terms of some basic geometric properties (e.g., smoothness, curvature, etc.) rather than specifying the types of curves (helices, etc.). In this paper we develop a theoretically exact shift-invariant FBP algorithm for a wide class of source trajectories.

[*] Received by the editors October 24, 2006; accepted for publication (in revised form) July 16, 2007; published electronically November 9, 2007. This research was supported in part by NSF grant DMS-0505494.

http://www.siam.org/journals/siap/68-2/67318.html

[†] Department of Mathematics, University of Central Florida, Orlando, FL 32816-1364 (akatsevi@pegasus.cc.ucf.edu, michael.kapralov@gmail.com).

The conditions describing our class are very natural. We consider curves C that are smooth, have no self-intersections, have positive curvature and torsion, do not bend too much, and do not admit lines which are tangent to C at one point and intersect C at another point. Our algorithm applies to any curve with these properties. The inversion algorithm of this paper is a generalization of the formula proposed for constant- and variable-pitch helices in [9, 12, 7].

The importance of our results is twofold. First, the algorithm can be used in a variety of applications. For example, in electron-beam CT/micro-CT there arise source trajectories that can be described as helices with variable radius and pitch [24]. The FBP algorithms of [9, 12, 7] do not work for such curves, but the new algorithm can easily handle them. Second, the results have theoretical value as well. They provide a deeper understanding of the available algorithms, put them into the context of a more general approach, and demonstrate which geometrical properties the curve is required to have for an efficient FBP algorithm to apply.

The paper is organized as follows. In section 2 we define PI lines for general curves, describe precisely the class of curves considered in the paper, and study properties of their PI segments. In section 3 we find the set U where PI lines are guaranteed to be unique. The result is based on the notions of maximal and minimal PI lines. These critical PI lines can be viewed as a generalization of the axial direction for regular helices. Also we find the special planes, such that the stereographic projection of C onto these planes has very useful properties. In section 4 we study more properties of the PI segments of C. Then the inversion formula is given. Finally, the results of numerical experiments are presented in section 5.

2. PI lines and their properties. The objective of this section is to define PI lines for a general class of smooth curves and study their properties. Let C be a smooth curve:

$$I := [a, b] \ni s \to y(s) \in \mathbb{R}^3, \ |\dot{y}(s)| \neq 0. \tag{2.1}$$

Here and below, the dot above a variable denotes differentiation with respect to s. Define the functions

$$\Phi(s, s_0) := [y(s) - y(s_0), \dot{y}(s), \ddot{y}(s)], \ Q(s, s_0) := [y(s) - y(s_0), \dot{y}(s_0), \dot{y}(s)], \tag{2.2}$$

where $[e_1, e_2, e_3] := e_1 \cdot (e_2 \times e_3)$ denotes the scalar triple product of three vectors. If C is a helix, then Φ and Q are precisely the functions that have been introduced under the same names in [7]. Similarly to [7], it turns out later that Φ is intimately related to the convexity of the projection of C onto a detector plane (cf. (4.6) below), and Q is related to the uniqueness of PI lines (cf. Definitions 2.1 and 2.2, (3.6), and the proof of Proposition 3.3). Given any $s_0, s_1 \in I$, $H(s_0, s_1)$ denotes the line segment with the endpoints $y(s_0), y(s_1) \in C$.

Definition 2.1. *Pick two points $y(s_0), y(s_1) \in C$, $s_0 < s_1$. The line segment $H(s_0, s_1)$ is called a PI segment if $Q(s_0, q) \neq 0$ for any $q \in (s_0, s_1)$.*

Definition 2.2. *Pick two points $y(s_0), y(s_1) \in C$, $s_0 < s_1$. The line segment $H(s_0, s_1)$ is called a maximal PI segment if $Q(s_0, s_1) = 0$, but $Q(s_0, q) \neq 0$ for any $q \in (s_0, s_1)$.*

If C is a helix, Definition 2.1 gives the usual PI segments $H(s, q), 0 < q - s < 2\pi$, and Definition 2.2 gives the maximal PI segments $H(s, s + 2\pi)$.

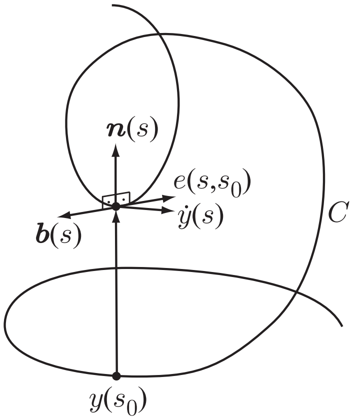


Fig. 1. *Critical case.*

Next we discuss how a smooth curve bends. Consider two points: $y(s_0), y(s) \in C$. Assume $y(s_0)$ is fixed, and $y(s)$ moves along C. The line segment joining $y(s_0)$ and $y(s)$ rotates about the instantaneous axis $e(s,s_0) = (y(s) - y(s_0)) \times \dot{y}(s)/|(y(s) - y(s_0)) \times \dot{y}(s)|$. The point $y(s)$ rotates also about the instantaneous axis, which is obtained by finding the circle of curvature of C at $y(s)$ (also known as the osculating circle). The corresponding axis of rotation is $\mathbf{b}(s)$, i.e., the binormal vector. If $s \to s_0$, then $e(s,s_0) \to \mathbf{b}(s)$. Thus, the difference in directions of the two vectors can measure how much the curve bends between the two points. The maximum possible "bent" occurs when the two axes point in the opposite directions: $e(s,s_0) = -\mathbf{b}(s)$ (see Figure 1).

For the convenience of the reader we recall the definitions of the curvature κ and torsion τ of a smooth curve:

$$\kappa(s) := \frac{|\dot{y}(s) \times \ddot{y}(s)|}{|\dot{y}(s)|^3}, \quad \tau(s) := \frac{[\dot{y}(s), \ddot{y}(s), \dddot{y}(s)]}{|\dot{y}(s) \times \ddot{y}(s)|^2}. \tag{2.3}$$

Now we can formulate the main assumptions on the curve C.

C1. C is smooth, and the curvature and torsion of C are positive;

C2. C does not self-intersect within any PI segment (or a maximal PI segment) of C;

C3. given any PI segment (or a maximal PI segment) $H(s_0,s)$ of C, there is no line tangent to C at $y(s_1)$ and intersecting C at $y(s_2)$, with $s_1, s_2 \in [s_0, s]$, $s_1 \neq s_2$;

C4. C does not bend too much; i.e., given any PI segment (or a maximal PI segment) $H(s_0,s)$ of C, one has $e(s_1,s_2) \neq -\mathbf{b}(s_2)$ for any $s_1, s_2 \in [s_0, s]$, $s_1 \neq s_2$.

If a curve satisfies conditions C1–C4, then its PI segments have a number of nice properties.

Proposition 2.3. *Let C be a curve which satisfies conditions* C1–C4*, and let $H(s_0,s_1)$ be its (possibly maximal) PI segment. Then for any $s,q \in [s_0,s_1]$ one has $\Phi(s,q) > 0$ if $s > q$ and $\Phi(s,q) < 0$ if $s < q$.*

Proof. By shrinking the PI line if necessary, the proposition follows if we show that $\Phi(s,s_0) \neq 0$ for any $s \in (s_0,s_1]$ and $\Phi(s,s_1) \neq 0$ for any $s \in [s_0,s_1)$. We prove only the first statement, because the proof of the second one is analogous.

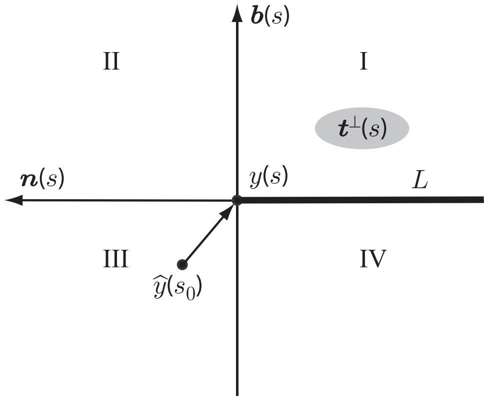


FIG. 2. *Projection of $y(s_0)$ onto the plane through $y(s)$ with normal vector $\dot{y}(s)$.*

Let us assume that the parameterization of $y(s)$ is natural, i.e., $|\dot{y}(s)| \equiv 1$. For convenience, recall the Frenet–Serret formulas:

$$\begin{bmatrix} \dot{\mathbf{t}} \\ \dot{\mathbf{n}} \\ \dot{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{bmatrix}, \tag{2.4}$$

where $\mathbf{t}(s), \mathbf{n}(s), \mathbf{b}(s)$ are the unit tangent and the normal and binormal vectors, respectively, $\kappa(s)$ is the curvature, and $\tau(s)$ is the torsion of the source trajectory (cf. (2.3)). Using (2.4), we get

$$\begin{aligned} \Phi(s, s_0) = [y(s) - y(s_0), \dot{y}(s), \ddot{y}(s)] &= \kappa(s)[y(s) - y(s_0), \mathbf{t}(s), \mathbf{n}(s)] \\ &= \kappa(s)\mathbf{b}(s) \cdot (y(s) - y(s_0)). \end{aligned} \tag{2.5}$$

Since we are interested in the sign of $\Phi(s, s_0)$ and $\kappa(s) > 0$, we determine the sign of

$$\begin{aligned} \mathbf{b}(s) \cdot (y(s) - y(s_0)) &= \int_{s_0}^{s} (\mathbf{b}(t) \cdot (y(t) - y(s_0)))'_t \, dt \\ &= -\int_{s_0}^{s} \tau(t)\mathbf{n}(t) \cdot (y(t) - y(s_0))dt. \end{aligned} \tag{2.6}$$

Let $\mathbf{t}^{\perp}(s)$ denote the plane passing through $y(s)$ and perpendicular to $\mathbf{t}(s)$. We assume that $\mathbf{n}(s)$ and $\mathbf{b}(s)$ are the coordinate axes on the plane, and $y(s)$ is the origin (see Figure 2).

Let $\Pi_{osc}(s)$ denote the osculating plane of C at $y(s)$. Recall that $\Pi_{osc}(s)$ contains $y(s)$ and is parallel to $\dot{y}(s)$ and $\ddot{y}(s)$. If $y(s_0)$ projects onto the ray $L := y(s) - \lambda\mathbf{n}(s), \lambda > 0$, then $y(s_0)$ belongs to $\Pi_{osc}(s)$. Moreover, the two rotation axes—one determined by rotation of $y(s)$ around $y(s_0)$ and the other, $\mathbf{b}(s)$, determined by rotation of $y(s)$ relative to the intrinsic center of rotation—are parallel and point in opposite directions. This is prohibited by the assumption that the curve does not bend too much (see C4), so $y(s_0)$ never projects onto L.

Let $\hat{y}(s_0)$ denote the projection of $y(s_0)$ onto $\mathbf{t}^{\perp}(s)$. The Taylor series expansions show that $\tau > 0$ and $\kappa > 0$ imply

$$\mathbf{n}(t) \cdot (y(s) - y(s_0)) < 0, \ \mathbf{b}(s) \cdot (y(s) - y(s_0)) > 0, \tag{2.7}$$

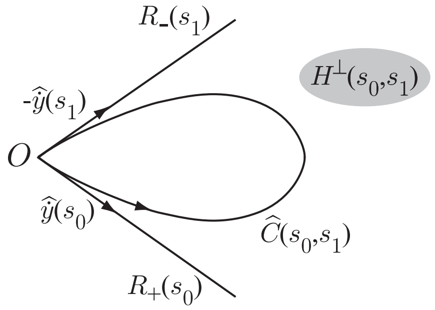


FIG. 3. *Illustration of the containment property: orthogonal projection onto $H^\perp(s_0,s_1)$.*

for $s-s_0>0$ small enough. Hence, initially $\hat{y}(s_0)$ is located in the third quadrant (see Figure 2). Suppose now s increases. If $\hat{y}(s_0)$ appears in the third quadrant, then $\mathbf{n}(t)\cdot(y(t)-y(s_0))<0$. So $\mathbf{b}(s)\cdot(y(s)-y(s_0))$ increases and $\hat{y}(s_0)$ moves down and does not cross the $\mathbf{n}$-axis. If $\hat{y}(s_0)$ appears in the fourth quadrant, then $\mathbf{n}(t)\cdot(y(t)-y(s_0))>0$ and $\mathbf{b}(s)\cdot(y(s)-y(s_0))$ decreases. This implies that in the fourth quadrant $\hat{y}(s_0)$ moves up. However, our assumption precludes $\hat{y}(s_0)$ from crossing L. Consequently, $\hat{y}(s_0)$ never crosses the $\mathbf{n}$-axis and $\Phi(s,s_0)>0$ for any $s\in(s_0,s_1]$. □

Let $H(s_0,s_1)$ be a PI segment (possibly maximal) and $C(s_0,s_1)$ the corresponding curve segment. Project $C(s_0,s_1)$, $\dot{y}(s_0)$, and $\dot{y}(s_1)$ orthogonally onto a plane perpendicular to $H(s_0,s_1)$. Such a plane is denoted $H^\perp(s_0,s_1)$. The corresponding projections are denoted $\hat{C}(s_0,s_1)$, $\hat{\dot{y}}(s_0)$, and $\hat{\dot{y}}(s_1)$, respectively (see Figure 3). Let O be the projection of $H(s_0,s_1)$. Because of condition C3, the vectors $\hat{\dot{y}}(s_0)$ and $\hat{\dot{y}}(s_1)$ determine two rays:

$$\begin{aligned} R_+(s_0,s_1) &:= \{x\in H(s_0,s_1)^\perp : x=O+\lambda\hat{\dot{y}}(s_0),\lambda\ge 0\},\\ R_-(s_0,s_1) &:= \{x\in H(s_0,s_1)^\perp : x=O+\lambda(-\hat{\dot{y}}(s_1)),\lambda\ge 0\}. \end{aligned}\tag{2.8}$$

PROPOSITION 2.4. *Let C be a curve which satisfies conditions* C1–C4. *If $H(s_0,s_1)$ is a (possibly maximal) PI segment of C, then one has the following:*

1. *$\hat{C}(s_0,s_1)$ is contained inside the wedge with vertex O and formed by the rays $R_+(s_0,s_1)$ and $R_-(s_0,s_1)$;*
2. *$\hat{C}(s_0,s_1)$ is smooth, and no line through O is tangent to $\hat{C}(s_0,s_1)$ at an interior point;*
3. *if $H(s_0,s_1)$ is not maximal, the angle between $R_+(s_0,s_1)$ and $R_-(s_0,s_1)$ is less than π. If $H(s_0,s_1)$ is maximal, the angle between the rays equals π;*
4. *no line through O intersects the interior of $\hat{C}(s_0,s_1)$ at more than one point.*

The property of C described in statement (1) of the proposition is important for us, so it will be given the name *containment property.* In other words, statement (1) says that PI segments of curves which satisfy conditions C1–C4 have the containment property.

Proof. To show that $\hat{C}(s_0,s_1)$ is contained inside the wedge, we first consider $\hat{C}(s_0,s_1)$, where $s_1=s_0+\epsilon$ for some $0<\epsilon\ll 1$. As is easily seen, containment follows from the two inequalities:

$$\begin{aligned} &[y(t)-y(s_0),y(s_1)-y(s_0),\dot{y}(s_0)]>0\ \forall t\in(s_0,s_1),\\ &[y(t)-y(s_0),y(s_1)-y(s_0),\dot{y}(s_1)]>0\ \forall t\in(s_0,s_1). \end{aligned}\tag{2.9}$$

To prove the first inequality introduce the function

$$\Psi(s_1,t) := \left[\frac{y(t)-y(s_0)-\dot y(s_0)(t-s_0)}{(t-s_0)^2}, \frac{y(s_1)-y(s_0)-\dot y(s_0)(s_1-s_0)}{(s_1-s_0)^2}, \dot y(s_0)\right]. \tag{2.10}$$

By using the Taylor series expansions we see that $\Psi(s_1,t)$ is smooth and bounded on compact sets. Notice also that

$$\Psi(s_1,s_1)=0,\ \Psi_t'(s_1,t)<\infty. \tag{2.11}$$

Hence $\Psi(s_1,t)/(s_1-t)$ is bounded as well, which implies

$$\begin{aligned}&[y(t)-y(s_0), y(s_1)-y(s_0), \dot y(s_0)]\\ &\qquad = \frac{(t-s_0)^2(s_1-s_0)^2(s_1-t)}{12}([\dot y(s_0), \ddot y(s_0), \dddot y(s_0)] + o(1)) > 0,\end{aligned} \tag{2.12}$$

where $o(1)\to 0$ as $s_1\to s_0$. We used the argument based upon the function Ψ, because we needed an asymptotic result that holds when $t\to s_0$ and when $t\to s_1$. The second inequality in (2.9) can be proven for small $s_1-s_0>0$ in a similar fashion.

Suppose now that s_1-s_0 is not necessarily small. Note that $\hat C(s_0,s_1)$ is tangent to the rays $R_+(s_0,s_1)$ and $R_-(s_0,s_1)$ at the point O of order precisely one. Consider, for example, the ray $R_+(s_0,s_1)$. To determine the order of tangency we need to find the asymptotics of the first expression in (2.9) as $t\to s_0$, with s_0 and s_1 fixed. We have

$$\begin{aligned}&[y(t)-y(s_0), y(s_1)-y(s_0), \dot y(s_0)]\\ &\qquad = [\ddot y(s_0), y(s_1)-y(s_0), \dot y(s_0)]\frac{(t-s_0)^2}{2} + O\left((t-s_0)^3\right)\\ &\qquad = -\Phi(s_0,s_1)\frac{(t-s_0)^2}{2} + O\left((t-s_0)^3\right).\end{aligned} \tag{2.13}$$

Similarly,

$$[y(t)-y(s_0), y(s_1)-y(s_0), \dot y(s_1)] = \Phi(s_1,s_0)\frac{(t-s_1)^2}{2} + O\left((t-s_1)^3\right),\ t\to s_1. \tag{2.14}$$

By Proposition 2.3, $\Phi(s_0,s_1)<0$, $\Phi(s_1,s_0)>0$, and the desired assertion follows.

Suppose $C(s_0,s_1)$ does not have the containment property. Assume, for example, that the first inequality in (2.9) is violated. A violation of the other inequality can be considered analogously. From (2.13) and Proposition 2.3, the inequality holds for some $t>s_0$, where $t-s_0$ is sufficiently small. Thus there exists $t\in(s_0,s_1)$ such that

$$[y(t)-y(s_0), y(s_1)-y(s_0), \dot y(s_0)] = 0. \tag{2.15}$$

Let us show that (2.15) defines t as a function of s_1. Formally differentiating (2.15) with respect to s_1 gives

$$\frac{dt}{ds_1} = -\frac{[y(t)-y(s_0), \dot y(s_1), \dot y(s_0)]}{[\dot y(t), y(s_1)-y(s_0), \dot y(s_0)]}. \tag{2.16}$$

The denominator in (2.16) does not vanish. Otherwise, from the linear independence of $\dot y(s_0)$ and $y(s_1)-y(s_0)$ (property C3) and (2.15) we get $Q(t,s_0)=[y(t)-$

$y(s_0), \dot{y}(s_0), \dot{y}(t)] = 0$. Since $H(s_0, s_1)$ is a PI line, this is a contradiction. Thus (2.15) does define t as a function of s_1, and $C(s_0, s_1)$ does not have the containment property for all s_1 in an open set. Hence we can consider the function $t(s)$ for some $s \le s_1$ using the fact that $Q(t, s_0) \neq 0$ for $t \in (s_0, s_1)$. As s decreases from s_1 towards s_0, one of the following must happen:

(a) $s, t \to s^* \neq s_0$. Replacing s_1 with s and t with $t(s)$ in (2.15) gives $Q(s^*, s_0) = [y(s^*)-y(s_0), \dot{y}(s_0), \dot{y}(s^*)] = 0$, which contradicts the assumption that $H(s_0, s_1)$ is a PI line.

(b) $t \to s_0$, $s \to s^* > s_0$. From (2.15), $\Phi(s_0, s^*) = [y(s_0)-y(s^*), \dot{y}(s_0), \ddot{y}(s_0)] = 0$, which contradicts Proposition 2.3.

Note that $s, t \not\to s_0$ because of (2.12). Thus the containment property is established.

To prove the second statement we argue by contradiction. Suppose there exists $t \in (s_0, s_1)$, where either $\hat{C}(s_0, s_1)$ is nonsmooth or where the line through O and $\hat{y}(t)$ is tangent to $\hat{C}(s_0, s_1)$. Here $\hat{y}(t)$ is the projection of $y(t)$ onto $H^{\perp}(s_0, s_1)$. In both cases

$$[y(s_1) - y(s_0), \dot{y}(t), y(t) - y(s_0)] = 0. \tag{2.17}$$

Just as in the proof of statement (1), (2.17) defines t as a function of s_1. Differentiating (2.17) with respect to s_1 gives

$$\frac{dt}{ds_1} = -\frac{[\dot{y}(s_1), \dot{y}(t), y(t) - y(s_0)]}{[y(s_1) - y(s_0), \ddot{y}(t), y(t) - y(s_0)]}. \tag{2.18}$$

The denominator in (2.18) does not vanish. Otherwise, together with (2.17) this gives $\Phi(t, s_0) = [y(t) - y(s_0), \dot{y}(t), \ddot{y}(t)] = 0$, which contradicts Proposition 2.3. Here we have used the fact that $y(s_1)-y(s_0)$ and $y(t)-y(s_0)$ are not parallel (cf. (2.9)). Hence we can consider the function $t(s)$ for some $s \le s_1$ using the fact that $\Phi(t, s_0) \neq 0$ for $t \in (s_0, s_1]$. As s decreases from s_1 towards s_0, one of the following must happen:

(a) $s, t \to s^* \neq s_0$. Replacing s_1 with s and t with $t(s)$ in (2.17) gives $[y(s^*) - y(s_0), \dot{y}(s^*), \ddot{y}(s^*)] = 0$, which contradicts Proposition 2.3.

(b) $t \to s_0$, $s \to s^* > s_0$. Then (2.17) implies $[y(s^*) - y(s_0), \dot{y}(s_0), \ddot{y}(s_0)] = 0$, which is again a contradiction.

(c) $s, t \to s_0$. Now (2.17) implies $[\dot{y}(s_0), \ddot{y}(s_0), \dddot{y}(s_0)] = 0$, i.e., $\tau(s_0) = 0$. This contradicts the assumption $\tau(s_0) > 0$.

Our argument proves that (2.17) does not happen, so statement (2) is established.

To prove statement (3), first consider $H(s_0, q)$ for $q - s_0 > 0$ sufficiently small. As follows from statements (1) and (2), $\hat{C}(s_0, q)$ is contained between the rays $R_+(s_0, q)$ and $R_-(s_0, q)$, which are close to each other. As q increases towards s_1, the two rays cannot collapse into one. Because of the containment, $\hat{C}(s_0, q)$ is always located between the rays. So if the two rays collapse into one for some $q > s_0$, then $C(s_0, q)$ is a planar curve, which contradicts the assumption $\tau > 0$. Hence $Q(s_0, s_1) = 0$ if and only if $R_+(s_0, s_1)$ and $R_-(s_0, s_1)$ point in opposite directions (see Figure 5).

Statements (1)–(3) imply that (i) whenever a line through O intersects $\hat{C}(s_0, s_1)$, then all of the intersection points (IPs) are on one side of O and (ii) neither $R_+(s_0, s_1)$ nor $R_-(s_0, s_1)$ intersects the interior of $\hat{C}(s_0, s_1)$. By (i) we can replace "line" with "ray" in statement (4). Suppose there is a ray γ with a vertex at O, which intersects $\hat{C}(s_0, s_1)$ at two interior points. Clearly, by rotating γ around O towards either $R_+(s_0, s_1)$ or $R_-(s_0, s_1)$ we can make the two IPs collide. As soon as the IPs collide, we get a ray tangent to $\hat{C}(s_0, s_1)$ at an interior point, which contradicts statement (2). □

Corollary 2.5. *No plane intersects $C(s_0, s_1)$ at more than three points.*

Proof. Suppose there is a plane Π that has at least four IPs with $C(s_0, s_1)$: $s_0 \le t_1 < t_2 < t_3 < t_4 \le s_1$. Consider $C(t_1, t_4)$, and project it onto the plane perpendicular to $H(t_1, t_4)$ (as was done in the proof of Proposition 2.4). As before, let O denote the projection of $H(t_1, t_4)$. The projection of Π gives the line through O which intersects $\hat{C}(t_1, t_4)$ at least at two points, which contradicts statement (3) of Proposition 2.4. □

Corollary 2.6. *Pick any $x \in H(s_0, s_1)$ and $s \in (s_0, s_1)$. Consider a plane Π rotating around the line through x and $y(s)$. The number of IPs of Π and $C(s_0, s_1)$ changes from one to three when Π passes through $H(s_0, s_1)$.*

Proof. Consider the critical case when Π contains $H(s_0, s_1)$. As follows from Proposition 2.4, the vectors $\dot{y}(s_0)$ and $-\dot{y}(s_1)$ point into the opposite half-planes relative to Π. Hence, a small rotation of Π around $\beta(s, x)$ in one direction gives one IP and in the opposite direction three IPs. See section 4 in [15] for more details. □

3. Establishing uniqueness of PI lines.

To establish uniqueness of PI lines, we generalize the standard argument from helices [17, 8, 7] to general curves.

Fix some reconstruction point $x \in \mathbb{R}^3 \setminus C$. For each $s \in I$, fix a vector $N(s)$, $|N(s)| \equiv 1$ (a specific $N(s)$ will be chosen later). Define the functions $q(s)$ and $\lambda(s)$ so that $q(s) > s$, $H(s, q(s))$ is a PI segment, $0 < \lambda(s) < 1$, and the point

$$x(s) := y(s) + \lambda(s)(y(q(s)) - y(s)) \in H(s, q(s)) \tag{3.1}$$

has the property

$$x(s) - x \parallel N(s). \tag{3.2}$$

We assume that the functions $q(s)$ and $\lambda(s)$ with the required properties exist. Later (see (3.11) and the proof of Proposition 3.3) we find an open set U such that for any $x \in U$ the functions $q(s)$ and $\lambda(s)$ do exist.

Condition (3.2) means that the parallel projection of $x(s)$ onto the plane through x with normal vector $N(s)$ coincides with x. Note that the vector-valued function $N(s)$ is determined independently of $q(s)$ and $\lambda(s)$. A similar idea is used in proving the uniqueness of PI lines for the standard helix, the difference being that the vector $N(s)$ is constant and directed along the axis of the helix.

Figure 4 illustrates the setup: The functions $q(s)$ and $\lambda(s)$ are defined in such a way as to ensure that the parallel projection of $x(s)$ onto the plane through x with normal $N(s)$ always coincides with x. Denote $\Delta y(s) := y(q(s)) - y(s)$. Thus,

$$\varepsilon(s) := N(s) \cdot \{(y(s) + \lambda(s)\Delta y(s)) - x\} \tag{3.3}$$

is the signed distance from $y(s) + \lambda(s)\Delta y(s)$ to x, i.e., $\varepsilon(s) = 0$ if and only if the chord $H(s, q(s))$ passes through x. We are interested in calculating $\varepsilon'(s)$.

Combining (3.1)–(3.3) gives

$$y(s) + \lambda(s)(y(q(s)) - y(s)) = x + \varepsilon(s)N(s). \tag{3.4}$$

Differentiate (3.4) with respect to s:

$$\dot{y}(s) + \lambda'(s)\Delta y(s) + \lambda(s)(\dot{y}(q(s))q'(s) - \dot{y}(s)) = \varepsilon'(s)N(s) + \varepsilon(s)\dot{N}(s). \tag{3.5}$$

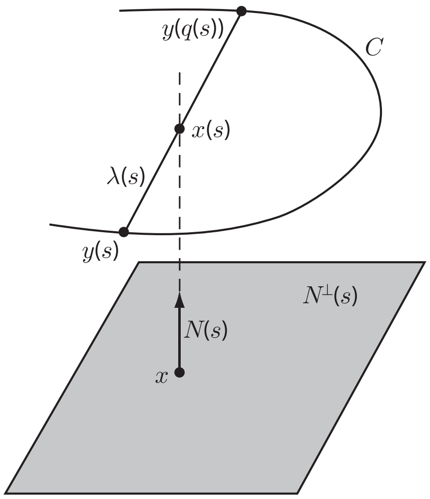


FIG. 4. *Parallel projection onto the plane $N^{\perp}(s)$ through x.*

Computing the dot product of (3.5) with $\Delta y(s)\times\dot y(q)$ on both sides gives the following expression:

$$\begin{aligned}&\varepsilon'(s)=A(s)+\varepsilon(s)B(s),\\ &A(s):=-(1-\lambda(s))\frac{Q(s,q(s))}{[N(s),\Delta y(s),\dot y(q(s))]},\ B(s):=-\frac{[\dot N(s),\Delta y(s),\dot y(q(s))]}{[N(s),\Delta y(s),\dot y(q(s))]},\end{aligned}\tag{3.6}$$

where we have used (2.2).

The goal is to obtain the uniqueness of PI lines. We start by choosing a vector $N(s)$ in such a way as to ensure that the denominator in (3.6) is never zero as long as $H(s,q(s))$ is a PI line. Denote the supremum (respectively, infimum) of all q such that $H(s,q)$ is a PI line by $q_{max}(s)$ (respectively, $q_{min}(s)$). Since $I=[a,b]$ is a compact interval, $q_{max}(s)$ and $q_{min}(s)$ are well-defined.

Now we study the properties of the function $q_{max}(s)$. Pick any $s_0\in(a,b)$ such that $q_{max}(s_0)<b$. Consider the equation

$$Q(q_{max}(s),s)=[y(q_{max}(s))-y(s),\dot y(s),\dot y(q_{max}(s))]=0\tag{3.7}$$

for s in a neighborhood of s_0. In particular, $q_{max}(s)$ satisfies (3.7) when $s=s_0$. Formally differentiating (3.7) with respect to s gives

$$q'_{max}(s_0)=-\frac{[y(q_{max}(s_0))-y(s_0),\ddot y(s_0),\dot y(q_{max}(s_0))]}{[y(q_{max}(s_0))-y(s_0),\dot y(s_0),\ddot y(q_{max}(s_0))]}.\tag{3.8}$$

By assumption C2, $y(q_{max}(s_0))-y(s_0)$ and $\dot y(s_0)$ are not parallel. Hence, if the denominator in (3.8) is zero, together with (3.7) this implies

$$[y(q_{max}(s_0))-y(s_0),\dot y(q_{max}(s_0)),\ddot y(q_{max}(s_0))]=0,\tag{3.9}$$

which contradicts Proposition 2.3. By the implicit function theorem, in a neighborhood of the point $(s_0,q_{max}(s_0))$ there exists a locally unique solution $q(s)$ to (3.7), and this solution satisfies $q(s_0)=q_{max}(s_0)$. By construction, $q(s)<b$ in a neighborhood of s_0. Clearly, $q_{max}(s)$ is continuous at s_0. Otherwise we can find $s_j\to s_0$ such that

$q_j := q_{max}(s_j) \not\to q_{max}(s_0)$. Thus we can choose a subsequence q_{j_k} which converges to some $\bar{q} \neq q_{max}(s_0)$. By the definition of q_{max}, $q_{max}(s_j) \leq q(s_j) < b$, so $Q(q_j, s_j) = 0$. By the continuity of Q, $Q(\bar{q}, s_0) = 0$. Since $\bar{q} \neq q_{max}(s_0)$, by the definition of q_{max} we must have $\bar{q} > q_{max}(s_0)$. This is a contradiction, because, by using the continuity of $q(s)$, there must exist K such that $q_{max}(s_{j_k}) > q(s_{j_k})$ for all $k > K$.

Using (3.7) and assumption C2 gives $\dot{y}(q_{max}(s_0)) = c_1\dot{y}(s_0) + c_2(y(q_{max}(s_0)) - y(s_0))$, $c_1 \neq 0$. Substituting into (3.8) we obtain after simple transformations: $q'_{max}(s_0) = -c_1^2\Phi(s_0, q_{max}(s_0))/\Phi(q_{max}(s_0), s_0)$. By Proposition 2.3, $q'_{max}(s_0) > 0$.

Our argument implies that the function $q_{max}(s)$ has the following properties: (1) $q_{max}(s)$ is continuous on $[a, b]$. If $q_{max}(a) < b$, then there exists $s^*_{max} \in (a, b)$ such that (2) $q_{max}(s) \in C^\infty([a, s^*_{max}])$, $q_{max}(s) < b$ and $q'_{max}(s) > 0$ on $[a, s^*_{max})$, and $q_{max}(s) \equiv b$ on $[s^*_{max}, b)$, and (3) to compute $q_{max}(s)$ on $[a, s^*_{max}]$ we can find $q_{max}(s_0)$ at any $s_0 \in [a, s^*_{max}]$ and then extend it to the entire interval by solving (3.7).

Properties of $q_{min}(s)$ are completely analogous: (1) $q_{min}(s)$ is continuous on $[a, b]$. If $q_{min}(b) > a$, then there exists $s^*_{min} \in (a, b)$ such that (2) $q_{min}(s) \in C^\infty([s^*_{min}, b])$, $q_{min}(s) > a$ and $q'_{min}(s) > 0$ on $(s^*_{max}, b]$, and $q_{min}(s) \equiv a$ on $(a, s^*_{min}]$, and (3) to compute $q_{min}(s)$ on $[s^*_{min}, b]$ we can find $q_{min}(s_0)$ at any $s_0 \in [s^*_{min}, b]$ and then extend it to the entire interval by solving (3.7) (with q_{max} replaced by q_{min}).

From the monotonicity of $q_{max}(s)$ we immediately obtain the following.

PROPOSITION 3.1. *Pick any $s_0 \in [a, b)$. If $s, q \in (s_0, q_{max}(s_0)]$ and $s \neq q$, then $H(s, q)$ is a PI line.*

Proposition 3.1 is a generalization of a similar property for helices: Any line segment connecting two different points within one turn of a helix is a PI line [4, 5, 7]. Note also the following immediate corollary to Proposition 3.1: $q_{min}(q_{max}(s_0)) = s_0$ if $q_{max}(s_0) < b$.

Define

$$(3.10)\quad N_{max}(s) := \frac{y(q_{max}(s)) - y(s)}{|y(q_{max}(s)) - y(s)|},\ N_{min}(s) := \frac{y(q_{min}(s)) - y(s)}{|y(q_{min}(s)) - y(s)|},\ s \in (a, b).$$

Thus, $N_{max}(s)$ (respectively, $N_{min}(s)$) is the unit vector along $H(s, q_{max}(s))$ (respectively, $H(q_{min}(s), s)$).

PROPOSITION 3.2. *Pick any $t \in (s, q_{max}(s))$. One has $[y(t) - y(s), \dot{y}(t), N_{max}(s)] \neq 0$, and the curve segments $C(s, t)$ and $C(t, q_{max}(s))$ are located on opposite sides of the plane containing $H(s, q_{max}(s))$ and $y(t)$. Similarly, pick any $t \in (q_{min}(s), s)$. One has $[y(t) - y(s), \dot{y}(t), N_{min}(s)] \neq 0$, and the curve segments $C(t, s)$ and $C(q_{min}(s), t)$ are located on opposite sides of the plane containing $H(q_{min}(s), s)$ and $y(t)$.*

Proof. We prove only the statements concerning $q_{max}(s)$. The other half of the proposition is completely analogous.

The assertion $[y(t) - y(s), \dot{y}(t), N_{max}(s)] \neq 0$ follows immediately from statement (2) of Proposition 2.4 (see also its proof). This proposition also implies that any line which contains O and passes between the rays $R_+(s, q_{max}(s))$ and $R_-(s, q_{max}(s))$ divides $\hat{C}(s, q_{max})$ into two segments located in the opposite half-planes (see Figure 5). This means that the curve segments $C(s, t)$ and $C(t, q_{max}(s))$ are located on opposite sides of the plane containing $H(s, q_{max}(s))$ and $y(t)$. □

Next we determine the region where PI lines, if exist, are unique. Even though the curve C is well-behaved locally, very little can be said about the global behavior of C. So we choose a "local" piece of C: $I_0 := [a_0, b_0] \subset (a, b)$. The word local is made precise later. For each $s \in I_0$ consider the curve $\hat{C}(s, q_{max})$ in the plane $N^\perp_{max}(s)$. By construction, $\hat{C}(s, q_{max})$ is closed. Let $Cyl_{max}(s)$ be the infinite open cylinder with

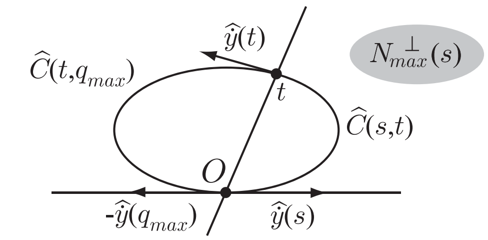


FIG. 5. *Projection onto the plane $N_{max}(s)^{\perp}$ in the case $q_{max}(s) < b$.*

axis $N_{max}(s)$, whose base is the interior of $\hat{C}(s, q_{max})$. In the same fashion we define the cylinders $Cyl_{min}(s)$ using $\hat{C}(q_{min}, s)$ and $N_{min}(s)$. Define U as the intersection of all such open cylinders:

$$U := \cap_{s\in I_0} \left(Cyl_{min}(s) \cap Cyl_{max}(s)\right). \tag{3.11}$$

If the curve turns too much, U can be empty. As an example, imagine a "slinky" toy. Locally it looks like a section of a helix. However if the slinky twists too much and the interval I_0 is sufficiently large, there can be no x that belongs to all of the cylinders. We assume that a sufficiently "local" piece of C is taken, so $U \neq \varnothing$. In other words, the only condition we assume for the interval I_0 is that the set U defined by (3.11) be nonempty. Note that in the case of a helix all cylinders $Cyl_{min}(s)$ and $Cyl_{max}(s)$ are identical, so (3.11) gives the usual domain inside the helix.

PROPOSITION 3.3. *Pick $x \in U$. If x admits a PI line, it is unique in the sense that there is no other PI line with an end point inside I_0.*

Proof. Choose $N(s) := N_{max}(s)$ in (3.2). Since $x \in U$, x projects along $N(s)$ into the interior of $\hat{C}(s, q_{max}(s))$ for any $s \in I_0$. Hence the functions $q(s)$ and $\lambda(s)$ and the map $s \to x(s)$ (cf. (3.1), (3.2)) are well-defined on I_0. By Proposition 3.2, $[\Delta y(s), \dot{y}(q(s)), N(s)] \neq 0$ for any $s \in I_0$. By Proposition 3.1, $H(s, q(s))$ are PI segments, so $Q(s, q(s)) \neq 0$ on I_0. By construction, $\lambda(s) < 1$ on I_0.

Our argument implies that $A(s)$ (cf. (3.6)) is bounded away from zero and of constant sign on I_0. Consider now $B(s)$ (cf. (3.6)). As we already know, the denominator is bounded away from zero. Differentiating (3.10) gives

$$\begin{aligned}\dot{N}_{max}(s) = &\frac{1}{|y(q_{max}(s)) - y(s)|}\\ &\times \left\{[\dot{y}(q_{max}(s))q'_{max}(s) - \dot{y}(s)] - N\left(N \cdot [\dot{y}(q_{max}(s))q'_{max}(s) - \dot{y}(s)]\right)\right\}.\end{aligned} \tag{3.12}$$

By assumption C1, C has no self-intersections, so $|y(q_{max}) - y(s)|$ is bounded away from zero. From (3.8) and the subsequent discussion, it follows that $q'_{max}(s)$ is bounded away from zero. Hence, $\dot{N}_{max}(s)$ is bounded, and $B(s)$ is bounded as well.

From the properties of $A(s)$ and $B(s)$ we get that $\varepsilon(s)$ cannot have more than one root on I_0. This follows immediately from the fact that the signs of $\varepsilon'(s)$ and $A(s)$ in a neighborhood of any s where $\varepsilon(s) = 0$ are the same. Hence x cannot have more than one PI segment with $s_b(x) \in I_0$.

Choosing $N(s) := N_{min}(s)$ in (3.2) and repeating the same argument gives that x cannot have more than one PI segment with $s_t(x) \in I_0$. □

4. Reconstruction algorithm. In order to derive an inversion formula we need to study the curve C some more.

Proposition 4.1. *Let $H(s_0, s_1)$ be a (possibly maximal) PI segment of C. Then $\hat{C}(s_0, s_1)$ has everywhere nonvanishing curvature.*

Proof. Recall that $\hat{C}(s_0, s_1)$ is smooth by Proposition 2.4. Pick any $t \in (s_0, s_1)$. Up to a nonzero factor the curvature of $\hat{C}(s_0, s_1)$ at the point t is given by $[y(s_1) - y(s_0), \dot{y}(t), \ddot{y}(t)]$. Using (2.2) we see that this expression equals $\Phi(t, s_0) - \Phi(t, s_1)$. Since $s_0 < t < s_1$, Proposition 2.3 gives the desired result. We can think of $\Phi(t, s)$ as a signed "distance" from $y(s)$ to $\Pi_{osc}(t)$, so Proposition 2.3 also gives that the line segment $H(s_0, s_1)$ intersects $\Pi_{osc}(t)$ for any $t \in (s_0, s_1)$. □

Corollary 4.2. *Let $H(s_0, s_1)$ be a (possibly maximal) PI segment of C. For any $x \in H(s_0, s_1)$ and $t \in (s_0, s_1)$, the vectors $\dot{y}(t)$ and $x - y(t)$ are not collinear.*

Proof. By Proposition 2.3, $\hat{C}(s_0, q_{max}(s_0))$ is strictly convex. $x \in H(s_0, s_1)$ implies that x projects into the domain bounded by $\hat{C}(s_0, q_{max}(s_0))$. Thus $\dot{y}(t)$ and $x - y(t)$ are not collinear. □

Proposition 4.3. *Let $H(s_0, s_1)$ be a (possibly maximal) PI segment of C. For any $x \in H(s_0, s_1)$ there exists the unique $s^*(x) \in (s_0, s_1)$ such that $x \in \Pi_{osc}(s^*(x))$.*

Proof. As follows from the proof of Proposition 4.1, $\Pi_{osc}(t)$ intersects $H(s_0, s_1)$ for any $t \in [s_0, s_1]$. Hence we can write

$$y(s_0) + \lambda(t)(y(s_1) - y(s_0)) = y(t) + a(t)\dot{y}(t) + b(t)\ddot{y}(t) \tag{4.1}$$

for some scalar functions λ, a, and b. Differentiate (4.1) with respect to t, multiply the resulting equation by $\dot{y}(t) \times \ddot{y}(t)$, and solve for λ':

$$\lambda'(t) = b(t)\frac{[\dot{y}(t), \ddot{y}(t), \dddot{y}(t)]}{[y(s_1) - y(s_0), \dot{y}(t), \ddot{y}(t)]}. \tag{4.2}$$

Since the torsion of C is nonzero, the numerator in (4.2) does not vanish. From the proof of Proposition 4.1, the denominator in (4.2) is nonzero. By Corollary 4.2, $b(t) \neq 0, t \in (s_0, s_1)$. Hence $\lambda(t)$ is a smooth monotone function on $[s_0, s_1]$. Obviously, $\Pi_{osc}(s_0)$ (respectively, $\Pi_{osc}(s_1)$) intersects $H(s_0, s_1)$ at $y(s_0)$ (respectively, $y(s_1)$). Thus $\lambda(s_0) = 0$, $\lambda(s_1) = 1$, and the proposition is proven. □

Due to the containment property (statement (1) of Proposition 2.4), the curve $C(s, q_{max}(s))$ (respectively, $C(s, q_{min}(s))$) is on one side of the plane passing through $y(s)$ and parallel to $\dot{y}(s)$ and $N_{max}(s)$ (respectively, $N_{min}(s)$). This makes it very convenient to project $C(s, q_{max}(s))$ (respectively, $C(s, q_{min}(s))$) onto a plane parallel to $\dot{y}(s)$ and $N_{max}(s)$ (respectively, $N_{min}(s)$). The corresponding projections turn out to be smooth. Let $DP_+(s)$ (respectively, $DP_-(s)$) denote a plane not passing through $y(s)$ and parallel to $\dot{y}(s)$ and $N_{max}(s)$ (respectively, $N_{min}(s)$). We think of $DP_+(s)$ and $DP_-(s)$ as detector planes, so they are chosen on the same side of $y(s)$ as the set U. More precisely, the rays with vertex $y(s)$ passing through U intersect $DP_+(s)$ and $DP_-(s)$. The stereographic projection of $C(s, q_{max}(s))$ onto $DP_+(s)$ is denoted Γ_+, while the stereographic projection of $C(q_{min}(s), s)$ onto $DP_-(s)$ is denoted Γ_-.

Proposition 4.4. *Γ_+ and Γ_- are smooth and have nonvanishing curvature at every point.*

Proof. We consider only Γ_+. The statement about Γ_- is proven analogously. Suppose, for simplicity, that the origin is at $y(s)$ and the equation of $DP_+(s)$ is $x_3 = 1$. Thus, x_1 and x_2 are the coordinates on $DP_+(s)$. Let $x_1(t)$ and $x_2(t)$ be the coordinates of the projection of $y(t), t \in (s, q_{max}(s))$, onto $DP_+(s)$. Then

$$x_1(t) = \frac{y_1(t)}{y_3(t)},\ x_2(t) = \frac{y_2(t)}{y_3(t)}. \tag{4.3}$$

Applying (2.3) to a planar curve gives

$$\kappa(t) = \frac{\dot{x}_1^2}{(\dot{x}_1^2+\dot{x}_2^2)^{3/2}}\left(\frac{\dot{x}_2}{\dot{x}_1}\right)'. \tag{4.4}$$

Differentiating (4.3) gives

$$\begin{aligned}\left(\frac{\dot{x}_2}{\dot{x}_1}\right)' &= \left(\frac{\dot{y}_2y_3-\dot{y}_3y_2}{\dot{y}_1y_3-\dot{y}_3y_1}\right)'\\ &= \frac{(\ddot{y}_2y_3-\ddot{y}_3y_2)(\dot{y}_1y_3-\dot{y}_3y_1)-(\dot{y}_2y_3-\dot{y}_3y_2)(\ddot{y}_1y_3-\ddot{y}_3y_1)}{(\dot{x}_1y_3^2)^2}\\ &= \frac{1}{(\dot{x}_1y_3^2)^2}\begin{vmatrix} y_1 & y_2 & y_3\\ \dot{y}_1 & \dot{y}_2 & \dot{y}_3\\ \ddot{y}_1 & \ddot{y}_2 & \ddot{y}_3\end{vmatrix}.\end{aligned} \tag{4.5}$$

Substituting (4.5) into (4.4) and using (4.3) (recall that $y(s)=0$ is the origin) gives the curvature of Γ_+:

$$\kappa(t) = \frac{\Phi(t,s)}{y_3^4(t)\,(\dot{x}_1^2(t)+\dot{x}_2^2(t))^{3/2}}. \tag{4.6}$$

By the properties of $C(s,q_{max}(s))$ mentioned prior to this proposition, $y_3(t)\neq 0, t\in (s,q_{max}(s))$. Also, $y_3(s)=0$, and, if $H(s,q_{max}(s))$ is maximal, $y_3(q_{max}(s))=0$. It remains to show that $\dot{x}_1^2(t)+\dot{x}_2^2(t)\neq 0$. This would also imply that Γ_+ is smooth. We argue by contradiction. Suppose $\dot{x}_1(t)=\dot{x}_2(t)=0$. Then $\dot{y}_2y_3=\dot{y}_3y_2$, $\dot{y}_1y_3=\dot{y}_3y_1$. Consequently, $y(t)\times\dot{y}(t)$ is parallel to the x_3-axis. Thus, either both $y(t)$ and $\dot{y}(t)$ are parallel to $DP_+(s)$ or $y(t)$ and $\dot{y}(t)$ are parallel to each other. Both cases are impossible because of the convexity of $\hat{C}(s,q_{max}(s))$ (cf. Proposition 4.1). Since $\Phi(t,s)\neq 0$ for $t\in[s,q_{max}(s)]$ (cf. Proposition 2.3), the desired assertion is proven. □

Denote $L_0^+ := DP_+(s)\cap\Pi_{osc}(s)$. It is clear that L_0^+ is an asymptote of Γ_+: $\text{dist}(\hat{y}(t),L_0^+)\to 0$ as $t\to s^+$. Similarly, $L_0^- := DP_-(s)\cap\Pi_{osc}(s)$ is an asymptote of Γ_-: $\text{dist}(\hat{y}(t),L_0^-)\to 0$ as $t\to s^-$.

Fix $x\in U$, which admits a PI line. Let $I_{PI}(x)=[s_b(x),s_t(x)]$ be the PI interval of x. Let $\hat{x}$ denote the projection of x onto a detector plane. Frequently it is convenient to identify detector planes by introducing systems of coordinates that depend smoothly on s. This allows one to identify all $DP_+(s)$ and, separately, all $DP_-(s)$. Since $x\in U$, x does not belong to any plane passing through $y(s)$ and parallel to $DP_+(s)$ or $DP_-(s)$, where $s\in I_{PI}(x)$. Hence Propositions 4.3 and 3.3 immediately imply the following statement.

Corollary 4.5. *As s moves along $I_{PI}(x)$, the point $\hat{x}$ traces smooth curves on $DP_+(s)$ and $DP_-(s)$. $\hat{x}$ is between $\Gamma_+(s)$ and L_0^+ on $DP_+(s)$ if and only if $s\in(s_b(x),s^*(x))$, and $\hat{x}$ is between L_0^- and $\Gamma_-(s)$ on $DP_-(s)$ if and only if $s\in(s^*(x),s_t(x))$.*

Loosely speaking, Corollary 4.5 can be stated as follows: $\hat{x}$ is between $\Gamma_+(s)$ and $\Gamma_-(s)$ if and only if $s\in I_{PI}(x)$.

Following [9, 12], choose any $\psi\in C^\infty(\mathbb{R}^+)$ with the properties

$$\begin{aligned}&\psi(0)=0;\ 0<\psi'(t)<1,\, t\geq 0,\\ &\psi'(0)=0.5;\ \psi^{(2k+1)}(0)=0,\, k\geq 1.\end{aligned} \tag{4.7}$$

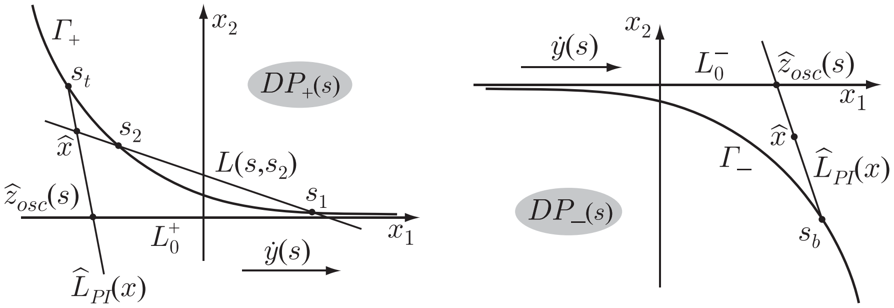


FIG. 6. *Detector planes $DP_+(s)$ (left panel) and $DP_-(s)$ (right panel).*

Suppose s, s_1, and s_2 are related by

$$s_1 = \begin{cases} \psi(s_2 - s) + s, & s_2 \geq s, \\ \psi(s - s_2) + s_2, & s_2 < s. \end{cases} \tag{4.8}$$

From (4.7), $s_1 = s_1(s, s_2)$ is a C^∞ function of s and s_2. Conditions (4.7) are easy to satisfy. One can take, for example, $\psi(t) = t/2$, and this leads to

$$s_1 = (s + s_2)/2. \tag{4.9}$$

Denote also

$$\begin{gathered} u(s, s_2) = \frac{(y(s_1) - y(s)) \times (y(s_2) - y(s))}{|(y(s_1) - y(s)) \times (y(s_2) - y(s))|} \operatorname{sgn}(s_2 - s), \\ q_{min}(s) < s_2 < q_{max}(s), s_2 \neq s, \\ u(s, s_2) = \frac{\dot{y}(s) \times \ddot{y}(s)}{|\dot{y}(s) \times \ddot{y}(s)|}, \ s_2 = s. \end{gathered} \tag{4.10}$$

In the same way as in [12], we prove that $u(s, s_2)$ is a C^∞ vector function of its arguments. Let $\Pi(s, s_2)$ be the plane through $y(s)$, $y(s_2)$, and $y(s_1(s, s_2))$. The intersection of $\Pi(s, s_2)$ with $DP_+(s)$ if $s < s_2 < q_{max}(s)$ or with $DP_-(s)$ if $q_{min}(s) < s_2 < s$ is called a filtering line and denoted $L(s, s_2)$.

Fix $x \in U$, which admits a PI line, and $s \in I_{PI}(x)$. Find $s_2 \in I_{PI}(x)$ such that $\Pi(s, s_2)$ contains x. More precisely, we have to solve for s_2 the following equation:

$$(x - y(s)) \cdot u(s, s_2) = 0, \ s_2 \in I_{PI}(x). \tag{4.11}$$

Recall that $\dot{y}(s)$ is parallel to $DP_+(s)$ and $DP_-(s)$. For convenience, we choose the x_1- and x_2-axes so that

1. $\dot{y}(s)$ and the x_1-axis are parallel and point in the same direction;
2. the equation of $\Pi_{osc}(s)$ is $x_2 = 0$;
3. on $DP_+(s)$, Γ_+ is located in the half-plane $x_2 > 0$;
4. on $DP_-(s)$, Γ_- is located in the half-plane $x_2 < 0$.

Figure 6 illustrates the two detector planes.

The advantage of planes $DP_+(s)$ and $DP_-(s)$ is that the segments $C(s, q_{max}(s))$ and $C(q_{min}(s), s)$ are projected onto them as continuous curves with positive curvature. If C is a helix, the two segments become the usual 2π-segments $C(s, s+2\pi)$ and $C(s - 2\pi, s)$. This makes it very convenient when describing how to choose filtering lines in a shift-invariant FBP algorithm. On the other hand, the disadvantage is that the two segments are projected onto two different planes. This makes it difficult to adapt the proofs from [12, 9] to the present more general situation. Fortunately, the difficulty can be resolved. Given $x \in U$ with the PI interval $I_{PI}(x) = [s_b(x), s_t(x)]$, we can find a family of "detector planes" such that for any $s \in I_{PI}(x)$ the entire PI segment of x, $C(s_b(x), s_t(x))$, projects onto them in exactly the same way as in the case of a regular constant-pitch helix. There is no guarantee that the larger segment $C(q_{min}(s), q_{max}(s))$ (which is equivalent to two adjacent turns of a helix) projects well onto the planes, but this is not needed.

Let $DP(s), s \in I_{PI}(x)$, be a plane not passing through $y(s)$ and parallel to $\dot{y}(s)$ and $N_{max}(s_b(x))$. Using the convexity of $\hat{C}(s_b(x), s_t(x)) \subset \hat{C}(s_b(x), q_{max}(s_b(x)))$ (cf. Proposition 4.1 and Figure 5) and repeating the proof of Proposition 4.4, we establish that the stereographic projection of $C(s_b(x), s_t(x))$ onto $DP(s)$ has all of the usual properties as in the constant-pitch helix case. More precisely, the projections of $C(s_b(x), s)$ and $C(s, s_t(x))$ are concave down and up, respectively, they share the usual asymptote $DP(s) \cap \Pi_{osc}(s)$, they are located on the opposite sides of the latter, etc. Thus, using the same argument as in [12, 7], we immediately obtain the following result.

Proposition 4.6. *The solution s_2 to* (4.11) *exists, is unique, and depends smoothly on s.*

The following result shows that filtering lines are shared by sufficiently many points $x \in U$. The planes $DP(s)$ used for the proof of Proposition 4.6 are selected separately for each x, so they do necessarily work for all x in a large subset of U. Thus we have to go back to the planes $DP_+(s)$ and $DP_-(s)$.

Proposition 4.7. *All $x \in U$ that project onto any line $L(s, s_2), s < s_2 < q_{max}(s)$, on $DP_+(s)$ to the left of s_2 or onto $L(s, s_2), q_{min}(s) < s_2 < s$, on $DP_-(s)$ to the right of s_2 share $L(s, s_2)$ as their filtering line.*

Proof. We consider only the case when $s_2 > s$, i.e., $\hat{x} \in DP_+(s)$. The other case can be considered analogously. We have $s_t(x) \in \Gamma_+$. By Corollary 4.5, $\hat{x}$ appears between L_0^+ and Γ_+. From the proof of Proposition 4.1, $\Pi_{osc}(s)$ intersects the PI segment of x, $H(s_b(x), s_t(x))$. Let $z_{osc}(s)$ denote the point of intersection. Let $\Pi_{max}(s)$ be the plane through $y(s)$ and parallel to $\dot{y}(s)$ and $N_{max}(s)$. Our first goal is to show that the line segment $[z_{osc}(s), y(s_t(x))]$ lies on one side of $\Pi_{max}(s)$ and, therefore, projects well onto $DP_+(s)$. Let $z_{max}(s)$ denote the intersection of the line through $L_{PI}(x)$ and $\Pi_{max}(s)$. Clearly, $z_{osc}(s) = z_{max}(s)$ when $s = s_b(x)$. From the proof of Proposition 4.3, $z_{osc}(s)$ moves toward $y(s_t(x))$ along $L_{PI}(x)$ as s increases from $s_b(x)$ to $s_t(x)$. From the convexity of $\hat{C}(s, q_{max}(s))$ (cf. Figure 5), it is easy to obtain that in a neighborhood of $s = s_b(x)$ the point $z_{max}(s)$ moves away from $H(s_b(x), s_t(x))$ as s increases. If for some $s \in (s_b(x), s_t(x))$ the points $z_{osc}(s)$ and $y(s_t(x))$ are on opposite sides of $\Pi_{max}(s)$, then the point $z_{max}(s)$ enters the line segment $[z_{osc}(s), y(s_t(x))]$ for some $s = s_0 \in (s_b(x), s_t(x))$. Hence, either (i) $z_{osc}(s_0) = z_{max}(s_0)$ or (ii) $y(s_t(x)) = z_{max}(s_0)$. From Proposition 2.3, $[y(q_{max}(s_0)) - y(s_0), \dot{y}(s_0), \ddot{y}(s_0)] \neq 0$, so (i) implies that $z_{osc}(s_0) - y(s_0)$ and $\dot{y}(s_0)$ are collinear, which contradicts Corollary 4.2. In case (ii), $y(s_t(x)) \in \Pi_{max}(s_0)$, which contradicts the containment property.

Hence $\hat{L}_{PI}(x)$, the projection of $H(s_b(x), s_t(x))$ onto $DP_+(s)$, intersects L_0^+. More precisely, the projection of the line segment $[z_{osc}(s), y(s_t(x))] \subset L_{PI}(x)$ is a continuous line segment that connects Γ_+ and L_0^+ (see Figure 6). Note that Proposition 4.3 implies $x \in [z_{osc}(s), y(s_t(x))]$ if $s < s^*(x)$. It turns out that $\hat{L}_{PI}(x)$ does not intersect Γ_+ at any point other than $s_t(x)$. Suppose there is an additional intersection point t. Thus the plane through $y(s)$ and $H(s_b(x), s_t(x))$ intersects $C_{PI}(x)$ at four points: $s_b(x), t, s,$

and $s_t(x)$, and this contradicts Corollary 2.5. Here $C_{PI}(x)$ is the section of the curve corresponding to the parametric interval $I_{PI}(x)$, i.e., $C_{PI}(x) := C(s_b(x), s_t(x))$.

If x projects onto $L(s, s_2)$ to the left of s_2, we make two observations: (i) $\hat{x}$ is between L_0^+ and Γ_+ on $DP_+(s)$, and (ii) $s_2 < s_t(x)$ (due to the properties of $\hat{L}_{PI}(x)$ that we just established). From (i) and Corollary 4.5, $s \in I_{PI}(x)$. From (ii), $s_2 \in (s, s_t(x))$, so by (i) $s_2 \in I_{PI}(x)$. By construction, s_2 was chosen to satisfy $(x - y(s)) \cdot u(s, s_2) = 0$. We have just shown that $s, s_2 \in I_{PI}(x)$. This proves that $L(s, s_2)$ is the filtering line for x. □

By Proposition 4.7, our construction defines $s_2 := s_2(s, x)$ and, consequently, $u(s, x) := u(s, s_2(s, x))$. Let $D_f(s, \Theta) = \int_0^\infty f(y(s) + t\Theta)dt$, $|\Theta| = 1$, denote the cone beam transform of f. The main result of the paper is the following theorem.

THEOREM 4.8. *Let C be a curve* (2.1)*, which satisfies conditions* C1–C4. *Let $I_0 \subset I$ be an interval such that the set U defined by* (3.11) *is nonempty. For any $f \in C_0^\infty(U)$ and $x \in U$ which admits a PI line such that $I_{PI}(x) \subset I_0$, one has*

$$f(x) = -\frac{1}{2\pi^2} \int_{I_{PI}(x)} \frac{1}{|x - y(s)|} \int_0^{2\pi} \frac{\partial}{\partial q} D_f(q, \Theta(s, x, \gamma)) \bigg|_{q=s} \frac{d\gamma}{\sin\gamma} ds, \tag{4.12}$$

where $\beta(s, x) = (x - y(s))/|x - y(s)|$, $e(s, x) := \beta(s, x) \times u(s, x)$, and $\Theta(s, x, \gamma) := \cos\gamma\beta(s, x) + \sin\gamma e(s, x)$.

Proof. Corollaries 2.5, 2.6, and 4.5 and Propositions 4.1, 4.3, 4.4, and 4.6 imply that locally, i.e., in a neighborhood of $I_{PI}(x)$, the curve C behaves in essentially the same way as the usual helix. Hence the same argument as in [12, 7] can be used to prove that (4.12) holds. For the convenience of the reader we recall the key steps in the proof. Fix $x \in U$ such that $I_{PI}(x) \subset I_0$ and $s \in I_{PI}(x)$. Suppose $s < s^*(x)$ (cf. Proposition 4.3). Consider the detector plane $DP(s)$ introduced prior to Proposition 4.6. Let d_0 be the unit vector perpendicular to $DP(s)$ and pointing from the source position $y(s)$ towards the detector. By the choice of U and the detector plane, this implies $\beta(s, x) \cdot d_0 > 0$ and $\dot{y}(s) \cdot d_0 = 0$. In the same way as in [10] we show that the stereographic projection of all of the relevant vectors onto $DP(s)$ preserves the sign of dot products. Let Π be a generic plane containing x and $y(s)$. By Corollary 2.5, there can be only either one or three IPs in the set $\Pi \cap C_{PI}(x)$. If s is the only IP or the middle of the three IPs, then the slope of the line $\Pi \cap DP(s)$ on $DP(s)$ is greater than the slopes of L_0 $(= \Pi_{osc}(s) \cap DP(s))$ and the filtering line through $\hat{x}$. Recall that the filtering line intersects the projection of $C(s, s_t(x))$ onto $DP(s)$ twice, and the projected curve is convex. Hence the IP s gets weight 1. Since $s < s^*(x)$, the only remaining alternative is that s is the smallest of the three IPs. In this case s gets weight 1 or -1 depending on the location of $\Pi \cap DP(s)$ relative to the filtering line through $\hat{x}$. Using exactly the same argument as in section 3 of [12], we prove that the largest of the three IPs gets weight -1 or 1, respectively. Hence the inversion formula (4.12) is exact. □

Proposition 4.7 implies that (4.12) is of the efficient shift-invariant FBP form. This means that filtering in (4.12) is convolution-based and is performed along a one-parametric family of lines.

5. Numerical experiments.

Numerical experiments are conducted using flat detector geometry. The simulation parameters are summarized in Table 1. The detector is located at the distance of 600 mm from the axis of rotation opposite to the source position. The algorithm is implemented in the native coordinates following [18]. The clock phantom (see, e.g., [7]) is chosen for reconstructions. The background cylinder is at 0 HU, the spheres are at 1000 HU, and the air is at -1000 HU.

Table 1
Simulation parameters.

Parameter	Value	Units
Views per rotation	1000	
Number of detector columns	1101	
Number of detector rows	111	
Actual detector pixel size	1 × 1	mm^2
Isocenter to detector distance	600	mm

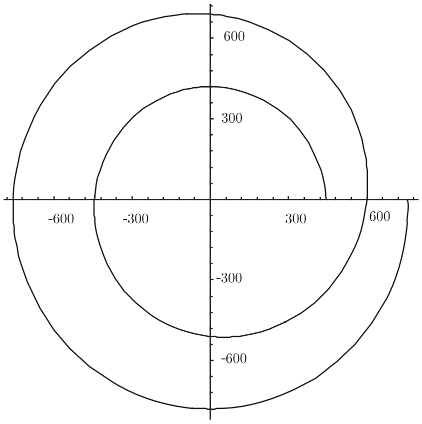


Fig. 7. *Projection of the source trajectory in* (5.1) *onto the xy-plane.*

Two source trajectories have been used. The first one is a variable radius helix given by the formula:

$$y(s) = \left(R(s)\cos s, R(s)\sin s, \frac{h_0}{2\pi}s \right), \;\; R(s) = R(1 + 0.3\sin(s/3)), \tag{5.1}$$

where $R = 600$ mm, and the table feed per turn is $h_0 = 35$ mm. The projection of this trajectory onto the plane $x_3 = 0$ for $s \in [-2\pi, 2\pi]$ is shown in Figure 7.

Because of the variable radius, detector usage is different for different source positions. The detector parameters given in Table 1 were chosen so as to accommodate all source positions. The horizontal size of the detector varied from 894 to 1098 mm; the average was 945 mm. The vertical size of the detector varied from 49 to 75 mm; the average was 59 mm. These values were calculated for the segment of the trajectory necessary to reconstruct the clock phantom.

The boundary of the set U is calculated according to (3.11). The cross section of the boundaries of cylinders $Cyl_{min}(s)$ and $Cyl_{max}(s)$ with the plane $x_3 = 0$ is shown in Figure 8 (left panel). The solid circle of radius $r = 240$ mm shows the boundary of the clock phantom, and the dashed circle is of the maximum radius $r \approx 374$ mm that fits inside the cross section of U. The result of reconstruction is shown in Figure 9. Here and in the experiment below we use voxels of size 1mm in each direction.

The second experiment is carried out using the variable-radius and variable-pitch helix given by:

$$y(s) = \left(R(s)\cos s, R(s)\sin s, \frac{h(s)}{2\pi}s \right), \;\; h(s) = h_0\left(1 + \frac{\sin(s/2)}{s}\right). \tag{5.2}$$

Here $R(s)$ and h_0 are the same as in (5.1). Because of the variable radius/variable pitch, detector usage is different for different source positions. The horizontal size

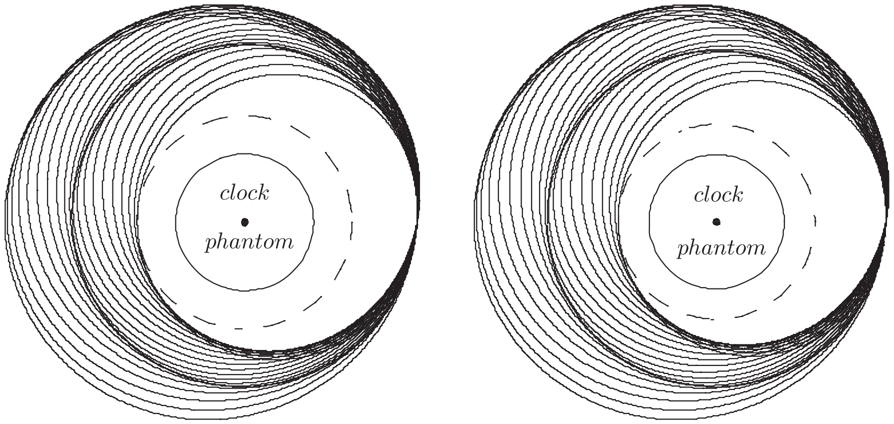


Fig. 8. *Cross section of boundaries of cylinders $Cyl(s)$ from* (3.11) *for trajectory* (5.1) *(left panel) and trajectory* (5.2) *(right panel).*

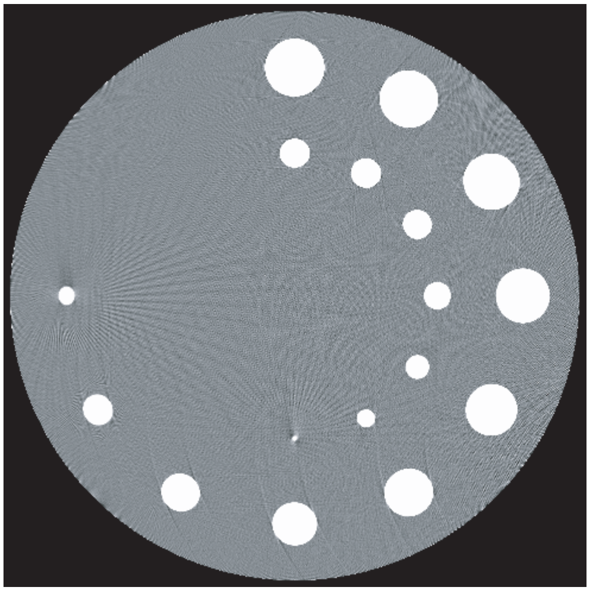


Fig. 9. *Reconstruction of the clock phantom from trajectory* (5.1)*: slice $x_3 = 0$, $WL = 0$* HU*, $WW = 100$* HU*.*

of the detector varied from 894 to 1098 mm; the average was 945 mm. The vertical size of the detector varied from 38 to 88 mm; the average was 71 mm. These values were calculated for the segment of the trajectory necessary to reconstruct the clock phantom. The cross section of the boundaries of cylinders $Cyl_{min}(s)$ and $Cyl_{max}(s)$ with the plane $x_3 = 0$ is shown in Figure 8 (right panel). Again, the solid circle of radius $r = 240$ mm shows the boundary of the clock phantom, and the dashed circle is of the maximum radius $r \approx 348$ mm that fits inside the cross section of U. The results of the reconstruction are shown in Figure 10.

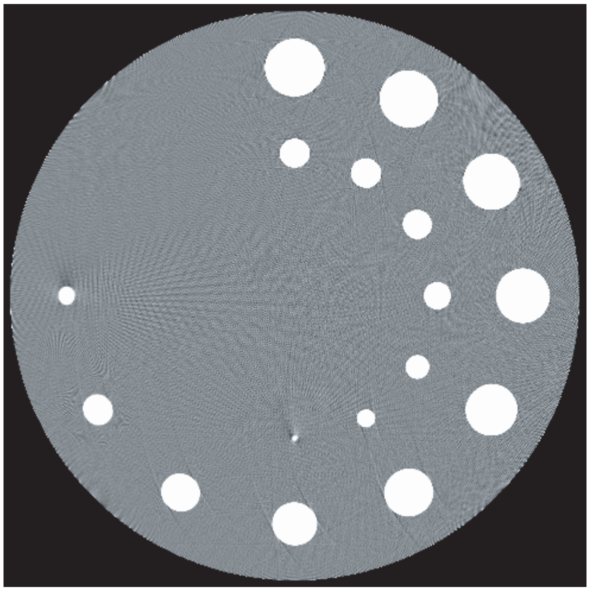


Fig. 10. *Reconstruction of the clock phantom from trajectory* (5.2)*: slice $z = 0$, $WL = 0$* HU*, $WW = 100$* HU*.*

REFERENCES

[1] C. Bontus, T. Köhler, and R. Proksa, *EnPiT: Filtered back-projection algorithm for helical CT using an n-Pi acquisition*, IEEE Trans. Medical Imaging, 24 (2005), pp. 977–986.

[2] C. Bontus, P. Koken, T. Köhler, and R. Proksa, *Circular CT in combination with a helical segment*, Physics in Medicine and Biology, 51 (2007), pp. 107–120.

[3] G.-H. Chen, T. Zhuang, B. E. Nett, and S. Leng, *A showcase of exact cone-beam image reconstruction algorithms for circle-based trajectories*, in Proceedings of the Eighth International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, F. Noo, L. G. Zeng, and H. Kudo, eds., Salt Lake City, Utah, 2005, pp. 295–299. Online access: www.ucair.med.utah.edu/3D05/proceedings.html.

[4] P. E. Danielsson, P. Edholm, J. Eriksson, and M. M. Seger, *Towards exact reconstruction for helical cone-beam scanning of long objects. A new detector arrangement and a new completeness condition*, in Proceedings of the 1997 Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine, Pittsburgh, D. W. Townsend and P. E. Kinahan, eds., 1997, Institute of Physics, Bristol, pp. 141–144.

[5] M. Defrise, F. Noo, and H. Kudo, *A solution to the long-object problem in helical cone-beam tomography*, Physics in Medicine and Biology, 45 (2000), pp. 623–643.

[6] M. Kapralov and A. Katsevich, *A one-PI algorithm for helical trajectories that violate the convexity condition*, Inverse Problems, 22 (2006), pp. 2123–2143.

[7] A. Katsevich, S. Basu, and J. Hsieh, *Exact filtered backprojection reconstruction for dynamic pitch helical cone beam computed tomography*, Physics in Medicine and Biology, 49 (2004), pp. 3089–3103.

[8] A. Katsevich and G. Lauritsch, *Filtered backprojection algorithms for spiral cone beam CT*, in Sampling, Wavelets, and Tomography, J. Benedetto and A. Zayed, eds., Birkhauser, Boston, 2003, pp. 255–287.

[9] A. Katsevich, *Analysis of an exact inversion algorithm for spiral cone-beam CT*, Physics in Medicine and Biology, 47 (2002), pp. 2583–2598.

[10] A. Katsevich, *Theoretically exact filtered backprojection-type inversion algorithm for spiral CT*, SIAM J. Appl. Math., 62 (2002), pp. 2012–2026.

[11] A. Katsevich, *Image reconstruction for the circle and line trajectory*, Physics in Medicine and Biology, 49 (2004), pp. 5059–5072.

[12] A. Katsevich, *An improved exact filtered backprojection algorithm for spiral computed tomography*, Adv. Appl. Math., 32 (2004), pp. 681–697.

[13] A. Katsevich, *On two versions of a 3π algorithm for spiral CT*, Physics in Medicine and Biology, 49 (2004), pp. 2129–2143.

[14] A. Katsevich, *Image reconstruction for the circle and arc trajectory*, Physics in Medicine and Biology, 50 (2005), pp. 2249–2265.

[15] A. Katsevich, *3PI algorithms for helical computer tomography*, Adv. Appl. Math., 21 (2006), pp. 213–250.

[16] T. Köhler, C. Bontus, and P. Koken, *The radon-split method for helical cone-beam CT and its application to nongated reconstruction*, IEEE Trans. Medical Imaging, 25 (2006), pp. 882–897.

[17] H. Kudo, F. Noo, and M. Defrise, *Quasi-exact filtered backprojection algorithm for long-object problem in helical cone-beam tomography*, IEEE Trans. Medical Imaging, 19 (2000), pp. 902–921.

[18] F. Noo, J. Pack, and D. Heuscher, *Exact helical reconstruction using native cone-beam geometries*, Physics in Medicine and Biology, 48 (2003), pp. 3787–3818.

[19] J. D. Pack, F. Noo, and R. Clackdoyle, *Cone-beam reconstruction using the backprojection of locally filtered projections*, IEEE Trans. Medical Imaging, 24 (2005), pp. 1–16.

[20] J. D. Pack and F. Noo, *Cone-beam reconstruction using 1D filtering along the projection of M-lines*, Inverse Problems, 21 (2005), pp. 1105–1120.

[21] E. Sidky, Y. Zou, and X. Pan, *Minimum data image reconstruction algorithms with shift-invariant filtering for helical, cone-beam CT*, Physics in Medicine and Biology, 50 (2005), pp. 1643–1657.

[22] H. Yang, M. Li, K. Koizumi, and H. Kudo, *Exact cone beam reconstruction for a saddle trajectory*, Physics in Medicine and Biology, 51 (2006), pp. 1157–1172.

[23] Y. Ye, S. Zhao, H. Yu, and G. Wang, *A general exact reconstruction for cone-beam CT via backprojection-filtration*, IEEE Trans. Medical Imaging, 24 (2005), pp. 1190–1198.

[24] Y. Ye, J. Zhu, and G. Wang, *Geometric studies on variable radius spiral cone-beam scanning*, Medical Physics, 31 (2004), pp. 1473–1480.

[25] H. Yu, S. Zhao, Y. Ye, and G. Wang, *Exact BPF and FBP algorithms for nonstandard saddle curves*, Medical Physics, 32 (2005), pp. 3305–3312.

[26] T. Zhuang, S. Leng, B. E. Nett, and G. Chen, *Fan-beam and cone-beam image reconstruction via filtering the backprojection image of differentiated projection data*, Physics in Medicine and Biology, 49 (2004), pp. 1643–1657.

[27] Y. Zou, X. Pan, D. Xia, and G. Wang, *PI-line-based image reconstruction in helical cone-beam computed tomography with a variable pitch*, Medical Physics, 32 (2005), pp. 2639–2648.

SIAM J. APPL. MATH.
Vol. 68, No. 2, pp. 354–391

EXPLOITING HISTORY-DEPENDENT EFFECTS TO INFER NETWORK CONNECTIVITY[*]

DUANE Q. NYKAMP[†]

Abstract. We present an approach to distinguish between causal connections and common input connections among nodes in a network. By modeling how the activity of a node depends on its own recent history, we demonstrate how this history dependence predicts different patterns of activity depending on the nature of the network connectivity. In particular, a causal connection between a pair of observed nodes can be distinguished from common input connections that originate from nodes whose activity remains unobserved. This work builds on previous results where this same distinction was made based on modeling how the activity of a node depends on measured external variables such as stimuli. The results have a potentially broad range of application as the analysis can be based on a fairly generic class of models.

Key words. neural networks, correlations, causality, maximum likelihood, point process, autocorrelation

AMS subject classification. 92C20

DOI. 10.1137/070683350

1. Introduction.

The determination of causal connections among nodes within a network is a difficult challenge. This challenge is magnified in the presence of hidden nodes, the effects of which can mimic the presence of causal connections among the set of measured nodes. For example, the connection from a hidden node onto two measured nodes could introduce correlations in the activity of the measured nodes that resemble the effect of a causal connection between the measured nodes (see Figure 1).

We have recently introduced [14, 13, 12] an approach for identifying causal connections in the presence of hidden nodes that is based on modeling the relationship between the activity of nodes and measurable external variables, such as those representing a stimulus. In the original formulation of this approach, the activity of any node could be only weakly dependent on the history of its activity. However, in general, the activity of a node could depend strongly on the recent activity of that node. For example, this approach was originally designed for neuronal networks, and the spiking activity of a neuron is strongly modulated by that neuron's spike history. After firing a spike, a neuron cannot immediately fire a second spike due to its refractory period. Some neurons tend to fire spikes in bursts so that, once the refractory period is over, the probability of firing a spike is transiently much higher than baseline. These history-dependent effects were neglected in our original formulation.

We have now discovered that, if one models how the activity of a node depends on its recent history, one's ability to distinguish causal connections within a network is enhanced. The reason that modeling history dependence can help determine causal connections is caricatured in Figure 2. For the purpose of illustration, imagine that the nodes are neurons and that the measured activity is the times of the neurons' output spikes. Moreover, imagine that neuron 1 tends to fire spikes in pairs (note the

[*] Received by the editors February 22, 2007; accepted for publication (in revised form) July 25, 2007; published electronically November 9, 2007. This research was supported by National Science Foundation grant DMS-0415409.

http://www.siam.org/journals/siap/68-2/68335.html

[†] School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (nykamp@math.umn.edu).

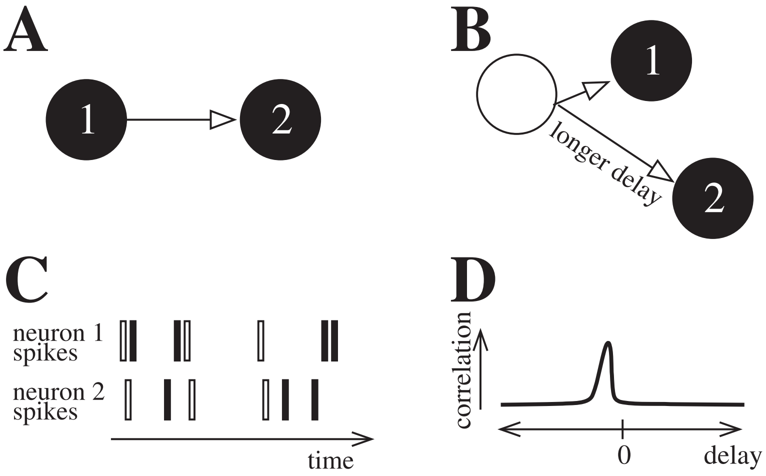


Fig. 1. *The effect of hidden nodes (unfilled circle) can be to mimic causal connections among measured nodes (filled circles).* (A) *A causal connection from measured node* 1 *onto measured node* 2*.* (B) *A common connection from a hidden node onto two measured nodes, where the connection onto measured node* 2 *has a longer delay.* (C) *Both the common input configuration (A) and the causal connection configuration (B) produce similar correlations in the activity of the measured nodes. For concreteness, let the nodes be neurons whose activity is a sequence of spike times illustrated by the temporal sequence of rectangles. Both the networks (A and B) will increase the probability that neuron* 2 *will fire a spike immediately after neuron* 1*. (These spike combinations are highlighted by the unfilled rectangles.)* (D) *Schematic of the correlation induced by either network (A or B). Neuron* 2 *is highly likely to fire a spike a certain delay after neuron* 1 *fires. There is a peak in the correlation measured at that delay. (We arbitrarily use a negative delay when neuron* 2 *follows neuron* 1*.) Since both the common input (A) and the causal connection (B) configurations induce similar correlations in the activity of the measured nodes, our goal is to distinguish which configuration underlies the measured activity of the two nodes.*

pairs of closely spaced spikes in the output of neuron 1 in the right panels of Figure 2). We argue that neuron 2 should respond differently to the spike pairs depending on whether the network contains a causal connection (Figure 2(A)) or common input connections (Figure 2(B)).

To further simplify the situation, imagine that the spike trains of neither neuron 2 nor the hidden neuron have a significant dependence on their history. Then, as portrayed in Figure 2(A), if neuron 1 has a causal connection onto neuron 2, neuron 2 will respond equally well to both spikes in the spike pairs emitted by neuron 1. Neuron 2 will receive both spikes in the pair as inputs, so neuron 2 will be likely to fire a spike immediately after both of these inputs. On the other hand, in the common input configuration of Figure 2(B), neuron 2 does not receive neuron 1's spike pairs as inputs. When the hidden neuron fires a single spike, it may elicit a spike pair from neuron 1. However, neuron 2 just receives the single input from the hidden neuron, so neuron 2 will not be driven to fire twice. When looking at just the spike trains of neuron 1 and 2, it may appear, for example, that neuron 2 is responding to just the first spike of each pair from neuron 1 and ignoring the second spike. The key intuition to gain from this example is that, for the common input configuration, it looks as though neuron 2 does not respond to spikes that can be predicted by the history dependence of neuron 1.

Of course, any real situation will be far more complicated than this exaggerated example. For instance, all of the nodes could have a strong history dependence to their activity, which will confound the simple reasoning given above. Moreover,

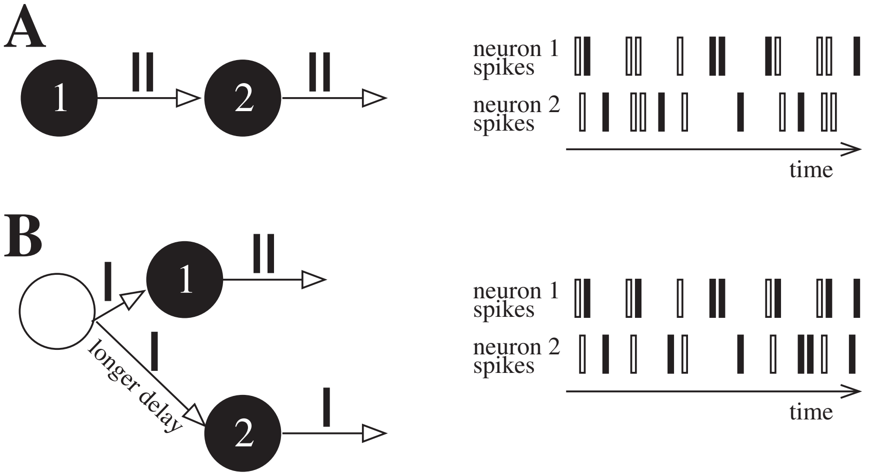


FIG. 2. *Illustration of the different effects of history dependence based on the underlying circuitry. Nodes are spiking neurons as in Figure* 1. *For this illustration, we assume neuron* 1*'s activity is strongly dependent on its spiking history; it is highly likely to fire spikes in pairs. We also assume that the spikes of neuron* 2 *and the unmeasured neuron are largely independent of their respective history.* (A) *In a causal connection configuration, neuron* 2 *may respond to all of neuron* 1*'s spikes. As schematized on the left, when neuron* 1 *fires a pair of spikes (black rectangles), neuron* 2 *is likely to spike after each one and so may spike twice. The right panel illustrates a possible temporal sequence of spikes from both neurons. Spike combinations where neuron* 2 *fires immediately after neuron* 1 *are highlighted by unfilled rectangles. Neuron* 2 *is highly likely to spike both after the first spike and after the second spike in each spike pair from neuron* 1. (B). *In a common input configuration, neuron* 2 *does not receive the spike pairs from neuron* 1. *Since a single spike from the hidden neuron can evoke the spike pair from neuron* 1, *neuron* 2 *receives only one input that is correlated with the spike pair from neuron* 1. *If the connection from the hidden neuron onto neuron* 2 *has a slightly longer delay than the connection onto neuron* 1, *neuron* 2 *will be likely to fire immediately after the first spike in each pair from neuron* 1. *It will not be likely to spike after the second spike of the pair, as illustrated in the right panel.*

the influence of the connections between a pair of nodes will typically be weaker than illustrated here, as input received via any one connection will be just one small influence on a node bathed with inputs from other nodes in the network. Hence, exploiting such history-dependent effects to infer connectivity requires some form of analysis that can synthesize the various ways in which internode connectivity and intranode history-dependent effects interact to influence nodes' activities. Nonetheless, the mathematical analysis we present will confirm that intuition gleaned from this exaggerated example does apply to the more complex situation (see section 3.4.1).

This paper presents a mathematical analysis through which one can employ a model of history-dependent effects to develop estimates of the network connectivity among measured nodes. In section 2, we describe the class of models that we consider. In section 3, we present the analysis to determine the connectivity. We demonstrate the results applied to simulated networks in section 4 and discuss the results in section 5.

2. The history-dependent model.

We present our model and analysis in fairly abstract terms. As detailed in [14], we employ a modular approach where the details of the single-node model are ignored in the network analysis. To employ the results to analyze a particular dataset, one must select an appropriate model, the form of which can be "plugged into" the network analysis.

2.1. The general model formulation. The model is formulated in discrete time. Let R_s^i be a random variable that represents the activity of node s at time point i. Since our examples will involve models of neurons, we will assume that R_s^i is a discrete random variable. However, the analysis proceeds analogously for a continuous random variable. Ignoring the activity of other nodes for a moment, the probability distribution of R_s^i will depend both on the history of node s and on some measurable external variables. Let $\mathbf{R}_s^{<i}$ be the vector of the history of the activity of node s (i.e., the vector with values of R_s^k for $k < i$). Denote the external variable vector by $\mathbf{X}$. The vector $\mathbf{X}$ could represent any quantity or set of quantities whose values are known and that modulate the activity of the nodes. For example, in neuroscience applications, $\mathbf{X}$ could correspond to a sequence of stimuli or a sequence of animal positions. (See [14] for a discussion on external variables. Note that $\mathbf{X}$ could depend on time, although the notation does not make that explicit.)

The activity of a given node on the network also depends on activity of other nodes. We denote the network connectivity by $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$, which indicates the magnitude of the effect of the activity of node $\tilde{s}$ at time $\tilde{\imath}$ (i.e., $R_{\tilde{s}}^{\tilde{\imath}}$) on the activity of node s at time i (i.e., R_s^i). We assume that all connections are causal so that $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} = 0$ for $\tilde{\imath} \geq i$. We model the effect of $R_{\tilde{s}}^{\tilde{\imath}}$ on the probability distribution of R_s^i as a function of the product $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} R_{\tilde{s}}^{\tilde{\imath}}$. Moreover, we simply linearly sum the coupling effects from all nodes and previous time steps, modeling the total coupling effect of all nodes on the probability distribution of R_s^i as a function of the sum

$$\sum_{\tilde{s}\neq s} \sum_{\tilde{\imath}<i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} R_{\tilde{s}}^{\tilde{\imath}}.$$

To summarize, we model the probability distribution of R_s^i as a parametric function of the history $\mathbf{R}_s^{<i}$ of node s, the external variables $\mathbf{X}$, and the past activity of all nodes as

$$\Pr(R_s^i = r_s^i \,|\, \mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) = P_s\left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\tilde{s}\neq s} \sum_{\tilde{\imath}<i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}; \bar{\theta}_s^i \right), \tag{2.1}$$

where P_s is some discrete probability distribution in its first argument and $\bar{\theta}_s^i$ is a vector of parameters. The quantity $\mathbf{R}^{<i}$ (without a subscript) is the history of all nodes, i.e., has components $R_{\tilde{s}}^k$ for all $\tilde{s}$ and all $k < i$. If $\mathbf{R}$ represents all of the activity of all nodes (i.e., has components $R_{\tilde{s}}^k$ for all $\tilde{s}$ and k), then, by Bayes' law, the probability distribution of $\mathbf{R}$, given the value of the external variable vector $\mathbf{X}$, is

$$\begin{aligned} \Pr(\mathbf{R} = \mathbf{r} \,|\, \mathbf{X} = \mathbf{x}) &= \prod_s \prod_i \Pr(R_s^i = r_s^i \,|\, \mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) \\ &= \prod_s \prod_i P_s\left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\tilde{s}\neq s} \sum_{\tilde{\imath}<i} \bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}; \bar{\theta}_s^i \right). \end{aligned} \tag{2.2}$$

To obtain (2.2), we exploited the fact that nodes influence each other only through causal connections. Hence, conditioned on the history $\mathbf{R}^{<i}$ of the network and the external variables, the activities R_s^i of nodes in a single time step i are independent. (In other words, we assume the time bins are small enough so that interactions involve a delay of at least one time bin.)

2.2. Assumptions required for analysis. With the exception of the linear coupling among nodes, (2.2) is a fairly generic description of a network in discrete time. (Recall that the equations could be trivially modified to allow the activity R_s^i to be a continuous random variable.) However, to proceed with our analysis we make a few strong assumptions about the network. These assumptions are similar to those detailed in [14]. (The biggest difference is that here we make no assumptions about the dependence of a node on its own history.) For this reason, we present only a brief discussion of these assumptions here and refer the reader to the more detailed discussion in the former article.

First, we assume that an algorithm exists to fit the activity of a single node to the same parametric model with the coupling factors $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ set to zero. Note that this particular assumption is only about choice of models; it is not an assumption about the network activity. We assume that, from measurements of the activity of just a single node s (i.e., of the vector $\mathbf{R}_s$ composed of R_s^i for all i), one has an algorithm to determine effective parameters θ_s^i by fitting the averaged model[1]

$$\Pr(\mathbf{R}_s = \mathbf{r}_s \,|\, \mathbf{X} = \mathbf{x}) = \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \theta_s^i\big). \tag{2.3}$$

Although the activity of all other nodes $R_{\tilde{s}}^{\tilde{\imath}}$ is ignored in this fitting procedure, we view the R_s^i as really generated from the full network via model (2.2). Therefore, the effective parameters θ_s^i do include the averaged effects of the coupling from other nodes. Our analysis will rely heavily on these effective parameters; hence, the results depend on having chosen a good model P_s and fitting algorithm so that the averaged model (2.3) captures key elements of the activity of each node. This assumption puts stringent limits on the model P_s. For example, one cannot use detailed biophysical models, as all of the parameters of such models cannot be determined by $\mathbf{R}_s$ and $\mathbf{X}$ alone. Neither could one allow the θ_s^i to be independent for each time i. (See [14] for more details.)

Second, we assume that the coupling $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ is weak so that we can expand the full model (2.2) in a Taylor series in $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ and retain terms only through second order. Since we assume that P_s is C^2 in its fourth argument, our analysis will have an error that is $O(\bar{W}^3)$. As detailed in [14], the assumption has the following important consequences: The average coupling strength must scale like $1/N$, where N is the network size; the identity of the nodes that appear in (2.2) must be regarded as "lumped" models that already incorporate effects of nodes projecting to them; and the network topology is highly simplified, as the second order Taylor series will represent combinations of at most two edges of the network graph. If the actual connectivity is too strong to strictly justify this assumption, the resulting connectivity estimates may need to be reinterpreted as an effective connectivity (see the discussion in section 5).

Third, once we have calculated θ_s^i by fitting the averaged model (2.3), we assume that the model is constructed so that we can calculate $P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ for any value of w. This is a strong assumption on the allowed form of the model function P_s, as the averaged model (2.3) is not based on $P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ for any nonzero w. This assumption also implies that we can calculate $\frac{\partial}{\partial w} P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$ and

[1] The probability of the left-hand side of model (2.3) is the marginal distribution of the probability of the left-hand side of model (2.2), averaged over the activity of all other nodes.

$\frac{\partial^2}{\partial w^2} P_s(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i)$. In fact, this assumption implies that these derivatives must be equivalent to derivatives with respect to some function of $\mathbf{r}_s^{<i}$, $\mathbf{x}$, and θ_s^i.

Last, unless one could repeatedly sample the activity of the nodes from the same time points, one couldn't hope to be able to determine arbitrary connectivity $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ that varies freely with the time point. (This is the same reason θ_s^i cannot be allowed to vary freely with the time point, as mentioned above.) When we actually implement the approach, we will eventually (see section 3.3.3) allow $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ to depend on $\tilde{\imath}$ and i only through the delay $i - \tilde{\imath}$. (One could also allow the coupling to adapt slowly with time.) During most of our analysis, we will keep the notation where $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ varies freely with the time point, as it adds no complexity to the equations.

3. The analysis. We begin by giving a short overview of the analysis. We operate under the assumption that model (2.2) gives the true probability distribution of the activity $\mathbf{R}$ of the entire network. However, we assume that one can observe just a small number of nodes with indices q in some subset $\mathcal{Q}$. We denote by $\mathbf{R}_\mathcal{Q}$ the activity of all of these measured nodes. (The components of $\mathbf{R}_\mathcal{Q}$ are a subset of those of $\mathbf{R}$.)

The first step of the analysis will be to derive an expression for the probability distribution of the activity of just the measured nodes. We will derive an expression for this probability, which we denote by $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$, by taking the expression for $\Pr(\mathbf{R}|\mathbf{X})$ given in (2.2) and averaging it over the activity of all hidden nodes. This step will rely heavily on the weak coupling assumption described above.

The resulting expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ will depend on all of the unknown parameters $\bar{\theta}_s^i$ and $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$. Given that many nodes remain hidden, we don't have any hope of obtaining estimates of the original parameters $\bar{\theta}_s^i$. However, we do, by assumption, have an algorithm for determining the effective parameters θ_q^i of any measured node q by fitting the averaged model (2.3) to the activity of that measured node. To take advantage of this information, our second step will be to derive an expression for the original parameters $\bar{\theta}$ in terms of the effective parameters θ.

Our third step is to combine the results of steps one and two to arrive at an expression for the probability distribution $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ of the measured node activity in terms of the effective parameters. With one further approximation, we can group all of the effects of the hidden nodes into a small number of parameters. In the end, our expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ will contain just two sets of unknown parameters: the effective causal connection parameters (which we'll denote by an unbarred W) and the effective common input parameters (which we'll denote by U). Given a measurement of the activity of the measured nodes, we use our expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ to compute maximum likelihood estimates of the W and U. The effective causal connection W will be our estimate of the connectivity among the measured nodes.

3.1. Step one: Average for measured node probability distribution. Our first step is to average the full model (2.2) over all possible values of the activity of hidden nodes to obtain an expression for the probability distribution of measured node activity. Before we compute the average, we use the weak coupling assumption described in section 2.2 to simplify (2.2).

We invoke the weak coupling assumption to expand the full model (2.2) as a Taylor series in $\bar{W}$. To simplify the presentation, we define the following shorthand notation for the probability distribution of the activity of node s (over all time points

i) that would result if all coupling was set to zero:

$$\bar{P}_s = \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \bar{\theta}_s^i, \big). \tag{3.1}$$

Similarly, we define shorthand notation for the derivatives of $\bar{P}_s$ with respect to the $r_{\tilde{s}}^{\tilde{\imath}}$ for $\tilde{s} \neq s$:

$$\frac{\partial \bar{P}_s}{\partial r_{\tilde{s}}^{\tilde{\imath}}} = \frac{\partial}{\partial r_{\tilde{s}}^{\tilde{\imath}}} \left(\prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) \right) \Bigg|_{\{r_{\acute{s}}^{\acute{\imath}}=0 | \acute{s} \neq s\}},$$

$$\frac{\partial^2 \bar{P}_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} = \frac{\partial^2}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} \left(\prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) \right) \Bigg|_{\{r_{\acute{s}}^{\acute{\imath}}=0 | \acute{s} \neq s\}}. \tag{3.2}$$

Since the $\bar{W}$ appear only in the combination $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i} r_{\tilde{s}}^{\tilde{\imath}}$, we can write our Taylor series in $\bar{W}$ as though it were a Taylor series in the $r_{\tilde{s}}^{\tilde{\imath}}$. This notation will turn out to be more convenient for the analysis because the key factors of $r_{\tilde{s}}^{\tilde{\imath}}$ will be written out explicitly. Note that, for each node s, we make no assumptions about the effect of its own history $\mathbf{r}_s^{<i}$ and do *not* expand out this history dependence in a Taylor series.

Using the above shorthand notation, the Taylor series of (2.2) is

$$\begin{aligned} \Pr(\mathbf{R} = \mathbf{r} | \mathbf{X} = \mathbf{x}) = \prod_s \bar{P}_s &+ \sum_{\substack{s_1, \tilde{s}_1 \\ \tilde{s}_1 \neq s_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &+ \frac{1}{2} \sum_{\substack{s_1, \tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s_1, \tilde{s}_2 \neq s_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\ &+ \frac{1}{2} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3} + O(\bar{W}^3). \end{aligned} \tag{3.3}$$

Note that the derivative $\partial \bar{P}_s / \partial r_{\tilde{s}}^{\tilde{\imath}}$ corresponds to the effect of a connection from node $\tilde{s}$ onto node s. If we wrote out the derivative explicitly, it would contain a sum of terms involving $\bar{W}_{\tilde{s},s}^{\tilde{\imath},i}$ for all $i > \tilde{\imath}$. It represents the change in the distribution of all R_s^i for $i > \tilde{\imath}$ given a change in $R_{\tilde{s}}^{\tilde{\imath}}$ (calculated at $R_{\tilde{s}}^{\tilde{\imath}} = 0$).

We can now write down an expression for the activity of all measured nodes by averaging over all possible values of the activity of the hidden nodes. As mentioned above, let $\mathcal{Q}$ denote the set of node indices corresponding to all measured nodes. Similarly, let $\mathcal{P}$ denote the set of node indices corresponding to all hidden nodes. Then $\mathcal{Q} \cup \mathcal{P}$ corresponds to the entire network. To simplify the notation, we will make the following notational conventions. We will use the index s and its variants to index all nodes in the network; i.e., we implicitly assume that $s \in \mathcal{Q} \cup \mathcal{P}$. We will use the indices p and q (and their variants) to index hidden and measured nodes, respectively; i.e., we implicitly assume that $p \in \mathcal{P}$ and $q \in \mathcal{Q}$. Last, we let $\mathbf{R}_{\mathcal{Q}}$ and $\mathbf{R}_{\mathcal{P}}$ represent all measured node activity R_q^i and all hidden node activity R_p^i, respectively.

To derive an expression for the probability distribution of all measured activity, we average (3.3) over all possible values of $\mathbf{R}_{\mathcal{P}}$. The probability distribution of $\mathbf{R}_{\mathcal{Q}}$ is therefore

$$\begin{aligned}
\Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}} | \mathbf{X} = \mathbf{x}) &= \sum_{\mathbf{r}_{\mathcal{P}}} \Pr(\mathbf{R} = \mathbf{r} | \mathbf{X} = \mathbf{x}) \\
&= \sum_{\mathbf{r}_{\mathcal{P}}} \prod_{s} \bar{P}_s + \sum_{\mathbf{r}_{\mathcal{P}}} \sum_{\substack{s_1, \tilde{s}_1 \\ \tilde{s}_1 \neq s_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\
&\quad + \frac{1}{2} \sum_{\mathbf{r}_{\mathcal{P}}} \sum_{\substack{s_1, \tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s_1, \tilde{s}_2 \neq s_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_2 \\ s_2 \neq s_1}} \bar{P}_{s_2} \\
&\quad + \frac{1}{2} \sum_{\mathbf{r}_{\mathcal{P}}} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} r_{\tilde{s}_1}^{\tilde{\imath}_1} r_{\tilde{s}_2}^{\tilde{\imath}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3} + O(\bar{W}^3),
\end{aligned} \tag{3.4}$$

where the sum over $\mathbf{r}_{\mathcal{P}}$ indicates a sum over all possible values of the hidden node activity.

It turns out that we can explicitly compute the sum over $\mathbf{r}_{\mathcal{P}}$. Note that the value r_s^i of a given random variable can appear in (3.4) either explicitly or in the probability distribution $\bar{P}_s$ (or its derivatives). It is not hidden in any other factors. Therefore, to compute a sum over all possible values of the activity of a node indexed by some s, we can factor out everything except one factor of $\bar{P}_s$ (or a derivative of $\bar{P}_s$) and a polynomial in the r_s^i. Hence, we need to derive expressions for the average of such quantities.

The average of a polynomial in the r_s^i multiplied by the undifferentiated $\bar{P}_s$ will simply be the expected value of that polynomial, under the probability distribution $\bar{P}_s$ with the W argument set to zero. Taking the average of expressions involving the derivatives of $\bar{P}_s$ is more complicated. In Appendix A.1, we outline how to compute such averages. The important point is that one can compute these averages explicitly in terms of the model parameters and the probability distributions $P_s(\cdot)$. We end up with the lengthy expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ given by (A.5).

3.2. Step two: Original parameters in terms of effective parameters. One of the assumptions given in section 2.2 is the existence of an algorithm to calculate the effective parameters θ_s^i by fitting the averaged model (2.3) to the activity of node s (while ignoring the activity of all other nodes). Hence, we can regard the effective parameters θ_q^i as known for all measured nodes $q \in \mathcal{Q}$. In the previous step, we obtained an expression for the probability distribution of the measured node activity $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ in terms of the unknown original model parameters $\bar{\theta}_s^i$. In this second step of the analysis, we will derive a relationship between the effective parameters θ_s^i and the original paramters $\bar{\theta}_s^i$. This relationship will allow us to rewrite our equation for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ in terms of the effective parameters.

We define equivalent shorthand notation for expressions involving the effective parameters as we did for expressions involving the original model parameters. We define P_s to be the probability distribution that we fit from the averaged model (2.3):

$$P_s = \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \theta_s^i,\big) = \Pr(\mathbf{R}_s = \mathbf{r}_s \,|\, \mathbf{X} = \mathbf{x}). \tag{3.5}$$

We then define the derivatives P_s just as we did for $\bar{P}_s$:

$$\frac{\partial P_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}} = \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} \left(\prod_i P_s \left(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}^{\acute{\imath},i}_{\acute{s},s} r^{\acute{\imath}}_{\acute{s}}; \theta^i_s \right) \right) \Bigg|_{\{r^{\acute{\imath}}_{\acute{s}}=0 | \acute{s} \neq s\}} \tag{3.6}$$

and analogously for the second derivatives. We also define the expected values:

$$E_0(g(\mathbf{R})) = \sum_{\mathbf{r}} g(\mathbf{r}) \prod_s P_s, \tag{3.7a}$$

$$E_0 \left(\frac{\partial R^i_s}{\partial R^{\tilde{\imath}}_{\tilde{s}}} \right) = \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} E(R^i_s | \mathbf{R}^{<i} = \mathbf{r}^{<i}) \Bigg|_{\{\mathbf{r}_{\acute{s}}=\mathbf{0} | \acute{s} \neq s\}} = \sum_{\mathbf{r}_s} r^i_s \frac{\partial P_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}}. \tag{3.7b}$$

These expected values are analogous to the barred versions given in Appendix A.1 ((A.2) and (A.4)) except that they are based on the averaged model (2.3). We assume that the chosen model and fitting algorithm for θ^i_s results in the averaged model (2.3) being a good approximation. Then (3.7a) does indeed represent the expected value of any function for the activity. For example, $E_0(R^i_s)$ is the expected value of the activity of node s at time i. We will also use the statistic $E_0(R^{i_1}_s R^{i_2}_s) - E_0(R^{i_1}_s) E_0(R^{i_2}_s)$, which represents the covariance of the activity of node s at the times i_1 and i_2.

The derivative of (3.7b) represents how the average activity of node s at time i changes with the activity of node $\tilde{s}$ at time $\tilde{\imath}$. (Since we assume causal connections, this is nonzero only if $\tilde{\imath} < i$.) See Appendix A.1 for further discussion on the properties of such derivatives.

Although we know the effective parameters only for measured nodes, we can still define the (unknown) effective parameters for hidden nodes using the averaged model (2.3). Using effective parameters for all nodes will simplify the form of our equation for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$.

It turns out that we have already done much of the work toward deriving an equation for effective parameters in step one, above. In that first step, we derived an expression for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, which is the marginal distribution (of the full distribution $\Pr(\mathbf{R}|\mathbf{X})$ given by model (2.2)) for the activity of a set of measured nodes. The averaged model (2.3) is based on $\Pr(\mathbf{R}_s|\mathbf{X})$, which we can regard as the marginal distribution for the activity of a single node. If we replace the set $\mathcal{Q}$ of measured nodes in (A.5) with just the single node s, then (A.5) becomes the marginal distribution for the activity of a single node. In this way, we obtain an expression for $\Pr(\mathbf{R}_s|\mathbf{X})$ in terms of the original model parameters. Given the definition (2.3) of the effective parameters, we have obtained an expression for the effective parameters θ in terms of the original model parameters $\bar{\theta}$.

However, we need to go the other direction: to transform expressions involving the original model parameters $\bar{\theta}$ in terms of the effective parameters θ. Using the procedure outlined in Appendix A.2, we can solve for the original uncoupled probability $\bar{P}_s$ (which is a function of the $\bar{\theta}^i_s$) in terms of the effective probability P_s (which is a function of the θ^i_s). We obtain the following relationship:

$$\begin{aligned} \bar{P}_s = P_s &- \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) \\ &- \frac{1}{2} \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_1}} [E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_1}) - E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_1})] \end{aligned}$$

$$\begin{aligned}
&-\sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}}\sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_s}\right)[r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)]\\
&+\frac{1}{2}\sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2}) + O(\bar{W}^3). \qquad (3.8)
\end{aligned}$$

Each term on the right-hand side of (3.8) has a significant meaning and illustrates the process of approximating a full network (2.2) by an averaged model (2.3). The sum on the first line is simply the change in the probability distribution of node s caused by the average effect of connections from other nodes $\tilde{s}_1$. Intuitively, this change is the average activity of node $\tilde{s}_1$ times the effect of node $\tilde{s}_1$ on the probability distribution of node s (i.e., the derivative of P_s). Since the effective distribution P_s includes the average influence of connections from other nodes, this term must be subtracted from P_s to regain the original uncoupled distribution $\bar{P}_s$.

The term from the second line accounts for second-order effects from a connection from node $\tilde{s}_1$. First consider the case where $\tilde{\imath}_1 = \tilde{\imath}_2$. Now imagine that the effect of the connection from node $\tilde{s}_1$ onto node s lasts multiple time steps.[2] Then the connection from node $\tilde{s}_1$ will introduce correlations in the activity of node s. (Recall how common input from a node onto two different nodes can introduce correlations between those two nodes. The effect of the second line of (3.8) is identical except that in this case we have "common input" onto the same node but at different times, which creates correlations within that one node's activity.) This correlation will be proportional to the variance of $R^{\tilde{\imath}_1}_{\tilde{s}_1}$.

The case with $\tilde{\imath}_1 \neq \tilde{\imath}_2$ is similar. If $R^{\tilde{\imath}_1}_{\tilde{s}_1}$ is correlated with $R^{\tilde{\imath}_2}_{\tilde{s}_1}$ (due to the history dependence of the activity of node $\tilde{s}_1$), then the combined effect of the activity of node $\tilde{s}_1$ at times $\tilde{\imath}_1$ and $\tilde{\imath}_2$ will induce correlations in the activity of node s. This correlation will be proportional to the covariance of $R^{\tilde{\imath}_1}_{\tilde{s}_1}$ and $R^{\tilde{\imath}_2}_{\tilde{s}_1}$.

The reason this source of correlation must be subtracted from P_s in the second line of (3.8) is as follows. When fitting the averaged model (2.3) for node s, one is averaging over the activity of all other nodes, including node $\tilde{s}_1$. The induced correlations due to the connection from node $\tilde{s}_1$ will still be present in the activity of node s. Hence, the averaged model P_s (and its parameters θ_s) will take into account this additional correlation, and the additional correlation will appear in the averaged model as part of the history dependence of node s. However, the original uncoupled model represented by $\bar{P}_s$ (and its parameters $\bar{\theta}_s$) will not include effects due to coupling from other nodes. This history dependence of $\bar{P}_s$ would not include these additional correlations due to the connection from node $\tilde{s}_1$. Hence, the effect of these correlations must be subtracted from the effective distribution P_s to regain the original distribution $\bar{P}_s$, as is done in the second line of (3.8).

The term from the third line of (3.8) is similar in that it accounts for additional correlations in the activity of node s due to connections involving other nodes. In this case, the correlations are induced by indirect connections from node s onto itself via one of the other nodes $\tilde{s}_1$. This effect has three components as shown by the three factors. The right factor is the deviation of the activity of node s at time $\tilde{\imath}_2$

[2] Since P_s, as defined in (3.5), models the activity of node s for all time steps, the derivative $\partial^2 P_s/(\partial r^{\tilde{\imath}_1}_{\tilde{s}_1})^2$ includes the effects of $R^{\tilde{\imath}_1}_{\tilde{s}_1}$ on the activity of node s at all times. In particular, this derivative captures how the activity of node $\tilde{s}_1$ at a single time point $\tilde{\imath}_1$ can influence the activity of node s at two different times, thus causing correlations in the activity of node s at those two times.

from its expected activity as predicted by the averaged model (2.3). The middle factor is the effect of the activity of node s at time $\tilde{\imath}_2$ on the activity of node $\tilde{s}_1$ at time $\tilde{\imath}_1$. The left factor is the effect of the activity of node $\tilde{s}_1$ at time $\tilde{\imath}_1$ on the probability distribution of node s. The resulting correlation in the activity of node s from this chain of connection would be included in the history dependence of the effective distribution P_s. But, since these correlations depend on connections, their effect would not be included in the original uncoupled distribution $\bar{P}_s$. Hence, their effect must be subtracted from P_s to regain the original distribution $\bar{P}_s$.

The sum from the last line of (3.8) is simply a second-order effect of single connections onto node s. Equation (3.8) is accurate up to second order in $\bar{W}$. The sum of the first line is only a first-order approximation of the change in P_s due to the average effect of connections from other nodes. The addition of the last line gives the correct second-order approximation.

3.3. Step three: Measured node distribution in terms of effective parameters.

Our third step is to derive an expression for the probability distribution $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ of the measured node activity in terms of the effective parameters θ_s^i. Once we have written down an initial form of this distribution, we can simplify it by grouping the effects of hidden nodes into two sets of parameters: an effective causal connection W and an effective common input U. Then, by making one further assumption, we can sufficiently reduce the degrees of freedom within W and U so that computing their solution becomes tractable.

3.3.1. The initial form of the measured node probability distribution.

In the first step of our analysis, we obtained a lengthy expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$, the probability distribution of the measured node activity. (It is given by (A.5) in Appendix A.1.) However, this expression is in terms of the original model parameters $\bar{\theta}_s^i$ which remain unknown. As outlined in Appendix A.3, we rewrite this expression in terms of the effective parameters. Appendix A.4 describes how we transform the result into the form of a true probability (which we need since we wish to use it to develop maximum likelihood estimates of network parameters). We show in Appendix A.4 that this step requires one small deviation from a true second-order approximation, so we will use the $\approx$ symbol in our result. We also use the shorthand notation[3]

$$\begin{aligned} P_s^i &= P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \theta_s^i\big), \\ \frac{\partial P_s^i}{\partial w} &= \frac{\partial}{\partial w} P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i\big)\Big|_{w=0}. \end{aligned} \tag{3.9}$$

In the end, we obtain the following expression for the probability distribution of the measured nodes' activity:

$$\Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q}|\mathbf{X} = \mathbf{x}) \approx \prod_q \prod_i P_q\big(r_q^i, \mathbf{r}_q^{<i}, \mathbf{x}, \widetilde{W}_q^i; \theta_q^i\big) + O(\bar{W}^3), \tag{3.10a}$$

[3] Note the subtle difference between the new notation P_s^i and $\partial P_s^i/\partial w$ (defined by (3.9)) on one hand and the similar notation P_s and $\partial P_s/\partial r_{\tilde{s}}^{\tilde{\imath}}$ (defined by (3.5) and (3.6)) on the other hand. One key difference is that the new notation contains a superscript i, which means it refers to the distribution of the activity of node s just at time point i.

where

$$
\begin{aligned}
\widetilde{W}_q^i = {} & \sum_{\substack{\tilde q\\ \tilde q\neq q}} \sum_{\substack{\tilde\imath_1\\ \tilde\imath_1<i}} \bar W^{\tilde\imath_1,i}_{\tilde q,q}[r^{\tilde\imath_1}_{\tilde q} - E_0(R^{\tilde\imath_1}_{\tilde q})] \\
& + \sum_{\substack{p,\tilde q\\ \tilde q\neq q}} \sum_{\substack{\tilde\imath_1,\tilde\imath_2\\ \tilde\imath_2<\tilde\imath_1<i}} \bar W^{\tilde\imath_1,i}_{p,q} E_0\left(\frac{\partial R_p^{\tilde\imath_1}}{\partial R_{\tilde q}^{\tilde\imath_2}}\right)[r_{\tilde q}^{\tilde\imath_2} - E_0(R_{\tilde q}^{\tilde\imath_2})] \\
& + \sum_{\substack{p,\tilde q\\ \tilde q\neq q}} \sum_{\substack{\tilde\imath_1,\tilde\imath_2,\tilde\imath_3\\ \tilde\imath_2<\tilde\imath_3<i\\ \tilde\imath_1<i}} \bar W^{\tilde\imath_1,i}_{p,q}\bar W^{\tilde\imath_2,\tilde\imath_3}_{p,\tilde q} \frac{\partial P^{\tilde\imath_3}_{\tilde q}}{\partial w}\frac{1}{P^{\tilde\imath_3}_{\tilde q}}[E_0(R_p^{\tilde\imath_1}R_p^{\tilde\imath_2}) - E_0(R_p^{\tilde\imath_1})E_0(R_p^{\tilde\imath_2})] \\
& + \sum_{\substack{p,\tilde q\\ \tilde q<q}} \sum_{\substack{\tilde\imath_1,\tilde\imath_2\\ \tilde\imath_1,\tilde\imath_2<i}} \bar W^{\tilde\imath_1,i}_{p,q}\bar W^{\tilde\imath_2,i}_{p,\tilde q} \frac{\partial P^{i}_{\tilde q}}{\partial w}\frac{1}{P^{i}_{\tilde q}}[E_0(R_p^{\tilde\imath_1}R_p^{\tilde\imath_2}) - E_0(R_p^{\tilde\imath_1})E_0(R_p^{\tilde\imath_2})] \\
& - \sum_{\substack{\tilde q\\ \tilde q\neq q}} \sum_{\substack{\tilde\imath_1,\tilde\imath_2,\tilde\imath_3\\ \tilde\imath_2<\tilde\imath_3<i\\ \tilde\imath_1<i}} \bar W^{\tilde\imath_1,i}_{\tilde q,q}\bar W^{\tilde\imath_2,\tilde\imath_3}_{\tilde q,q} \frac{\partial P^{\tilde\imath_3}_{q}}{\partial w}\frac{1}{P^{\tilde\imath_3}_{q}}[E_0(R_{\tilde q}^{\tilde\imath_1}R_{\tilde q}^{\tilde\imath_2}) - E_0(R_{\tilde q}^{\tilde\imath_1})E_0(R_{\tilde q}^{\tilde\imath_2})] \\
& - \sum_{\substack{\tilde q\\ \tilde q\neq q}} \sum_{\substack{\tilde\imath_1,\tilde\imath_2\\ \tilde\imath_2<\tilde\imath_1<i}} \bar W^{\tilde\imath_1,i}_{\tilde q,q} E_0\left(\frac{\partial R_{\tilde q}^{\tilde\imath_1}}{\partial R_{q}^{\tilde\imath_2}}\right)[r_{q}^{\tilde\imath_2} - E_0(R_{q}^{\tilde\imath_2})]. \qquad \text{(3.10b)}
\end{aligned}
$$

Though this expression is somewhat lengthy, each line of (3.10b) represents the effect of a connection or combination of connections on the probability distribution of the measured nodes' activity. Just as we did for the single-node results (3.8), we briefly describe the effects of the connections as embedded in (3.10).

The sum from the first line of (3.10b) represents a direct causal connection from measured node $\tilde q$ onto measured node q. The second line represents an indirect causal connection from measured node $\tilde q$ onto measured node q via a hidden node p. Both lines describe a change in the distribution of the activity of node q due to a deviation in the activity of node $\tilde q$ from that predicted by the averaged model.

The third and fourth lines are the common input onto measured nodes q and $\tilde q$ from a hidden node p. The common input effect is proportional to the (unknown) (co)variance of the activity of node p (compare to the second line of (3.8)). We separated out the common input that reaches nodes q and $\tilde q$ simultaneously (fourth line of (3.10b)). We arbitrarily put this common input effect into the $\widetilde{W}_q$ of the node with the higher index (as we restrict the sum to $\tilde q < q$). The goal of this analysis will be to distinguish the common input from the third line from the causal connections of the first two lines. (We will assume any correlations at zero delay are due to the common input described on the fourth line.)

The fifth and sixth lines of (3.10b) involve only measured nodes. These lines are similar to the second and third lines of (3.8), and their presence in (3.10) has a similar origin. When the effective parameters of node q were determined, the activity of node $\tilde q$ was ignored. Nonetheless, connections from node $\tilde q$ onto node q still influenced the activity of node q. As we described in the context of (3.8), the activity of node $\tilde q$ could induce correlations in the activity of node q if it had connections onto node q that lasted multiple time steps. Similarly, node $\tilde q$ could induce correlations in node q via an indirect connection from node q onto itself through node $\tilde q$.

If the network contains such a pattern of connections, the effective distribution P_q of node q would already contain such correlations as part of the history dependence of the model. Hence, the probability distribution of $R_{\mathcal{Q}}$ in (3.10) would contain these correlations in the activity of node q even if all $\bar{W}$ were zero. However, these correlations in the activity of node q were caused by connections (i.e., nozero $\bar{W}$) between node q and $\tilde{q}$, as described above. When these individual connections are added to (3.10) via the direct causal connections of the first line of (3.10b), the resulting correlations in the activity of node q will have been added to (3.10) twice. To correct for this, we need to explicitly subtract them off via the fifth and sixth lines of (3.10b).

3.3.2. Grouping the effects of hidden nodes. Once the effective parameters θ_q^i have been determined for all measured nodes q, the only unknowns in (3.10) are the connectivity factor $\bar{W}$ and all expressions involving hidden nodes p. We group these unknowns into two expressions:

$$\begin{aligned} W^{i_2,i_1}_{q_2,q_1} &= \bar{W}^{i_2,i_1}_{q_2,q_1} + \sum_{p} \sum_{\substack{\tilde{i} \\ i_1>\tilde{i}>i_2}} \bar{W}^{\tilde{i},i_1}_{p,q_1} E_0\left(\frac{\partial R_p^{\tilde{i}}}{\partial R_{q_2}^{i_2}}\right), \\ U^{i_2,i_1}_{q_2,q_1} &= \sum_{p} \sum_{\substack{\tilde{i}_1,\tilde{i}_2 \\ \tilde{i}_1<i_1,\tilde{i}_2<i_2}} \bar{W}^{\tilde{i}_1,i_1}_{p,q_1} \bar{W}^{\tilde{i}_2,i_2}_{p,q_2} [E_0(R_p^{\tilde{i}_1} R_p^{\tilde{i}_2}) - E_0(R_p^{\tilde{i}_1}) E_0(R_p^{\tilde{i}_2})], \end{aligned} \tag{3.11}$$

defined for $q_2 \neq q_1$. The causal connection factor $W^{i_2,i_1}_{q_2,q_1}$ is the effective causal connection from node q_2 onto node q_1. It includes an indirect causal connection via a hidden node p. The direct and indirect causal connections are lumped together as we cannot distinguish between them. The common input factor $U^{i_2,i_1}_{q_2,q_1}$ is the effective common input from hidden nodes that arrives at node q_2 at time i_2 and at node q_1 and time i_1. Both $W^{i_2,i_1}_{q_2,q_1}$ and $U^{i_2,i_1}_{q_2,q_1}$ include sums over arbitrary hidden nodes. Although we cannot resolve the individual contributions of the hidden nodes, we will be able to solve for these effective parameters.

We also rewrite the expression for the expected value of the derivative to pull out the hidden factor of $\bar{W}$ contained in it. From the definition (3.7) as well as the definitions of the derivatives (3.6) and (3.9), we write[4]

$$\begin{aligned} E_0\left(\frac{\partial R_s^i}{\partial R_{\tilde{s}}^{\tilde{i}}}\right) &= \sum_{\mathbf{r}_s} r_s^i \frac{\partial P_s}{\partial r_{\tilde{s}}^{\tilde{i}}} = \sum_{\mathbf{r}_s} r_s^i \sum_{\substack{i_2 \\ \tilde{i}<i_2\leq i}} \bar{W}^{\tilde{i},i_2}_{\tilde{s},s} \frac{\partial P_s^{i_2}}{\partial w} \frac{1}{P_s^{i_2}} \prod_{i_3} P_s^{i_3} \\ &= \sum_{\substack{i_2 \\ \tilde{i}<i_2\leq i}} \bar{W}^{\tilde{i},i_2}_{\tilde{s},s} E_0\left(R_s^i \frac{\partial P_s^{i_2}}{\partial w} \frac{1}{P_s^{i_2}}\right). \end{aligned} \tag{3.12}$$

With these definitions of W and U, $\widetilde{W}_q^i$ becomes

$$\begin{aligned} \widetilde{W}_q^i = {} & \sum_{\substack{\tilde{q} \\ \tilde{q}\neq q}} \sum_{\substack{\tilde{i} \\ \tilde{i}<i}} W^{\tilde{i},i}_{\tilde{q},q} [r_{\tilde{q}}^{\tilde{i}} - E_0(R_{\tilde{q}}^{\tilde{i}})] \\ & + \sum_{\substack{\tilde{q} \\ \tilde{q}\neq q}} \sum_{\substack{\tilde{i} \\ \tilde{i}<i}} U^{\tilde{i},i}_{\tilde{q},q} \frac{\partial P_{\tilde{q}}^{\tilde{i}}}{\partial w} \frac{1}{P_{\tilde{q}}^{\tilde{i}}} + \sum_{\substack{\tilde{q} \\ \tilde{q}<q}} U^{i,i}_{\tilde{q},q} \frac{\partial P_{\tilde{q}}^{i}}{\partial w} \frac{1}{P_{\tilde{q}}^{i}} \end{aligned}$$

[4] One subtlety in (3.12) is the fact that we restrict $i_2 \leq i$. If $i_2 > i$, then the term disappears due to a similar argument as underlying the identities in (A.3).

$$-\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}}\sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3<i\\ \tilde{\imath}_1<i}} W_{\tilde{q},q}^{\tilde{\imath}_1,i}W_{\tilde{q},q}^{\tilde{\imath}_2,\tilde{\imath}_3}\frac{\partial P_q^{\tilde{\imath}_3}}{\partial w}\ \frac{1}{P_q^{\tilde{\imath}_3}}[E_0(R_{\tilde{q}}^{\tilde{\imath}_1}R_{\tilde{q}}^{\tilde{\imath}_2})-E_0(R_{\tilde{q}}^{\tilde{\imath}_1})E_0(R_{\tilde{q}}^{\tilde{\imath}_2})]$$

$$-\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}}\sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2,\tilde{\imath}_3\\ \tilde{\imath}_2<\tilde{\imath}_3\leq\tilde{\imath}_1<i}} W_{q,\tilde{q}}^{\tilde{\imath}_2,\tilde{\imath}_3}W_{\tilde{q},q}^{\tilde{\imath}_1,i}E_0\left(R_{\tilde{q}}^{\tilde{\imath}_1}\frac{\partial P_{\tilde{q}}^{\tilde{\imath}_3}}{\partial w}\frac{1}{P_{\tilde{q}}^{\tilde{\imath}_3}}\right)[r_q^{\tilde{\imath}_2}-E_0(R_q^{\tilde{\imath}_2})]. \tag{3.13}$$

Note that, according to (3.12), the quantity $E_0\big(\partial R_{\tilde{s}}^{\tilde{\imath}}/\partial R_s^i\big)$ is $O(\bar{W})$. Hence, the definition (3.11) of W shows that W is a first-order approximation to $\bar{W}$ (i.e., $W_{\tilde{q},q}^{\tilde{\imath},i} = \bar{W}_{\tilde{q},q}^{\tilde{\imath},i}+O(\bar{W}^2)$). This means that, in terms that are quadratic in $\bar{W}$, we can replace $\bar{W}$ with W and still maintain our second-order approximation (as the error is cubic in $\bar{W}$). This allowed us to write (3.13) in terms of just the effective W.

Our expression for the probability distribution $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ of the measured activity is now (3.10a) combined with (3.13). Given the effective parameters θ_q^i, we can calculate the P_q^i and the $\partial P_q^i/\partial w$ via (3.9). We can also calculate, in principle, all of the expressions involving $E_0(\cdot)$ using the definitions in (3.7). (We estimate these via Monte Carlo simulations, as described in Appendix C.) Therefore, the only remaining unknown factors are the causal connection factors W and the common input factors U.

3.3.3. A further assumption for a tractable solution. Our goal is to estimate W and U by finding their values that maximize our approximation of the probability distribution of measured activity. In other words, we seek maximum likelihood estimators of W and U. However, there are still too many unknowns to make the solution tractable, as we still have more unknowns than we would have data points (we have only one measurement of activity per measured node per time point).[5] To reduce the number of unknowns, we assume that W and U depend only on the difference between their temporal indicies, i.e.,

$$W_{q_1,q_2}^{i-j,i}=W_{q_1,q_2}^{j} \qquad \text{and} \qquad U_{q_1,q_2}^{i-j,i}=U_{q_1,q_2}^{j}. \tag{3.14}$$

(One could presumably weaken this assumption by allowing W and U to change slowly over time at the cost of additional computational complexity and increased data requirements.)

This assumption for W has no hidden surprises, as it is equivalent to assuming that the underlying connectivity $\bar{W}$ depends only on the difference in temporal indicies.[6] However, this assumption for U is more significant than may appear at first glance. It turns out that this assumption is really about the hidden nodes and affects how one can interpret the meaning of W and U.

To demonstrate this, we rewrite the definition of U from (3.11) using the index j to indicate the difference between temporal indices:

$$U_{q_2,q_1}^{i_1-j_1,i_1} = \sum_{p}\sum_{\substack{j_2,j_3\\ j_2>0,j_3>0}} \bar{W}_{p,q_1}^{i_1-j_2,i_1}\bar{W}_{p,q_2}^{i_1-j_1-j_3,i_1-j_1}[E_0(R_p^{i_1-j_2}R_p^{i_1-j_1-j_3})-E_0(R_p^{i_1-j_2})E_0(R_p^{i_1-j_1-j_3})].$$

[5] Perhaps one could solve for W and U in full generality if one could repeatedly sample from a small number of time bins and one assumed that the $\bar{W}$ could vary over the time bins but were identical for each repetition.

[6] The effect of the connectivity, however, could vary with time, as each $P_s(\cdot)$ could change with time.

Our assumption on $U_{q_2,q_1}^{i_1-j_1,i_1}$ is that it is independent of i_1. If the $\bar{W}$ depend only on the difference in temporal indices, the only place on the right-hand side where i_1 doesn't immediately drop out is in the (co)variance of activity of the hidden node p. Hence, by insisting that $U_{q_2,q_1}^{i_1-j_1,i_1}$ be independent of i_1, we are really approximating the covariance of each hidden node p as though it were independent of time bin i_1. Equivalently, we could view this approximation as replacing the covariance of node p with its average over all time bins i_1.

As detailed in [14], such an approximation leads to a certain degree of ambiguity in the identification of causal connections, which we refer to as *subpopulation ambiguity.* This ambiguity contains subtleties that are out of the scope of this article and are discussed extensively in [14]. We illustrate the basic consequences of the ambiguity with simulation results (see section 4.4). Note also that, as described in [14], this ambiguity is already present in many experimental contexts (such as those commonly used in neuroscience); hence, in those contexts, this approximation does not add additional ambiguity.

Putting this all together, our procedure to construct the causal connections among measured nodes is as follows. For each measured node indexed by $q \in \mathcal{Q}$, determine the effective parameters θ_q^i by fitting the averaged model (2.3) to the external variables $\mathbf{X}$ and the activity R_q^i of node q. (We assume such an algorithm for determining the θ_q^i is known.) Then determine the effective causal connections W_{q_1,q_2}^j and the effective common input U_{q_1,q_2}^j from the external variables and the activity $\mathbf{R}_\mathcal{Q}$ of all measured nodes by finding the values of W_{q_1,q_2}^j and U_{q_1,q_2}^j that maximize the log-likelihood modeled by the equation

$$\log \Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q} | \mathbf{X} = \mathbf{x}) = \sum_q \sum_i \log P_q\big(r_q^i, \mathbf{r}_q^{<i}, \mathbf{x}, \widetilde{W}_q^i; \theta_q^i\big), \tag{3.15a}$$

where

$$\begin{aligned}\widetilde{W}_q^i = &\sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} W_{\tilde{q},q}^j [r_{\tilde{q}}^{i-j} - E_0(R_{\tilde{q}}^{i-j})] \\ &+ \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j\\ j>0}} U_{\tilde{q},q}^j \frac{\partial P_{\tilde{q}}^{i-j}}{\partial w} \frac{1}{P_{\tilde{q}}^{i-j}} + \sum_{\substack{\tilde{q}\\ \tilde{q}<q}} U_{\tilde{q},q}^0 \frac{\partial P_{\tilde{q}}^i}{\partial w} \frac{1}{P_{\tilde{q}}^i} \\ &- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j_1,j_2,j_3\\ j_1,j_2,j_3>0}} W_{\tilde{q},q}^{j_1} W_{\tilde{q},q}^{j_2} \frac{\partial P_q^{i-j_3}}{\partial w} \frac{1}{P_q^{i-j_3}} [E_0(R_{\tilde{q}}^{i-j_1} R_{\tilde{q}}^{i-j_3-j_2}) - E_0(R_{\tilde{q}}^{i-j_1}) E_0(R_{\tilde{q}}^{i-j_3-j_2})] \\ &- \sum_{\substack{\tilde{q}\\ \tilde{q}\neq q}} \sum_{\substack{j_1,j_2,j_3\\ j_1,j_2>0\\ j_3\geq 0}} W_{q,\tilde{q}}^{j_2} W_{\tilde{q},q}^{j_1} E_0\bigg(R_{\tilde{q}}^{i-j_1} \frac{\partial P_{\tilde{q}}^{i-j_1-j_3}}{\partial w} \frac{1}{P_{\tilde{q}}^{i-j_1-j_3}}\bigg) [r_q^{i-j_1-j_3-j_2} - E_0(x R_q^{i-j_1-j_3-j_2})].\end{aligned} \tag{3.15b}$$

Unfortunately, especially with the terms that are quadratic in W, one cannot be certain that the log-likelihood is free of nonglobal local maxima. So, in general, one needs to be aware that one could get trapped in such a local maximum in the process of looking for the global maximum. The likelihood surface may be better behaved if one ignores the quadratic terms (the final two terms in (3.15b)). We next present an example probability distribution P_s where the likelihood surface has no nonglobal local maxima in the absence of the quadratic terms. We use that fact to find a local maximum of the full log-likelihood that, at least in our tests, gives good results.

3.4. Special case: A Poisson distribution. We present a special case of the results when the activity of each node at each time step is drawn from a Poisson distribution. We use such a distribution because, for small time bins, the averaged model approximates a generic history-dependent point process [4, 5], which one can use to model the spike times of a neuron. Moreover, the results with the Poisson distribution illustrate how history dependence can distinguish common input from causal connections, as discussed below. We use the Poisson model when we demonstrate the results via simulations.

Since we assume that R^i_s, the activity of node s at time bin i, is a Poisson random variable, we simply need to specify its mean. We can write the probability distribution of R^i_s as

$$P_s(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s) = \Gamma(r^i_s, \lambda_s(\mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s)), \tag{3.16a}$$

where

$$\Gamma(n, \lambda) = \frac{1}{n!}\lambda^n e^{-\lambda}. \tag{3.16b}$$

The function $\lambda_s(\mathbf{r}^{<i}_s, \mathbf{x}, w; \theta^i_s)$ defines how the expected value of R^i_s depends on the history $\mathbf{r}^{<i}_s$ of node s, the external variables $\mathbf{x}$, and the total input w from other neurons. We rewrite the log-likelihood (3.15) as

$$\log \Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q} | \mathbf{X} = \mathbf{x}) = \sum_{q,i} r^i_q \log \lambda_q(\mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q) - \sum_{q,i} \lambda_q(\mathbf{r}^{<i}_q, \mathbf{x}, \widetilde{W}^i_q; \theta^i_q) + C, \tag{3.17a}$$

where

$$\begin{aligned}\widetilde{W}^i_q = {} & \sum_{\substack{\tilde{q} \\ \tilde{q} \neq q}} \sum_{\substack{j \\ j>0}} W^j_{\tilde{q},q}[r^{i-j}_{\tilde{q}} - E_0(\lambda_{\tilde{q}}(\mathbf{R}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}}))] \\ & + \sum_{\substack{\tilde{q} \\ \tilde{q} \neq q}} \sum_{\substack{j \\ j>0}} U^j_{\tilde{q},q}[r^{i-j}_{\tilde{q}} - \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})] \frac{\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})}{\lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, 0; \theta^{i-j}_{\tilde{q}})} \\ & + \sum_{\substack{\tilde{q} \\ \tilde{q} < q}} U^0_{\tilde{q},q}[r^i_{\tilde{q}} - \lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})] \frac{\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})}{\lambda_{\tilde{q}}(\mathbf{r}^{<i}_{\tilde{q}}, \mathbf{x}, 0; \theta^i_{\tilde{q}})} \\ & + \text{quadratic terms.}\end{aligned} \tag{3.17b}$$

The constant $C = -\sum_{q,i} \log((r^i_q)!)$ can be ignored since we simply want to maximize (3.17) over W and U with everything else fixed. We use the notation $\partial_w \lambda_{\tilde{q}}(\mathbf{r}^{<i-j}_{\tilde{q}}, \mathbf{x}, w; \theta^{i-j}_{\tilde{q}})$ for the partial derivative of $\lambda_{\tilde{q}}(\cdot)$ with respect to w.

The quadratic terms are the last two lines of (3.15b); we gain no insight by rewriting them in terms of the Poisson distribution. As detailed in section 3.3.1, they are needed to have a correct second-order expression. However, they don't directly contribute to the distinction between causal connections and common input.

3.4.1. Different effects of causal connections and common input. From (3.17), we see two important differences in the way that the causal connections W and the common input U affect the probability distribution $\Pr(R_{\mathcal{Q}}|\mathbf{X})$ of the measured node activity. Our ability to successfully distinguish causal connections from common input connections is based on these two differences.

The first difference is that the common input terms have an additional $\partial_w \lambda_{\tilde{q}}/\lambda_{\tilde{q}}$ factor. In previous work [14], this factor was the only difference that appeared because the analysis did not exploit history-dependent effects. As detailed in [14], this difference alone can distinguish causal connections from common input in many cases. Even if one did not model history-dependent effects, the relationship among external variables (such as a stimulus) and the activity of measured nodes would distinguish common input from causal connection, and this difference is captured by the $\partial_w \lambda_{\tilde{q}}/\lambda_{\tilde{q}}$ factor.

The second difference in the way W and U appear in (3.17) is due to the history-dependent effects. This second difference is the focus of this paper. It turns out that this difference is exactly what we observed in the exaggerated example presented in the introduction and illustrated in Figure 2. For both the causal connection W term and the common input U term of (3.17b), a certain quantity is subtracted from the activity $r_{\tilde{q}}^{i-j}$. The difference between these quantities can distinguish a causal connection from common input. In what follows, we will show that the key difference is that the activity predicted by a node's history dependence is subtracted only from the common input term.

Equation (3.17) shows that a causal connection from node $\tilde{q}$ onto node q induces a change in the probability distribution of node q proportional to the deviation of the activity of node $\tilde{q}$ from that predicted by the averaged model (2.3). That is, in the causal connection term (first line of (3.17b)), a contribution is added to $\widetilde{W}_q^i$ when the measured activity of node $\tilde{q}$ (i.e., $r_{\tilde{q}}^{i-j}$) differs from its expected value $E_0(R_{\tilde{q}}^{i-j}) = E_0(\lambda_{\tilde{q}}(\mathbf{R}_{\tilde{q}}^{<i-j}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i-j}))$ given by the averaged model.

An important point is that, once the effective parameters ($\theta_{\tilde{q}}^{\tilde{\imath}}$ for all $\tilde{\imath}$) have been determined, this expected value $E_0(R_{\tilde{q}}^{i-j})$ does not depend on the actual history $\mathbf{r}_{\tilde{q}}^{<i-j}$ of node $\tilde{q}$. This expected value is an average over all possible histories of node $\tilde{q}$, given the effective parameters and the external variables.[7] For example, imagine that $R_{\tilde{q}}^i$ corresponds to the number of spikes of neuron $\tilde{q}$ in time bin i (with a sufficiently small time bin so that $R_{\tilde{q}}^i > 1$ with vanishingly small probability). Imagine, moreover, that (similar to neuron 1 in Figure 2) neuron $\tilde{q}$ tended to spike in pairs so that if it spiked in time bin $i-1$ but not in time bin $i-2$, it was very likely to spike in time bin i: $\Pr(R_{\tilde{q}}^i = 1 | R_{\tilde{q}}^{i-1} = 1 \ \& \ R_{\tilde{q}}^{i-2} = 0) \approx 1$. If one used an appropriate model, then the averaged model (2.3) would capture this tendency to fire in pairs once the parameters $\theta_{\tilde{q}}$ were fit to the spikes $\mathbf{R}_{\tilde{q}}$ of neuron $\tilde{q}$. Even so, the expected value $E_0(R_{\tilde{q}}^i)$ would not depend on the presence or absence of spikes in the previous two time bins; it is independent of the specific history of node $\tilde{q}$. Even if $r_{\tilde{q}}^{i-1} = 1$ and $r_{\tilde{q}}^{i-2} = 0$, the expected value $E_0(R_{\tilde{q}}^i)$ would not be close to one. If indeed $r_{\tilde{q}}^i = 1$, then both the spike at time bin $i-1$ and the spike at time bin i would contribute equally to the causal connection term in the first line of (3.17b).[8]

[7] We calculate this value via Monte Carlo. We repeatedly generate a realization of the activity of node $\tilde{q}$ for all time points according to the averaged model (2.3). The average activity at each time point i over many such realizations is our estimate of $E_0(R_{\tilde{q}}^i)$. See Appendix C.

[8] One would get a similar result if neuron $\tilde{q}$ had a *refractory period* where, for example, it could not spike in time i if it spiked in time bin $i-1$: $\Pr(R_{\tilde{q}}^i = 1 | R_{\tilde{q}}^{i-1} = 1) = 0$. Even if neuron $\tilde{q}$ did spike at time bin $i-1$, the presence of the refractory period would not affect $E_0(R_{\tilde{q}}^i)$.

We contrast this observation with the common input term from the second line of (3.17b). In the common input term, the activity $r_{\tilde{q}}^{i-j}$ of node $\tilde{q}$ is subtracted by the mean $\lambda_{\tilde{q}}$ of the Poisson distribution, given the specific history $\mathbf{r}_{\tilde{q}}^{<i-j}$ measured from node $\tilde{q}$. Unlike the causal connection term, this quantity is the expected value of $R_{\tilde{q}}^{i-j}$, *conditioned on the measured history* $\mathbf{r}_{\tilde{q}}^{<i-j}$: $\lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i-j}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i-j}) = E_0(R_{\tilde{q}}^{i-j} \,|\, \mathbf{R}_{\tilde{q}}^{<i-j} = \mathbf{r}_{\tilde{q}}^{<i-j})$. This is still an expected value based on the averaged model, but it is not an average over all possible histories of node $\tilde{q}$. As above, imagine that node $\tilde{q}$ was a neuron that tended to fire pairs of spikes and that one used a model that accurately captured this firing pattern. Then, if $r_{\tilde{q}}^{i-1} = 1$ and $r_{\tilde{q}}^{i-2} = 0$, the expected value $\lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i})$ would be close to one because the model predicts that neuron $\tilde{q}$ should immediately fire a second spike. If indeed $r_{\tilde{q}}^{i} = 1$, this spike would have little contribution to the second line of (3.17b) as $r_{\tilde{q}}^{i} - \lambda_{\tilde{q}}(\mathbf{r}_{\tilde{q}}^{<i}, \mathbf{x}, 0; \theta_{\tilde{q}}^{i})$ would be small.

Equation (3.17) therefore demonstrates that the intuition we gain from the exaggerated example of Figure 2 is applicable to the more realistic situation we used to derive (3.17). For example, although it isn't intuitively obvious what should happen when all nodes have strong history dependence, (3.17) shows that one may estimate the connectivity even in that case, provided one has a model through which one can accurately capture the history dependence of the measured nodes.

3.4.2. Tractable computation of maximum likelihood estimators. In order to efficiently compute maximum likelihood estimators of W and U, we'd like to make sure that any local maxima of the log-likelihood (3.17) are indeed global maxima. As discussed below, if one ignores the quadratic terms in (3.17b), one can develop a condition on the form of λ_s to ensure all local maxima are global maxima. One can then use the solution to the reduced problem (without quadratic terms) to guide the search for a solution to the full problem.

If we ignore the quadratic terms from (3.17b), then $\widetilde{W}_q^i$ is linear in W and U, and the log-likelihood (3.17) has the same form of dependence on W and U as discussed in [14]. Since concavity is preserved under addition and $r_q^i \geq 0$, the log-likelihood will be concave in W and U if $\lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, w; \theta_q^i)$ is convex in w and $\log \lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, w; \theta_q^i)$ is concave in w. Reference [17] describes this condition in a more general setting and outlines the ensuing requirements on λ_q, such as the fact that λ_q must be monotonically increasing in w and must grow at least linearly in w. If the log-likelihood is concave in W and U, there can be no nonglobal local maxima.

We base our search for a maximizer of the full log-likelihood (3.17) on the maximizer for reduced log-likelihood (ignoring the quadratic terms of (3.17b)). We form a homotopy from the reduced problem to the full problem by multiplying the quadratic terms by some number $\gamma \in [0, 1]$. After maximizing the reduced log-likelihood ($\gamma = 0$), we form a series of log-likelihoods with increasing γ. For each problem, we use the maximizer of the previous problem as the initial condition. We end up with a maximizer of the full log-likelihood ($\gamma = 1$). Although we cannot guarantee that we have found a global maximizer, we have achieved good results using this algorithm in our simulation tests. (To calculate the maximizer for a given γ, we iterate to a critical point of (3.17) using a modified version of Powell's hybrid method as implemented in the GNU Scientific Library [6].)

4. Results.

4.1. Overview of simulations. To test the performance of our analysis, we simulated small networks of simplified neurons responding to a stimulus $\mathbf{X}$. Our goal is to demonstrate that we can distinguish the common input and causal connection networks schematized in Figure 1, under the condition that the common input neuron is unmeasured.

4.1.1. The stimulus. The external variables $\mathbf{X}$ represented the same one-dimensional (i.e., constant along vertical lines) visual stimulus as detailed in [14]. This stimulus was a movie of a sequence of sinusoidal gratings $\mathbf{I}^k$ with wave number k. The jth line of $\mathbf{I}^k$ was $I_j^k = \operatorname{cas}(2\pi kj/N_0)$, where $\operatorname{cas} x = \cos x + \sin x$, $N_0 = 100$, and $0 \leq j \leq N_0 - 1$. Every 10 simulated milliseconds, a new image was selected, with replacement, from the set composed of the $\mathbf{I}^k$ and $-\mathbf{I}^k$, for $k = -10, -9, \dots, 9, 10$. The movie was one simulated minute long.

4.1.2. The simulated neurons. In the simulated networks, we let each neuron be a generalized linear model (also called a linear-nonlinear model). We discretized time into $\Delta t_{\text{sim}} = 0.5$ ms time bins. In each time bin i, we let the probability that a neuron spiked be a linear function of its spiking history, the stimulus $\mathbf{X}$, and previous spikes of other neurons, composed with a half-squaring nonlinearity,

$$\begin{aligned} &\Pr(R_p^i = 1 | \mathbf{R}^{<i} = \mathbf{r}^{<i}, \mathbf{X} = \mathbf{x}) \\ &\qquad = A \Delta t_{\text{sim}} \left[\sum_{j>0} \bar{h}^j_{\text{hist},p} r_p^{i-j} + \bar{\mathbf{h}}^i_{\text{ext},p} \cdot \mathbf{x} + \sum_{q \neq p} \sum_{j>0} \bar{W}^j_{q,p} r_q^{i-j} + \bar{y}_p \right]_+^2, \end{aligned} \tag{4.1}$$

where $[y]_+^2 = y^2$ if $y > 0$ and is zero otherwise. The activity variable $R_p^i = 1$ if neuron p spiked in time bin i and $R_p^i = 0$ otherwise. We set $A = 0.01$ ms^{-1}. The value of the threshold parameters $\bar{y}_p$, coupling parameters $\bar{W}^j_{q,p}$, and other parameters that appear below are given in the context of specific simulations. If (4.1) resulted in a probability greater than one, it was truncated to one.

The linear kernel $\bar{\mathbf{h}}_{\text{hist},p}$ specified the spike-history dependence of neuron p. We included a refractory period of length τ_p^{ref} by setting $\bar{h}^j_{\text{hist},p} = -100$ for $j \Delta t_{\text{sim}} \leq \tau_p^{\text{ref}}$. (Since -100 was much larger in magnitude than other parameters in (4.1), $\Pr(R_p^i = 1)$ was zero for an interval of τ_p^{ref} after each spike.) After the refractory period, we let the history-dependent term transiently increase the probability of a spike by setting

$$\bar{h}^j_{\text{hist},p} = a_{\text{hist},p} e^{-j \Delta t_{\text{sim}} / \tau_{\text{hist},p}} \qquad \text{for } j \Delta t_{\text{sim}} > \tau_p^{\text{ref}}.$$

As our purpose is to demonstrate the effect of history dependence, we included strong history dependence in each model neuron, setting $a_{\text{hist},p}$ relatively large and positive. Hence, the history-dependence term created a tendency for spikes to occur in bursts, leading to significant peaks in autocorrelation, such as shown in Figure 3.

We used the same spatiotemporal kernels $\bar{\mathbf{h}}_{\text{ext},p}$ as in [14], retaining the convention that $\bar{\mathbf{h}}^i_{\text{ext},p}$ was the kernel $\bar{\mathbf{h}}_{\text{ext},p}$ shifted for time point i. For line $j = 0, 1, \dots, N_0$ and temporal index t, we used the form

$$\bar{h}_{\text{ext},p}(j,t) = (t - b_p) \exp\left(-\frac{t - b_p}{\tau_{\text{ext},p}} - \frac{(j-c)^2}{2\sigma_p^2} \right) \cos(2\pi f_p (j - c) + \phi_p)$$

for $t > b_p$ and $\bar{h}_{\text{ext},p}(j,t) = 0$ otherwise [10]. To center the kernels on the image, we set $c = (N_0 - 1)/2$. The vector $\bar{\mathbf{h}}_{\text{ext},p}$ corresponded to $\bar{h}_{\text{ext},p}(j, k\Delta t_{\text{sim}})$ for integer k with $k \Delta t_{\text{sim}} < 200$ ms. We normalized $\bar{\mathbf{h}}_{\text{ext},p}$ so that the standard deviation of $\bar{\mathbf{h}}^i_{\text{ext},p} \cdot \mathbf{X}$ was equal to the parameter $a_{\text{ext},p}$; hence, $a_{\text{ext},p}$ specified how strongly neuron p responded to the stimulus.

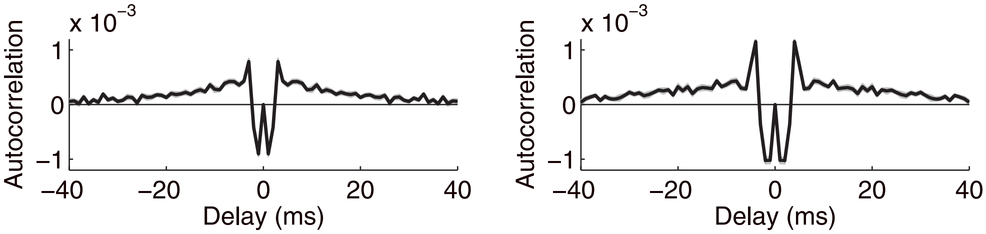


FIG. 3. *Examples of the large autocorrelations due strong history dependence included in simulated models. The autocorrelation of neuron* p *at delay* j *is* $\langle R_p^i R_p^{i-j}\rangle - \langle\langle R_p^i|\mathbf{X}\rangle\langle R_p^{i-j}|\mathbf{X}\rangle\rangle$*, where* $\langle\cdot|\mathbf{X}\rangle$ *indicates averaging over all repeats of the stimulus and* $\langle\cdot\rangle$ *indicates averaging over all time points. Shown are the autocorrelations from neuron* 1 *(left) and neuron* 2 *(left) in the simulation of Figure* 4(A)*. Autocorrelation at zero delay has been truncated to zero.*

We used interneuronal coupling of the form

$$\bar{W}_{pq}^j = B_{pq}\frac{j\Delta t_{\mathrm{sim}} - d_{pq}}{\tau_w^2}\exp\left(-\frac{j\Delta t_{\mathrm{sim}} - d_{pq}}{\tau_W}\right)$$

for $j\Delta t_{\mathrm{sim}} > d_{pq}$ and $\bar{W}_{pq}^j = 0$ otherwise. Hence, d_{pq} represented the delay and B_{pq} the strength of the connection. For all connections, we set the time scale to $\tau_W = 0.5$ ms.

4.1.3. The model used in the analysis. We also used a generalized linear model (or linear-nonlinear model) for the analysis. (In [14], we test an earlier version of the analysis for stronger deviations from the simulated model.) In the analysis, we uses a temporal discretization of $\Delta t = 1$ ms.

We modeled the activity of each neuron in time bin i as a Poisson distribution (section 3.4) with the expected value given by

$$(4.2)\quad \lambda_s(\mathbf{r}_s^{<i}, \mathbf{x}, w; \theta_s^i) = A_s \log\left(1 + \exp\left[\sum_{j>0} h_{\mathrm{hist},s}^j r_s^{i-j} + \mathbf{h}_{\mathrm{ext},s}^i \cdot \mathbf{x} + w + y_s\right]\right).$$

The parameters θ_s^i correspond to A_s and y_s, as well as the parameters within $\mathbf{h}_{\mathrm{hist},s}$ and $\mathbf{h}_{\mathrm{ext},s}$. We used this form of the nonlinearity so that λ_s would be convex and $\log\lambda_s$ would be concave, a requirement for tractable numerical computations discussed in section 3.4.2 and [17]. We discuss how to determine the parameters θ_s^i in Appendix B.

4.2. Distinguishing common input from direct connection. We simulated two networks analogous to those schematized in Figure 1. In the first network, neuron 2 had a direct connection onto neuron 1. In the second network, a third, unmeasured neuron had a direct connection onto both neurons 1 and 2, with a longer delay onto neuron 1. In both cases, the spikes of neuron 1 were correlated with a delayed version of the spikes of neuron 2.

We simulated the response of each network to ten repetitions of the minute-long movie described above. Then we set the thresholds $\bar{y}_p$ so that each neuron spiked approximately 1,000 times during each presentation of the movie, obtaining approximately 10,000 spikes per neuron. The spikes from the third, common input neuron were discarded, as we treated that neuron as an unmeasured neuron.

Since we analyze just the spikes of two neurons, we will plot both the causal connection factor W and the common input factor U as a function of the delay j

defined as spike time of neuron 1 minus spike time of neuron 2. Hence, our plots will use the convention

$$W^j = \begin{cases} W_{12}^{-j} & \text{for } j < 0, \\ 0 & \text{for } j = 0, \\ W_{21}^{j} & \text{for } j > 0, \end{cases}$$

$$U^j = \begin{cases} U_{12}^{-j} & \text{for } j \leq 0, \\ U_{21}^{j} & \text{for } j > 0. \end{cases}$$

As shown in Figure 4, we were able to successfully distinguish the common input network from the direct connection network, despite the fact that the correlations between neurons 1 and 2 looked the same in both cases. The causal connection measure W was positive in the direct connection network; the common input measure U was positive in the common input network.

Section 3.4.1 outlines two differences between causal connections and common input that our analysis exploits to make this distinction. Only one of those differences was due exclusively to the history-dependence modeling that is the focus of this paper. To test the relative importance of the history-dependent factor, we reanalyzed the simulation of Figure 4 while ignoring any history-dependent effects. We set $\mathbf{h}_{\text{hist},p}$ in (4.2) to zero, essentially reverting our analysis back to an earlier version [14]. In this case, we model the expected activity of a node as independent of its measured history (conditioned on the external variables $\mathbf{X}$), so we remove the difference between the causal connection and common input terms of (3.17b) that is attributed to this history dependence.

The results after ignoring history-dependent effects (not shown) differed only slightly from the results when employing the full model. As in Figure 4, W was positive in the direct connection network, and U was positive in the common input network. Note that the simulations were generated with strong history dependence (yielding autocorrelations as in Figure 3). The fact that we achieved good results even while assuming no history dependence indicates that the analysis is at least somewhat robust to deviations from model assumptions.

4.3. Improvement from modeling history dependence. Although the above results do indicate that the analysis that includes history dependence can succeed in distinguishing causal connections from common input, we wish to demonstrate that we have gained analytic power from our history-dependent modeling. Adding history-dependent effects to our modeling introduced significant complexity compared to an earlier version of the analysis [14]. To justify such complexity, we must demonstrate an improved ability to distinguish connectivity.

One limitation of earlier versions [13, 14] of this analysis is that they require that the neural activity be strongly related to measurable external variables (such as stimuli) in a manner that one can capture with a model. In many experimental contexts, such as when recording from brain areas that are not closely linked to a stimulus, such a strong relationship between external variables and neuronal activity may not be available. In such cases, the earlier versions of the analysis may not apply. On the other hand, if such neurons have a strong history dependence that can be captured by a model, the additional handle provided by history-dependent modeling may allow one to apply the analysis to these systems.

To demonstrate how the history-dependent modeling can improve the results, we repeated the simulation of Figure 4 but weakened the relationship between the

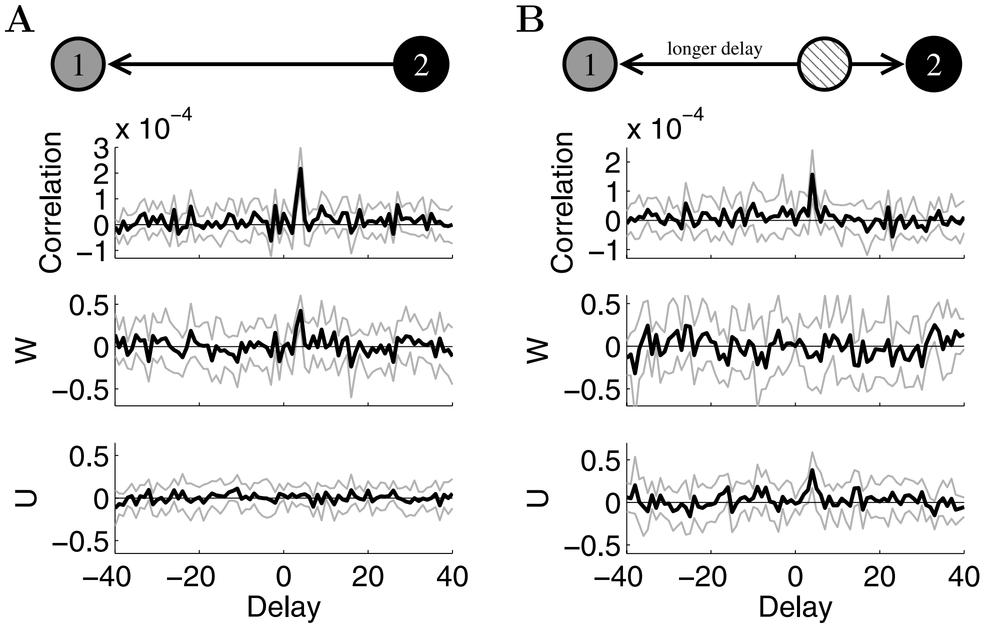


FIG. 4. *Successfully distinguishing a causal connection from common input.* (A) *Results from analyzing a network where neuron* 2 *has a direct connection onto neuron* 1*, as schematized at top. The correlation (shuffle-corrected correlogram or covariogram* [19, 1, 16]*) at delay* j *is* $\langle R_1^i R_2^{i-j}\rangle - \langle\langle R_1^i|\mathbf{X}\rangle\langle R_2^{i-j}|\mathbf{X}\rangle\rangle$*, where the averaging* $\langle\cdot\rangle$ *is defined as in Figure* 3*. The direct connection leads to a peak in the correlation at a positive delay. The causal connection measure* W *(but not the common input measure* U*) has a positive peak at the same delay, indicating the presence of a causal connection from neuron* 2 *onto neuron* 1*. (At the peak,* W *was seven standard errors from zero.) Thin gray lines indicate a bootstrap estimate of three standard errors, calculated by resampling from the set of stimulus repetitions* 50 *times. Simulation parameters:* $a_{\text{hist},1} = 1.2$, $a_{\text{hist},2} = 1.5$, $\tau_{\text{hist},1} = 10$ ms, $\tau_{\text{hist},2} = 12$ ms, $\bar{y}_1 = 0.5$, $\bar{y}_2 = 0.7$, $b_1 = b_2 = 0$, $a_{\text{ext},1} = a_{\text{ext},2} = 1$, $\tau_{\text{ext},1} = 40$ ms, $\tau_{\text{ext},2} = 50$ ms, $\sigma_1 = 10$, $\sigma_2 = 15$, $f_1 = 0.08$, $f_2 = 0.04$, $\phi_1 = 0$, $\phi_2 = 2\pi/3$, $B_{21} = 1.2$, $d_{21} = 3$, $B_{11} = B_{12} = B_{22} = 0$. (B) *Results from analyzing a network where an unmeasured neuron (hatched circle in schematic at top) has a connection onto neuron* 1 *and onto neuron* 2*. Since the connection onto neuron* 1 *has a longer delay, there is a peak in the correlation at a positive delay that is indistinguishable from a peak in correlation due to a direct connection from neuron* 2 *onto neuron* 1*. Only* U*, and not* W*, has a positive peak at the same delay, indicating that the correlation was due to common input rather than any causal connection from neuron* 2 *onto neuron* 1*. (At the peak,* U *was five standard errors from zero.) Most parameters as in panel A. Exceptions and additional parameters (the unmeasured neuron is indexed by* 3*):* $a_{\text{hist},3} = 1.0$, $\tau_{\text{hist},3} = 6$ ms, $\bar{y}_2 = 0.6$, $\bar{y}_3 = 0.8$, $b_3 = 0$, $a_{\text{ext},3} = 1$, $\tau_{\text{ext},3} = 45$ ms, $\sigma_3 = 20$, $f_3 = 0.06$, $\phi_3 = 4\pi/3$, $d_{31} = 4$, $d_{32} = 0$, $B_{31} = B_{32} = 4.5$, $B_{ij} = 0$ *for all other* i *and* j*.*

neuronal activity and the stimulus. We reduced the magnitude of the external variable terms $\mathbf{h}_{\text{ext},p} \cdot \mathbf{X}$ by a factor of 5 (reducing their standard deviation $a_{\text{ext},p}$ from 1 to 0.2). As this greatly increased the difficulty of the network analysis, we also doubled the simulation length to 20 simulated minutes (20 repeats of the movie), obtaining around 20,000 spikes from each neuron.

The results of the analysis based on the full model (4.2) are shown in Figure 5. Despite the weak dependence on the stimulus, the analysis was still able to determine which network contained the causal connection and which network contained common input.

In this case, since the neurons' activities were only weakly related to the stimulus, the history-dependent effects played a bigger role in determining the connectivity. To

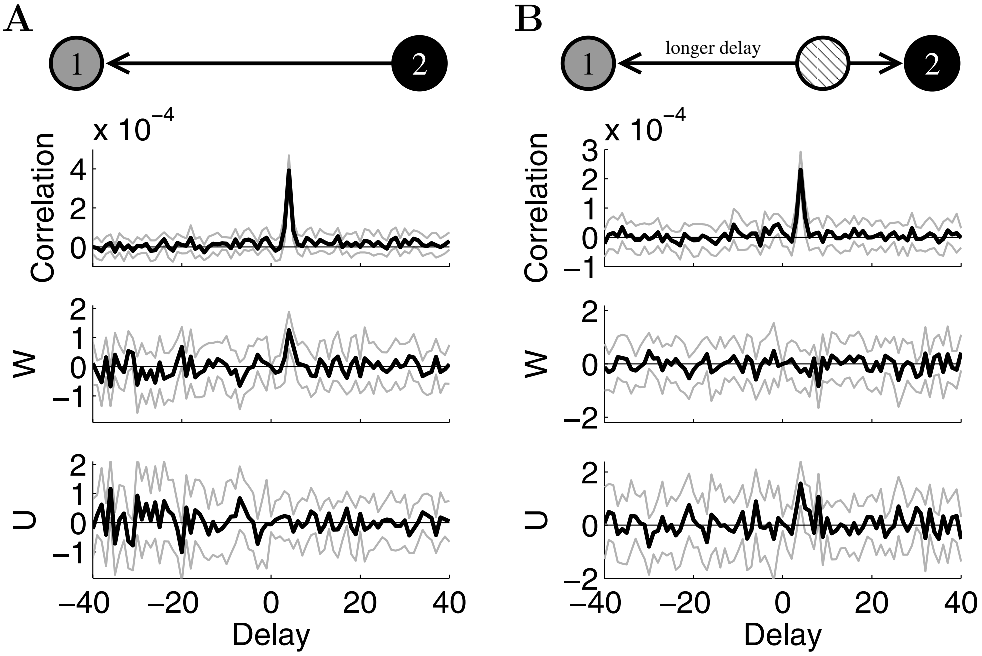


FIG. 5. *Determining circuitry even when neuronal activity is only weakly related to the stimulus. The same networks as in Figure* 4 *were simulated, except that the magnitude of the stimulus input was decreased by a factor of* 5 *and the simulation length was doubled. The causal connection network was still distinguished from the common input network, primarily due to exploitation history-dependent effects (cf. Figure* 6*, where these effects were ignored). Panels as in Figure* 4*.* (A) *The causal connection measure W has a peak at the same delay as the correlation, indicating the correlation was due to a causal connection from neuron* 2 *onto neuron* 1*. (At the peak, W was six standard errors from zero.) Parameters as in Figure* 4(A)*, except that $a_{\text{ext},1} = a_{\text{ext},2} = 0.2$, $\bar{y}_1 = 1.1$, and $\bar{y}_2 = 1.1$.* (B) *The common input measure U has a peak at the same delay as the correlation, indicating the correlation was due to common input. (At the peak, U was five standard errors from zero.) Parameters as in Figure* 4(B)*, except that $a_{\text{ext},1} = a_{\text{ext},2} = a_{\text{ext},3} = 0.2$, $\bar{y}_1 = 1.0$, $\bar{y}_2 = 1.0$, $\bar{y}_3 = 1.2$, and $B_{31} = B_{32} = 4$.*

demonstrate the role of the history-dependent model, we reanalyzed the simulation results of Figure 5 while ignoring history-dependent effects (as above, we set $\mathbf{h}_{\text{hist},p}$ in (4.2) to zero). This time, the analysis was unable to make a clear distinction between the direct connection network and the common input network, as shown in Figure 6. In the direct connection network of Figure 6(A), both W and U were positive so that the result was ambiguous. In the common input network of Figure 6(B), only U was positive at the delay corresponding to the correlation, but U was barely above the noise, and the result was much weaker than in Figure 5(B). (If we quadrupled the simulation to 80 simulated minutes, then the network analysis was able to determine the connectivity even with ignoring history-dependent effects.)

4.4. Subpopulation ambiguity. In section 3.3.3, we described an assumption we made about the hidden nodes in order to complete our analysis. We briefly mentioned that this assumption resulted in a certain degree of ambiguity in the identity of causal connections. This ambiguity is described in detail in [14], where we refer to it as *subpopulation ambiguity.*

The nature of the subpopulation ambiguity is illustrated by Figure 7. Here we repeated the simulation of the common input network of Figure 4(B), except we

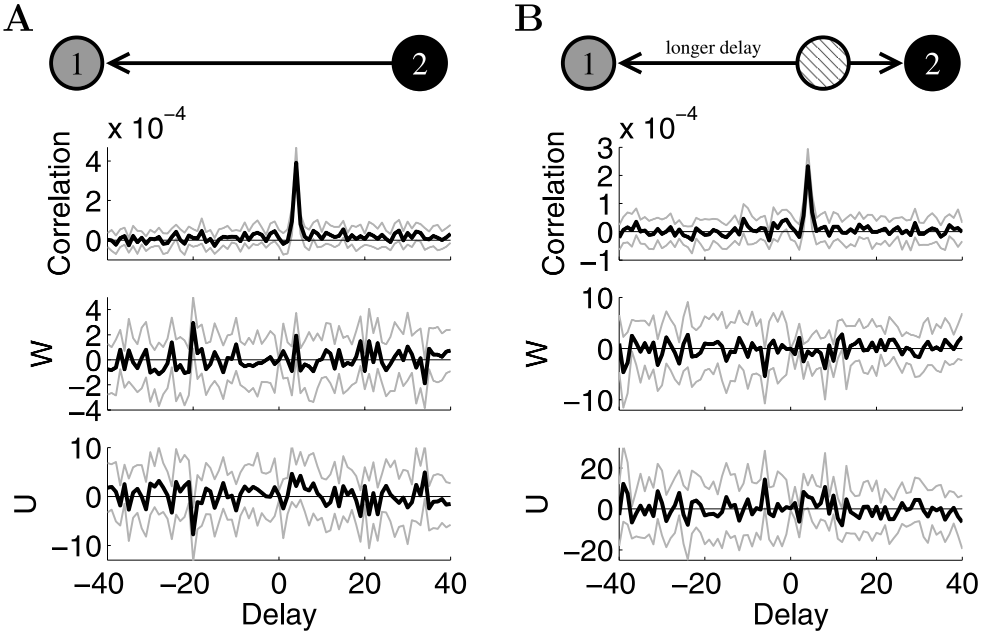


FIG. 6. *Reanalyzing the simulations of Figure* 5 *while ignoring all history-dependent effects. The history kernel* $\mathbf{h}_{\text{hist}}$ *of* (4.2) *was set to zero, so the analysis could not exploit the differences between causal connection and common input networks that are caused by history dependence. In this case, since the neural activity was only weakly related to the stimulus, the analysis failed to cleanly distinguish the circuitry. Panels as in Figure* 4. (A) *The causal connection measure* W *did have a (small) positive peak at the delay of the peak in the correlation. However, the common input measure* U *also had a positive peak at that delay, so that the identity of the causal connection could not be clearly determined. (At the peak* W *was three standard errors and* U *was over two standard errors from zero.)* (B) *Only the causal connection measure* U *had a peak at the delay of the correlation peak, so the results do correctly point to the presence of common input. However, the peak in* U *at that delay is small (though it was three standard errors from zero), especially compared to Figure* 5(B), *indicating that ignoring history dependence hampered the ability to determine circuitry.*

changed the kernel $\bar{\mathbf{h}}_{\text{ext},3}$ of the unmeasured neuron first to match the kernel $\bar{\mathbf{h}}_{\text{ext},2}$ of neuron 2 and then to match the kernel $\bar{\mathbf{h}}_{\text{ext},1}$ of neuron 1. As shown in Figure 7(A), the analysis misidentifies the common input as a causal connection when the kernel of the unmeasured common input neuron matched neuron 2. The analysis does not have any trouble correctly identifying the common input when the kernel of the unmeasured common input neuron matched neuron 1, as shown in Figure 7(B).

We argue that the misidentification in Figure 7(A) merely introduces a relatively modest ambiguity into the interpretation of the results. Clearly, one cannot justify a strict interpretation that the peak in W always indicates a causal connection from neuron 2 itself onto neuron 1. However, note that in the network of Figure 7(A) there is a causal connection from the unmeasured neuron onto neuron 1 and that this unmeasured neuron has similar properties to neuron 2 (one might use the language that the unmeasured neuron has a *receptive field* that is similar to that of neuron 2). Hence, one can make a looser interpretation of the peak in W to indicate the presence of a causal connection onto neuron 1 from some neuron with properties (or receptive field) similar to neuron 2.

In experiments where one measures only the spike times of individual neurons, neurons are identified only by their properties, such as the relationship between their

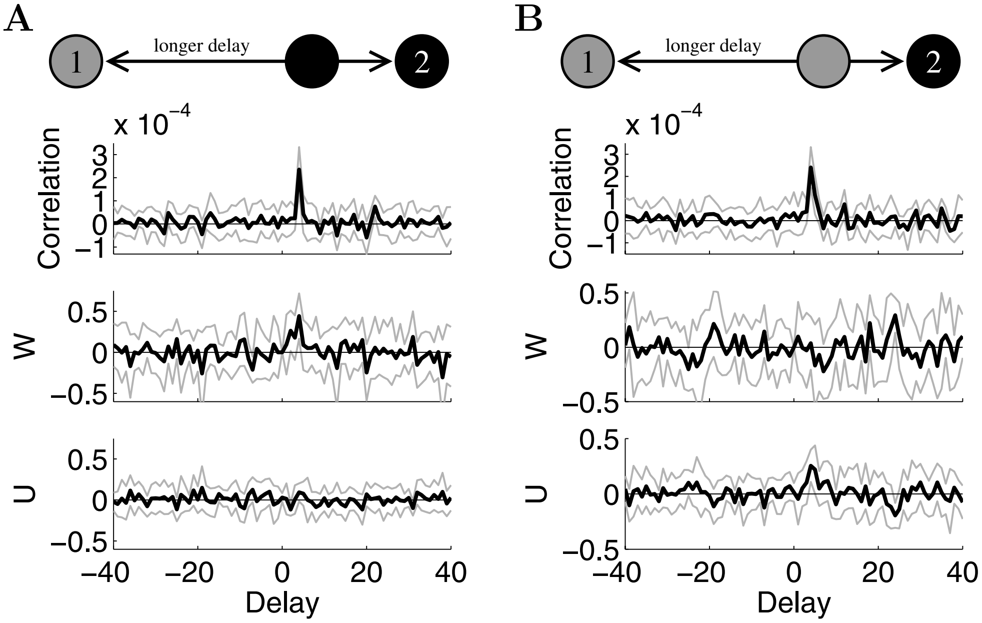


FIG. 7. *An illustration of the subpopulation ambiguity in the identification of the individual neurons involved in a connection. Note that, in both networks shown, the delays are set up so that the correlations mimic a connection from neuron* 2 *onto neuron* 1*. Hence neuron* 1 *and neuron* 2 *do not play symmetric roles. Panels as in Figure* 4. (A) *When the unmeasured neuron has similar properties as neuron* 2 *(as schematized by the black circles at top), the causal connection factor W has a peak at the delay of the correlation peak, incorrectly indicating a connection from neuron* 2 *onto neuron* 1*. (At the peak, W is four standard errors from zero.) However, there is a connection onto neuron* 1 *from a neuron similar to neuron* 2 *(a black neuron in the schematic). Hence, W must be interpreted as indicating a causal connection onto neuron* 1 *from a neuron with properties similar to those of neuron* 2*. Parameters as in Figure* 4(B)*, except that $\bar{y}_1 = 0.4$, $\bar{y}_2 = 0.5$, $\bar{y}_3 = 0.9$, $b_2 = 1$ ms, $\tau_{\text{ext},3} = 50$ ms, $\sigma_3 = 16$, $f_3 = 0.038$, and $\phi_3 = 2\pi/3$.* (B). *When the unmeasured neuron has properties similar to neuron* 1 *(as schematized by the gray circles at top), the results correctly indicate a common input connection as U has a positive peak at the delay of the correlation peak. (At the peak, U is five standard errors from zero.) In this case, it is important that the analysis obtained the correct results, as there is no connection from a neuron similar to neuron* 2 *(a black neuron in the schematic) onto a neuron similar to neuron* 1 *(a gray neuron). Parameters as in Figure* 4(B)*, except that $\bar{y}_1 = 0.4$, $\bar{y}_2 = 0.5$, $\bar{y}_3 = 0.7$, $b_1 = 5$ ms, $\tau_{\text{ext},3} = 40$ ms, $\sigma_3 = 11$, $f_3 = 0.078$, and $\phi_3 = 0$.*

spikes and external variables or stimuli. In this case, if two neurons had similar properties (such as the unmeasured neuron and neuron 2 in Figure 7(A)), those two neurons would be indistinguishable. Hence, it would not make a difference if one concluded that neuron 2 had a connection onto neuron 1 or concluded that an unmeasured neuron with similar properties had a connection onto neuron 1. In either case, the conclusion would be that a neuron with the properties of neuron 2 had a connection onto a neuron with the properties of neuron 1.

In Figure 7(B), there is no connection from a neuron with properties similar to neuron 2 onto a neuron with properties similar to neuron 1. Even with the looser interpretation of the causal connection W, this network cannot be identified as having a causal connection. It is critical that the analysis correctly identified the correlation as arising from common input.

In [14], we refer to a group of neurons with similar properties as a *subpopulation* of neurons. Since the identity of the presynaptic neuron involved in a connection is narrowed down only to an individual member within a supopulation, we use the language that our analysis can determine connectivity only with subpopulation ambiguity. In using such a term, one must be careful to recognize that one is not assuming connections between groups of neurons but only ambiguity in the identity of individual neurons. See [14] for more details, including more intuition behind the subpopulation ambiguity.

5. Discussion. The present work represents a continuation of our development of methods to determine the pattern of causal connections among measured neurons while controlling for the effects of unmeasured neurons [14, 13, 12]. We have successfully eliminated the limitation of earlier versions that the activity of a neuron could depend only weakly on its history. In the process, we have discovered that one can exploit such history dependence to increase one's ability to distinguish common input from causal connections.

Although the analysis involved a fair number of technical manipulations, it turns out that the intuition developed in the introduction does hold for the class of models we consider. In the common input configuration (Figure 2(B)), but not in the causal connection configuration (Figure 2(A)), spikes that can be accounted for by the first neuron's history dependence do not influence the second neuron's spiking probability (see (3.17)). This difference is exploited by our analysis in order to distinguish common input from causal connections.

Successfully exploiting history dependence requires a strong dependence on history in a manner that one can capture by a model. In our simulations, we included such history dependence and demonstrated that we could use it to improve our estimates of connectivity. It is well known that the spike times of neurons are not well approximated by a Poisson process [23, 21] and hence contain history dependence. However, it remains unclear if this history dependence is sufficiently strong and if it can be sufficiently well modeled to aid in the determination of connectivity.

The analysis was justified by a weak coupling assumption (section 2.2) where the original coupling $\bar{W}$ was assumed to be a small parameter. However, even if the original coupling $\bar{W}$ were large and only the perturbation $\widetilde{W}$ due to coupling (see (3.15b)) were small, the analysis might still indicate the effective connectivity of the network. To interpret the analysis under these conditions, one could reinterpret the likelihood equation (3.15) as a perturbation off the effective models (2.3) rather than off the original network (2.2). In this case, one cannot assume that the causal connectivity obtained with W actually corresponds to the underlying connectivity of the network. Such a reinterpretation of W as an effective connectivity would allow application of the results to networks where the weak coupling assumption cannot be justified.

Although the analysis depends on selecting appropriate single-neuron models of the form (2.3), flexibility is given by the modular approach [14] employed in our analysis. One can develop additional single-neuron models and include them in the analysis without modification of the network analysis. In the simulation tests, we used only generalized linear models. Such a model of the dependence of neural activity on spiking history is, of course, only roughly approximated by such a model. One future goal is to implement more sophisticated models of history dependence, such as a stochastic integrate-and-fire model [18]. Paninski, Pillow, and Simoncelli have already developed efficient numerical schemes for determining the parameters of the stochastic integrate-and-fire model [18], and the model does fit into the formalism of (2.3). Such a model may more closely approximate history dependence observed in biological neurons.

Although there is a large literature focused on analyzing interactions among neurons [19, 1, 16, 2, 3, 11, 25, 15, 7, 20, 18, 24, 9], we are aware of only one other attempt to explicitly control for the effects of common input from *unmeasured* sources. Kulkarni and Paninski [9] have recently developed an expectation-maximization algorithm for fitting a neuronal model that contains a latent noise source that could correspond to such unmeasured common input. In their approach, the common input (i.e., latent noise) is assumed to be a Gaussian process (justified by thinking of the common input as a sum of a large number of small inputs). Hence, in place of a point process model (2.2) for a network containing unmeasured neurons, their model is a doubly stochastic process or Cox process [22]. As their approach differs significantly from ours, one future task will be to compare the results of the two methods to understand their relative strengths and weaknesses.

Earlier versions of our analysis relied exclusively on models of the relationship between neuron spikes and external variables such as stimuli. As we have demonstrated via simulations, modeling history-dependent effects may allow one to apply the analysis even in cases where the activity of neurons is not strongly related to external variables. Especially with the implementation of more sophisticated models of history dependence, our analysis may become applicable to a large variety of neuronal systems (or other networks), regardless of whether or not they are strongly linked to a stimulus or other external variable.

Appendix A. Calculations underlying analysis.

A.1. Averaging over hidden node activity. We outline how to simplify (3.4) for the probability distribution $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ of measured node activity $\mathbf{R}_\mathcal{Q}$ by explicitly computing the averages over hidden node activity $\mathbf{R}_\mathcal{P}$. We argued in the context of (3.4) that the value r_s^i of any random variable appears in (3.4) only as a polynomial in r_s^i times $\bar{P}_s$ (or times a derivative of $\bar{P}_s$).

Expressions involving the undifferentiated $\bar{P}_s$ are simple. Let a sum over $\mathbf{r}_s$ denote the sum over all possible values of the activity $\mathbf{r}_s$ (i.e., r_s^i for all i) of a given node s. Then, since $\bar{P}_s$ is shorthand for a probability distribution in the $\mathbf{r}_s$, we can conclude that

$$\text{(A.1)}\quad \sum_{\mathbf{r}_s} \bar{P}_s = 1, \qquad \sum_{\mathbf{r}_s} r_s^i \bar{P}_s = \bar{E}_0(R_s^i), \qquad \text{and} \qquad \sum_{\mathbf{r}_s} r_s^{i_1} r_s^{i_2} \bar{P}_s = \bar{E}_0(R_s^{i_1} R_s^{i_2}).$$

$\bar{E}_0(\cdot)$ denotes the expected value under the probability distribution defined by the P_s with W arguments set to zero, i.e., for any function g of the activity of nodes,

$$\text{(A.2)}\qquad \bar{E}_0(g(\mathbf{R})) = \sum_{\mathbf{r}} g(\mathbf{r}) \prod_s \prod_i P_s\big(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, 0; \bar{\theta}_s^i\big).$$

(The sum over $\mathbf{r}$ indicates the sum over all possible values of the activity of all nodes.) Note that $\bar{E}_0(R_s^i)$ is not the expected value of R_s^i under model (2.2); it is the expected value of R_s^i only if the coupling happened to be zero.

The expressions involving the derivatives of $\bar{P}_s$ are more subtle. First, note that, for any node s,

$$\sum_{\mathbf{r}_s} \prod_i P_s \left(r_s^i, \mathbf{r}_s^{<i}, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}_{\acute{s},s}^{\acute{\imath},i} r_{\acute{s}}^{\acute{\imath}}; \bar{\theta}_s^i \right) = 1$$

independent of any value of the $r^{\tilde{\imath}}_{\tilde{s}}$ for $\tilde{s} \neq s$. (The case when all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$ was the first identity of (A.1).) So, if we differentiate with respect to any $r^{\tilde{\imath}}_{\tilde{s}}$ with $\tilde{s} \neq s$, we will get zero:

$$\sum_{\mathbf{r}_s} \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} \left(\prod_i P_s \left(r^i_s, \mathbf{r}^{<i}_s, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath} < i} \bar{W}^{\acute{\imath},i}_{\acute{s},s} r^{\acute{\imath}}_{\acute{s}}; \bar{\theta}^i_s \right) \right) = 0.$$

In particular, this derivative is zero if we set all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$, so that

$$\text{(A.3)} \qquad \sum_{\mathbf{r}_s} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}} = 0, \qquad \text{and} \qquad \sum_{\mathbf{r}_s} \frac{\partial^2 \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_1}} = 0.$$

On the other hand,

$$E(R^i_s | \{\mathbf{R}^{<i}_{\acute{s}} = \mathbf{r}^{<i}_{\acute{s}}\}_{\acute{s} \neq s}) = \sum_{\mathbf{r}_s} r^i_s \prod_{\acute{\imath}_1} P_s \left(r^{\acute{\imath}_1}_s, \mathbf{r}^{<\acute{\imath}_1}_s, \mathbf{x}, \sum_{\acute{s} \neq s} \sum_{\acute{\imath}_2 < \acute{\imath}_1} \bar{W}^{\acute{\imath}_2,\acute{\imath}_1}_{\acute{s},s} r^{\acute{\imath}_2}_{\acute{s}}; \bar{\theta}^{\acute{\imath}_1}_s \right)$$

does depend on the values of the $r^{\tilde{\imath}}_{\tilde{s}}$ for $\tilde{s} \neq s$ and $\tilde{\imath} < i$. Due to network connections, the expected value of R^i_s could indeed depend on the value of the past activity of another node. So, if we differentiate with respect to any $r^{\tilde{\imath}}_{\tilde{s}}$, we won't necessarily get zero. Denote this derivative, once we set all $r^{\tilde{\imath}}_{\tilde{s}} = 0$ for $\tilde{s} \neq s$, as

$$\text{(A.4)} \qquad \bar{E}_0 \left(\frac{\partial R^i_s}{\partial R^{\tilde{\imath}}_{\tilde{s}}} \right) = \frac{\partial}{\partial r^{\tilde{\imath}}_{\tilde{s}}} E(R^i_s | \{\mathbf{R}^{<i}_{\acute{s}} = \mathbf{r}^{<i}_{\acute{s}}\}_{\acute{s} \neq s}) \Big|_{\{\mathbf{r}_{\acute{s}} = \mathbf{0} | \acute{s} \neq s\}} = \sum_{\mathbf{r}_s} r^i_s \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}}_{\tilde{s}}}.$$

The notation captures that this expression represents how a change in the activity of node $\tilde{s}$ at time $\tilde{\imath}$ affects the average activity of node s at time i. This is nonzero, of course, only if $\tilde{\imath} < i$. Note that this expression doesn't depend on $R^{\tilde{\imath}}_{\tilde{s}}$, as the derivative is calculated around $R^{\tilde{\imath}}_{\tilde{s}} = 0$. Note also that this expression need not be zero even if $R^{\tilde{\imath}}_{\tilde{s}}$ does not directly influence R^i_s, i.e., if $\bar{W}^{\tilde{\imath},i}_{\tilde{s},s} = 0$. Because we have allowed R^i_s to depend arbitrarily on its history $R^{i_2}_s$ for $i_2 < i$, this derivative could be nonzero just because $\bar{W}^{\tilde{\imath},i_2}_{\tilde{s},s} \neq 0$.

We use the identities in (A.1), (A.3), and (A.4) to simplify all of the sums over $\mathbf{r}_{\mathcal{P}}$ in the marginal distribution of $\mathbf{R}_{\mathcal{Q}}$ given in (3.4). To use these identities, we need to distinguish all of the subsets of the various s indicies that could correspond to a hidden node. We do this by enumerating all of the possible ways in which each s index could be either a hidden or a measured node, as well as all of the possible ways in which hidden node indices in a given term could correspond to the same node. Hence each term in (3.4) will be expanded into many different terms.

However, due to the identities in (A.3), most terms involving derivatives of hidden nodes disappear. Recall that a derivative of $\bar{P}_s$ represents a connection onto node s. A connection onto a hidden node p should not directly affect the marginal distribution of the measured nodes $R_{\mathcal{Q}}$; such a connection should have an effect only through a connection from that hidden node onto a measured node. Indeed, the only place where derivatives of hidden nodes survive is in the last term:

$$\frac{1}{2} \sum_{\mathbf{r}_{\mathcal{P}}} \sum_{\substack{s_1, s_2, \tilde{s}_1, \tilde{s}_2 \\ s_2 \neq s_1, \tilde{s}_1 \neq s_1 \\ \tilde{s}_2 \neq s_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{s_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \frac{\partial \bar{P}_{s_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} r^{\tilde{\imath}_1}_{\tilde{s}_1} r^{\tilde{\imath}_2}_{\tilde{s}_2} \prod_{\substack{s_3 \\ s_3 \neq s_1 \\ s_3 \neq s_2}} \bar{P}_{s_3}.$$

If we set s_1 to a hidden node $s_1 = p_1$, then by identity (A.4), the term will survive only if $\tilde{s}_2$ also corresponds to the same hidden node $\tilde{s}_2 = p_1$ and if $\tilde{\imath}_2$ corresponds to an earlier time $\tilde{\imath}_2 < \tilde{\imath}_1$. If we then tried to set s_2 to another hidden node $s_2 = p_2$, we would need the contradictory condition of $\tilde{\imath}_2 > \tilde{\imath}_1$ for the term to survive. In this case, we must set s_2 to a measured node $s_2 = q_2$. Hence, with these substitutions, the term represents the cascade of the effect of a connection from node $\tilde{s}_2$ (which could be hidden or measured) onto hidden node p_1 combined with the effect of a connection from hidden node p_1 onto measured node q_2. (We must double the effect of this term because we could swap s_1 and s_2 and obtain the same result.)

In all other cases, only the effects of connections onto measured nodes survive. The connections from measured and hidden nodes must still be distinguished. We describe this process in [14] as identifying all possible subnetworks of two or fewer edges. When this process is completed, we end up with the following lengthy expression:

$$
\begin{aligned}
\Pr(\mathbf{R}_{\mathcal{Q}} = \mathbf{r}_{\mathcal{Q}} | \mathbf{X} = \mathbf{x}) = {} & \prod_{q} \bar{P}_q + \sum_{\substack{q_1, \tilde{q}_1 \\ q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} r^{\tilde{\imath}_1}_{\tilde{q}_1} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{q_1, \tilde{p}_1} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{q}_1, \tilde{q}_2 \\ q_1 \neq \tilde{q}_1, q_1 \neq \tilde{q}_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1} \partial r^{\tilde{\imath}_2}_{\tilde{q}_2}} r^{\tilde{\imath}_1}_{\tilde{q}_1} r^{\tilde{\imath}_2}_{\tilde{q}_2} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{\substack{q_1, \tilde{q}_1, \tilde{p}_2 \\ q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1} \partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} r^{\tilde{\imath}_1}_{\tilde{q}_1} \bar{E}_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{q_1, \tilde{p}_1, \tilde{p}_2} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1} \partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{p}_1} R^{\tilde{\imath}_2}_{\tilde{p}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{\substack{q_1, \tilde{p}_1, \tilde{p}_2 \\ \tilde{p}_1 \neq \tilde{p}_2}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} \bar{E}_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{p}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{p}_2}} \right) \bar{E}_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \sum_{q_1, \tilde{p}_1, \tilde{q}_2} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} \bar{E}_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{p}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{q}_2}} \right) r^{\tilde{\imath}_2}_{\tilde{q}_2} \prod_{\substack{q_2 \\ q_2 \neq q_1}} \bar{P}_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{p}_1, q_2, \tilde{p}_2 \\ q_2 \neq q_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{p}_1} R^{\tilde{\imath}_2}_{\tilde{p}_2}) \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} \\
& + \sum_{\substack{q_1, \tilde{q}_1, q_2, \tilde{p}_2 \\ q_2 \neq q_1, q_1 \neq \tilde{q}_1}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} r^{\tilde{\imath}_1}_{\tilde{q}_1} \bar{E}_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} \\
& + \frac{1}{2} \sum_{\substack{q_1, \tilde{q}_1, q_2, \tilde{q}_2 \\ q_2 \neq q_1, q_1 \neq \tilde{q}_1 \\ q_2 \neq \tilde{q}_2}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial \bar{P}_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} \frac{\partial \bar{P}_{q_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{q}_2}} r^{\tilde{\imath}_1}_{\tilde{q}_1} r^{\tilde{\imath}_2}_{\tilde{q}_2} \prod_{\substack{q_3 \\ q_3 \neq q_1, q_3 \neq q_2}} \bar{P}_{q_3} + O(\bar{W}^3).
\end{aligned}
\tag{A.5}
$$

A.2. Obtaining an effective parameter equation. We obtained (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ by averaging the full model (2.2) over the activity of all hidden nodes. The averaged model (2.3) is equivalent to the full model (2.2) averaged over the activity of all nodes except for a single node s. Hence, the averaged model (2.3) must be equal to (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$ where the set $\mathcal{Q}$ of measured nodes is replaced by the single node s.

Given definition (3.5) for the effective probability distribution P_s, we obtain

$$\begin{aligned}
P_s &= \Pr(\mathbf{R}_s = \mathbf{r}_s|\mathbf{X} = \mathbf{x}) \\
&= \bar{P}_s + \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) \\
&\quad + \frac{1}{2} \sum_{\substack{\tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s, \tilde{s}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2}) \\
&\quad + \sum_{\substack{\tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s, \tilde{s}_2 \neq s \\ \tilde{s}_1 \neq \tilde{s}_2}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{s}_2}} \right) \bar{E}_0(R^{\tilde{\imath}_2}_{\tilde{s}_2}) \\
&\quad + \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{s}} \right) r^{\tilde{\imath}_2}_s + O(\bar{W}^3),
\end{aligned} \tag{A.6}$$

where we simply replaced all variations of q in (A.5) with s and replaced the $\tilde{p}$ in (A.5) with the corresponding $\tilde{s}$. Equation (A.6) relates the effective parameters θ (hidden in P) to the original model parameters $\bar{\theta}$ (hidden in $\bar{P}$). Since, by assumption, we have an algorithm to determine the effective parameters θ (at least for measured nodes), we want to be able to rewrite everything in terms of the effective parameters. To accomplish this, we need an expression for the original model parameters $\bar{\theta}$ in terms of the effective parameters θ.

Recall that each derivative with respect to r implicitly includes a factor of $\bar{W}$. Hence (A.6) shows that P_s deviates from $\bar{P}_s$ by an amount that is $O(\bar{W})$. Since we are computing only a second-order approximation in $\bar{W}$, we can replace $\bar{P}_s$ with P_s in any terms that are second-order in $\bar{W}$ (i.e., contain two derivatives with respect to r) without affecting the order of our approximation. Similarly expressions with E_0 differ by the equivalent expressions with $\bar{E}_0$ by an amount that is $O(\bar{W})$ (compare (3.7) with (A.2) and (A.4)), so we can also replace $\bar{E}_0$ with E_0 in terms that are second-order in $\bar{W}$. Then (A.6) becomes (after solving for $\bar{P}_s$)

$$\begin{aligned}
\bar{P}_s &= P_s - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial \bar{P}_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} \bar{E}_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}) \\
&\quad - \frac{1}{2} \sum_{\substack{\tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s, \tilde{s}_2 \neq s}} \sum_{\tilde{\imath}_1, \tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1} \partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1} R^{\tilde{\imath}_2}_{\tilde{s}_2}) \\
&\quad - \sum_{\substack{\tilde{s}_1, \tilde{s}_2 \\ \tilde{s}_1 \neq s, \tilde{s}_2 \neq s \\ \tilde{s}_1 \neq \tilde{s}_2}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{\tilde{s}_2}} \right) E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2}) \\
&\quad - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\substack{\tilde{\imath}_1, \tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0 \left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{s}_1}}{\partial R^{\tilde{\imath}_2}_{s}} \right) r^{\tilde{\imath}_2}_s + O(\bar{W}^3).
\end{aligned} \tag{A.7}$$

To write the right-hand side of (A.7) solely in terms of effective parameters θ, we need to change only the sum from the first line. Since this sum is $O(\bar{W})$, we need approximations to $\partial \bar{P}_s / \partial r_{\tilde{s}_1}^{\tilde{\imath}_1}$ and $\bar{E}_0(R_{\tilde{s}_1}^{\tilde{\imath}_1})$ that are accurate to first order in $\bar{W}$. We start with the first-order approximation of $\bar{P}_s$ (the first line of (A.7)):

$$\bar{P}_s = P_s - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1}} E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) + O(\bar{W}^2). \tag{A.8}$$

Here we could replace $\bar{P}_s$ and $\bar{E}_0$ with P_s and E_0 in the terms that are first-order in $\bar{W}$, since we are computing only a first-order approximation.

When we differentiate $\bar{P}_s$ with respect to $r_{\tilde{s}_2}^{\tilde{\imath}_2}$, we are, by (3.6), essentially differentiating with respect to the $\bar{W}_{\tilde{s}_2,s}^{\tilde{\imath}_2,i}$. Hence, if we differentiate the left-hand side of (A.8) with respect to $r_{\tilde{s}_2}^{\tilde{\imath}_2}$, we need to differentiate only those terms on the right-hand side of (A.8) that are functions of P_s or $\bar{P}_s$. We obtain the following expression for the derivative $\partial \bar{P}_s / \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}$ in terms of effective parameters:

$$\frac{\partial \bar{P}_s}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} = \frac{\partial P_s}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} - \sum_{\substack{\tilde{s}_1 \\ \tilde{s}_1 \neq s}} \sum_{\tilde{\imath}_1} \frac{\partial^2 P_s}{\partial r_{\tilde{s}_1}^{\tilde{\imath}_1} \partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) + O(\bar{W}^2). \tag{A.9}$$

To find an expression for $\bar{E}_0(R_{\tilde{s}_1}^{\tilde{\imath}_1})$ in terms of effective parameters, we simplify its definition based on (A.2) to

$$\bar{E}_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) = \sum_{\mathbf{r}_{\tilde{s}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} \bar{P}_{\tilde{s}_1}.$$

We similarly simplify the definition of $E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1})$ (based on (3.7a)) to

$$E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) = \sum_{\mathbf{r}_{\tilde{s}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} P_{\tilde{s}_1}.$$

Then, by using (A.8) along with (3.7b), we can write $\bar{E}_0(R_{\tilde{s}_1}^{\tilde{\imath}_1})$ as

$$\begin{aligned}
\bar{E}_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) &= \sum_{\mathbf{r}_{\tilde{s}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} \bar{P}_{\tilde{s}_1} \\
&= \sum_{\mathbf{r}_{\tilde{s}_1}} r_{\tilde{s}_1}^{\tilde{\imath}_1} P_{\tilde{s}_1} - \sum_{\mathbf{r}_{\tilde{s}_1}} \sum_{\substack{\tilde{s}_2 \\ \tilde{s}_2 \neq \tilde{s}_1}} \sum_{\tilde{\imath}_2} r_{\tilde{s}_1}^{\tilde{\imath}_1} \frac{\partial P_{\tilde{s}_1}}{\partial r_{\tilde{s}_2}^{\tilde{\imath}_2}} E_0(R_{\tilde{s}_2}^{\tilde{\imath}_2}) + O(\bar{W}^2) \\
&= E_0(R_{\tilde{s}_1}^{\tilde{\imath}_1}) - \sum_{\substack{\tilde{s}_2 \\ \tilde{s}_2 \neq \tilde{s}_1}} \sum_{\substack{\tilde{\imath}_2 \\ \tilde{\imath}_2 < \tilde{\imath}_1}} E_0\left(\frac{\partial R_{\tilde{s}_1}^{\tilde{\imath}_1}}{\partial R_{\tilde{s}_2}^{\tilde{\imath}_2}} \right) E_0(R_{\tilde{s}_2}^{\tilde{\imath}_2}) + O(\bar{W}^2).
\end{aligned} \tag{A.10}$$

We substitute (A.9) and (A.10) into the first line of (A.7) and obtain the following second-order expression of $\bar{P}_s$ in terms of effective parameters:

$$\begin{aligned}\bar{P}_s =&P_s - \sum_{\substack{\tilde{s}_1\\ \tilde{s}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})\\ &- \frac{1}{2} \sum_{\substack{\tilde{s}_1,\tilde{s}_2\\ \tilde{s}_1\neq s,\tilde{s}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{s}_1}\partial r^{\tilde{\imath}_2}_{\tilde{s}_2}} [E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}R^{\tilde{\imath}_2}_{\tilde{s}_2}) - 2E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})]\\ &- \sum_{\substack{\hat{s}_2\\ \hat{s}_2\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\hat{s}_2}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\hat{s}_2}}{\partial R^{\tilde{\imath}_2}_{s}}\right)[r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)].\end{aligned}$$

Since for $\tilde{s}_1 \neq \tilde{s}_2$, $E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1}R^{\tilde{\imath}_2}_{\tilde{s}_2}) = E_0(R^{\tilde{\imath}_1}_{\tilde{s}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{s}_2})$, we can simplify this expression to obtain (3.8).

A.3. Measured node activity in terms of effective parameters. Equation (A.5) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, the probability distribution of the measured node activity, is given in terms of the original model parameters $\bar{\theta}$. Our next step is to use (3.8) to rewrite (A.5) in terms of effective parameters θ.

First, we rewrite (3.8) to replace the sums over all nodes in the network by two sums: one over the measured nodes and one over the hidden nodes. Recall that we use q (and its variants) to denote measured node indices and p (and its variants) to denote hidden node indices (i.e., implicitly restrict $q \in \mathcal{Q}$ and $p \in \mathcal{P}$).

$$\begin{aligned}\bar{P}_s = P_s &- \sum_{\substack{\tilde{q}_1\\ \tilde{q}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}) - \sum_{\substack{\tilde{p}_1\\ \tilde{p}_1\neq s}} \sum_{\tilde{\imath}_1} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})\\ &- \frac{1}{2} \sum_{\substack{\tilde{q}_1\\ \tilde{q}_1\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}\partial r^{\tilde{\imath}_2}_{\tilde{q}_1}} [E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1}) - E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\\ &- \frac{1}{2} \sum_{\substack{\tilde{p}_1\\ \tilde{p}_1\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}\partial r^{\tilde{\imath}_2}_{\tilde{p}_1}} [E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1}) - E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\\ &- \sum_{\substack{\tilde{q}_1\\ \tilde{q}_1\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{q}_1}}{\partial R^{\tilde{\imath}_2}_{s}}\right)[r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)]\\ &- \sum_{\substack{\tilde{p}_1\\ \tilde{p}_1\neq s}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_2<\tilde{\imath}_1}} \frac{\partial P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}} E_0\left(\frac{\partial R^{\tilde{\imath}_1}_{\tilde{p}_1}}{\partial R^{\tilde{\imath}_2}_{s}}\right)[r^{\tilde{\imath}_2}_s - E_0(R^{\tilde{\imath}_2}_s)]\\ &+ \frac{1}{2} \sum_{\substack{\tilde{q}_1,\tilde{q}_2\\ \tilde{q}_1\neq s,\tilde{q}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}\partial r^{\tilde{\imath}_2}_{\tilde{q}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_2})\\ &+ \sum_{\substack{\tilde{q}_1,\tilde{p}_2\\ \tilde{q}_1\neq s,\tilde{p}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}\partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_2})\\ &+ \frac{1}{2} \sum_{\substack{\tilde{p}_1,\tilde{p}_2\\ \tilde{p}_1\neq s,\tilde{p}_2\neq s}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_s}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}\partial r^{\tilde{\imath}_2}_{\tilde{p}_2}} E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_2}) + O(\bar{W}^3).\end{aligned}\tag{A.11}$$

The first term on the right-hand side of (A.5) for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ is $\prod_q \bar{P}_q$. This is the only term that is zeroth-order in $\bar{W}$ and so is the only term where we need a second-order conversion from original paremeters $\bar{\theta}$ to effective parameters θ. We derive a second-order approximation of $\prod_q \bar{P}_q$ by taking the product of (A.11) (ignoring terms that are third- or higher-order in $\bar{W}$) and substitute this expression into (A.5). We use a first-order approximation of $\bar{P}_s$, $\bar{E}_0(R_s^i)$ (A.10), and $\partial \bar{P}_s / \partial r_{\tilde{s}}^{\tilde{\imath}}$ (A.9) to rewrite the first-order terms of (A.5) in terms of effective parameters. After simplification, (A.5) becomes

$$
\begin{aligned}
\Pr(\mathbf{R}_\mathcal{Q} = \mathbf{r}_\mathcal{Q}|\mathbf{X} = \mathbf{x}) = {} & \prod_q P_q + \sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}} \sum_{\tilde{\imath}_1} \frac{\partial P_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} [r_{\tilde{q}_1}^{\tilde{\imath}_1} - E_0(R_{\tilde{q}_1}^{\tilde{\imath}_1})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \sum_{\substack{q_1,\tilde{p}_1,\tilde{q}_2\\ \tilde{q}_2\neq q_1}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_1>\tilde{\imath}_2}} \frac{\partial P_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} E_0\left(\frac{\partial R_{\tilde{p}_1}^{\tilde{\imath}_1}}{\partial R_{\tilde{q}_2}^{\tilde{\imath}_2}}\right) [r_{\tilde{q}_2}^{\tilde{\imath}_2} - E_0(R_{\tilde{q}_2}^{\tilde{\imath}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial P_{q_1}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_1}} \frac{\partial P_{q_2}}{\partial r_{\tilde{p}_1}^{\tilde{\imath}_2}} [E_0(R_{\tilde{p}_1}^{\tilde{\imath}_1} R_{\tilde{p}_1}^{\tilde{\imath}_2}) - E_0(R_{\tilde{p}_1}^{\tilde{\imath}_1}) E_0(R_{\tilde{p}_1}^{\tilde{\imath}_2})] \prod_{\substack{q_3\\ q_3\neq q_1, q_3\neq q_2}} P_{q_3} \\
& - \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1} \partial r_{\tilde{q}_1}^{\tilde{\imath}_2}} [E_0(R_{\tilde{q}_1}^{\tilde{\imath}_1} R_{\tilde{q}_1}^{\tilde{\imath}_2}) - E_0(R_{\tilde{q}_1}^{\tilde{\imath}_1}) E_0(R_{\tilde{q}_1}^{\tilde{\imath}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& - \sum_{\substack{q_1,\tilde{q}_1\\ \tilde{q}_1\neq q_1}} \sum_{\substack{\tilde{\imath}_1,\tilde{\imath}_2\\ \tilde{\imath}_1>\tilde{\imath}_2}} \frac{\partial P_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} E_0\left(\frac{\partial R_{\tilde{q}_1}^{\tilde{\imath}_1}}{\partial R_{q_1}^{\tilde{\imath}_2}}\right) [r_{q_1}^{\tilde{\imath}_2} - E_0(R_{q_1}^{\tilde{\imath}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1,q_2,\tilde{q}_2\\ q_2\neq q_1,\tilde{q}_1\neq q_1\\ \tilde{q}_2\neq q_2}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial P_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1}} \frac{\partial P_{q_2}}{\partial r_{\tilde{q}_2}^{\tilde{\imath}_2}} [r_{\tilde{q}_1}^{\tilde{\imath}_1} - E_0(R_{\tilde{q}_1}^{\tilde{\imath}_1})][r_{\tilde{q}_2}^{\tilde{\imath}_2} - E_0(R_{\tilde{q}_2}^{\tilde{\imath}_2})] \prod_{\substack{q_3\\ q_3\neq q_1, q_3\neq q_2}} P_{q_3} \\
& + \frac{1}{2} \sum_{\substack{q_1,\tilde{q}_1,\tilde{q}_2\\ q_1\neq\tilde{q}_1,q_1\neq\tilde{q}_2}} \sum_{\tilde{\imath}_1,\tilde{\imath}_2} \frac{\partial^2 P_{q_1}}{\partial r_{\tilde{q}_1}^{\tilde{\imath}_1} \partial r_{\tilde{q}_2}^{\tilde{\imath}_2}} [r_{\tilde{q}_1}^{\tilde{\imath}_1} - E_0(R_{\tilde{q}_1}^{\tilde{\imath}_1})][r_{\tilde{q}_2}^{\tilde{\imath}_2} - E_0(R_{\tilde{q}_2}^{\tilde{\imath}_2})] \prod_{\substack{q_2\\ q_2\neq q_1}} P_{q_2} + O(\bar{W}^3).
\end{aligned}
$$

(A.12)

A.4. Transforming back to probability distribution. Equation (A.12) is a second-order approximation to a probability distribution, but it is not exactly a probability distribution. Since we wish to use $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ to compute maximum likelihood estimators of coupling parameters (i.e., find values of certain parameters that maximize $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$), we need to use an expression for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$ that is a true probability distribution. For most terms of (A.12), one can simply reverse the Taylor expansion to pull terms back into the product of P_q.

However, one cannot simply reverse the Taylor expansion for the common input terms, i.e., the third line (common input from a hidden node onto two measured nodes) and the fourth line ("common input" from a measured node onto a single measured node). For those two terms, we'll need to tease apart the effects from different time points. We use the notation defined in (3.9) for P_s^i, the probability distribution at a single time point i (as well as its second derivative, defined analogously by (3.9)). We rewrite the derivatives with respect to r in terms of the P_s^i and its derivatives. We also separate out the common input effects at a single time point, rewriting the third and fourth lines of (A.12) as[9]

$$
\begin{aligned}
&+\frac{1}{2}\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\frac{\partial P_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{p}_1}}\frac{\partial P_{q_2}}{\partial r^{\tilde{\imath}_2}_{\tilde{p}_1}}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\prod_{\substack{q_3\\ q_3\neq q_1,q_3\neq q_2}}P_{q_3}\\
&-\frac{1}{2}\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\frac{\partial^2 P_{q_1}}{\partial r^{\tilde{\imath}_1}_{\tilde{q}_1}\partial r^{\tilde{\imath}_2}_{\tilde{q}_1}}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\prod_{\substack{q_2\\ q_2\neq q_1}}P_{q_2}\\
&=\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\substack{\hat{\imath}_1,\hat{\imath}_2\\ \hat{\imath}_2<\hat{\imath}_1}}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{p}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_2}_{\tilde{p}_1,q_2}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_2}_{q_2}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_2}_{q_2}}\\
&+\frac{1}{2}\sum_{\substack{q_1,\tilde{p}_1,q_2\\ q_2\neq q_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\hat{\imath}_1}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{p}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_1}_{\tilde{p}_1,q_2}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_1}_{q_2}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1}R^{\tilde{\imath}_2}_{\tilde{p}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{p}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{p}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_1}_{q_2}}\\
&-\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\substack{\hat{\imath}_1,\hat{\imath}_2\\ \hat{\imath}_2<\hat{\imath}_1}}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{q}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_2}_{\tilde{q}_1,q_1}\frac{\partial P^{\hat{\imath}_1}_{q_1}}{\partial w}\frac{\partial P^{\hat{\imath}_2}_{q_1}}{\partial w}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}P^{\hat{\imath}_2}_{q_1}}\\
&-\frac{1}{2}\sum_{\substack{q_1,\tilde{q}_1\\ q_1\neq\tilde{q}_1}}\sum_{\tilde{\imath}_1,\tilde{\imath}_2}\sum_{\hat{\imath}_1}\bar{W}^{\tilde{\imath}_1,\hat{\imath}_1}_{\tilde{q}_1,q_1}\bar{W}^{\tilde{\imath}_2,\hat{\imath}_1}_{\tilde{q}_1,q_1}\frac{\partial^2 P^{\hat{\imath}_1}_{q_1}}{\partial w^2}[E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1}R^{\tilde{\imath}_2}_{\tilde{q}_1})-E_0(R^{\tilde{\imath}_1}_{\tilde{q}_1})E_0(R^{\tilde{\imath}_2}_{\tilde{q}_1})]\frac{\prod_{q_3}P_{q_3}}{P^{\hat{\imath}_1}_{q_1}}.
\end{aligned}
$$

The last line in the above equation corresponds to the second-order effect of a single connection between two measured nodes. For this term, we cannot reverse the Taylor expansion to fold the term back into the product of the P_q and create a probability distribution. However, this term represents a second-order effect that is not summed over all nodes of the network (it is simply summed over the measured nodes, which we view as a small subset). If we modify our weak coupling assumption to allow us to ignore second-order terms that are not summed over all nodes, we can neglect this last term. Since we no longer have exactly a second-order approximation in $\bar{W}$, we denote the approximation by $\approx$.

With this approximation, we can reverse the Taylor expansion of the remaining terms of (A.12) and obtain (3.10) for $\Pr(\mathbf{R}_{\mathcal{Q}}|\mathbf{X})$, which is written as a probability distribution in terms of effective parameters.

Appendix B. Estimation of single-node parameters. We sketch our algorithm for determining the single-node parameters θ^i_s of model (4.2) that we used to analyze the results of our simulations. The parameters θ^i_s correspond to A_s, y_s, $\mathbf{h}_{\text{hist},s}$, and $\mathbf{h}_{\text{ext},s}$. We calculated maximum likelihood estimators of these parameters from measurements of R^i_s, the spikes of neuron s, and the stimulus $\mathbf{X}$.

We chose our form (4.2) of $\lambda_s(\cdot)$ so that $\lambda_s(\cdot)$ is convex and $\log\lambda_s(\cdot)$ is concave as a function of y_s, $\mathbf{h}_{\text{hist},s}$, and $\mathbf{h}_{\text{ext},s}$. In this way, for a fixed A_s, the log-likelihood surface (logarithm of (2.3)) is free of nonglobal local maxima [17], and we could use standard gradient ascent algorithms to find the maximum, conditioned on a value of A_s. (We used the Polak–Ribiere conjugate gradient algorithm as implemented in the GNU Scientific Library [6].)

[9] Note that all of the probabilities P^i_q that appear in a denominator are also a factor in the corresponding numerator. If a P^i_q that appears in a denominator were to be zero, one could still define the ratio by canceling the factor in the numerator.

Before calculating these maximum likelihood estimators, we calculated the absolute refractory period τ_s^{absref} as the minimum number of $\Delta t = 1$ ms time bins observed between spikes. Then, so that our model predicts absolutely no firing for τ_s^{absref} time steps after each spike, we set $h^i_{\text{hist},s} = -10^{100}$ for $i \leq \tau_s^{\text{absref}}$. To reduce the dimension of the parameter space, we restricted the remainder of the history kernel $\mathbf{h}_{\text{hist},s}$ to be in the subspace spanned by the vectors

$$B^k_{s,1}(i) = \sin\left(\pi k \left[2\frac{i-\tau_s^{\text{absref}}}{\tau_{s,1}} - \left(\frac{i-\tau_s^{\text{absref}}}{\tau_{s,1}}\right)^2\right]\right)$$

for $0 < i - \tau_s^{\text{absref}} < \tau_{s,1}$ and $B^k_{s,1}(i) = 0$ otherwise. (These vector are not orthogonal, so we applied Gram–Schmidt orthonormalization to obtain basis vectors.) We set $\tau_{s,1} = 60 - \tau_s^{\text{absref}}$ time bins. These basis vectors are analagous to those used in [8]; they can represent fine temporal structure for the time immediately after the spike but are smoother for longer time scales. We used 29 basis vectors $1 \leq k \leq 29$ (viewing the 30th basis vector as capturing the absolute refractory period).

We similarly reduced the dimension of $\mathbf{h}_{\text{ext},s}$ by using basis vectors that are a product of a Hartley basis function in space (to match the stimulus) and temporal basis functions similar to the $B^k_{s,1}$. The basis function indexed by k and l evaluated at time bin i and space bin j was based on

$$B^{k,l}_{s,2}(i,j) = \operatorname{cas}(2\pi l j/N_0) \sin\left(\pi k \left[2i/\tau_{s,2} - (i/\tau_{s,2})^2\right]\right)$$

for $0 < i < \tau_{s,2}$ and $B^{k,l}_{s,2}(i,j) = 0$ otherwise (again, we obtained orthogonal basis functions through Gram–Schmidt orthonormalization). As in the definition of the stimulus (section 4.1.1), $\operatorname{cas} x = \cos x + \sin x$ and $N_0 = 100$. We set $\tau_{s,1} = 200$ time bins. We used the 210 basis vectors $-10 \leq l \leq 10$ and $1 \leq k \leq 10$.

As mentioned above, we calculated y_0 and the coefficients of the basis functions to maximize the log-likelihood, given a fixed value of A_s. This defines all parameters as a function of A_s. We then search for a value of A_s that maximizes the log-likelihood while keeping the other parameters set at this function of A_s. We use this procedure since the log-likelihood surface may not be well-behaved as a function of A_s.

Recall that the causal connection measure W and the common input measure U are maximum likelihood estimators based on (3.15), which depends on these values of θ^i_s. To reduce bias at this stage, we calculate the θ^i_s using cross-validation. We divided the data into 4 segments. For each time bin i from one of these segments, we calculated the parameters θ^i_q using only the data in the other 3 segments. (For computation efficiency, we don't recalculate A_s four times but base A_s from calculations using all of the data.)

Appendix C. Monte Carlo estimates of single-node expected values. The estimation of connectivity parameters is based on (3.15) for $\Pr(\mathbf{R}_\mathcal{Q}|\mathbf{X})$, the probability distribution of measured node activity. Once the effective parameters θ_q of the measured nodes have been estimated, the only unknown quantities in (3.15) are the causal connection W and common input U parameters. However, some of the known quantities are given as expected values of functions of the measured node activities as predicted by the averaged model (2.3). Although these expected values are completely determined by the averaged model and the known effective parameters θ_q, computing them explicitly would be impractical, as one would need to enumerate all possible sequences of the history of each node and average over them all.[10] Instead, for each measured node, we use the averaged model (2.3) to randomly generate sequences of activity in order to estimate these expected values using Monte Carlo.

There are three different expected values that appear in (3.15b). They are the average activity $E_0(R_q^i)$ at a given time bin, the second moment $E_0(R_q^i R_q^{i-j})$, and the expected value involving the derivative $E_0(R_q^i(\partial P_q^{i-j}/\partial w)/P_q^{i-j})$. To estimate these expected values via Monte Carlo, we randomly generate a sequence $\mathbf{R}_q$ of the activity of the node from the averaged model (2.3). Then, at each time point i (ignoring initial time points for which we don't have enough preceeding history), we make a sample estimate of each expected value, as described below. We repeat this process 1000 times, setting our final estimates to be averages of these 1000 samples.

To compute the average activity $E_0(R_q^i)$, we could simply record the sampled R_q^i and average these. However, we improve our estimate by taking advantage of the fact that we have an analytic expression for the mean of R_q^i conditioned on the history $\mathbf{R}_q^{<i}$ (for the Poisson distribution it is simply $\lambda_q(\mathbf{R}_q^{<i}, \mathbf{x}, 0; \theta_q^i)$). Our estimate of $E_0(R_q^i)$ is the average of such conditioned means.

In our examples, we use the Poisson distribution (section 3.4) for the probability distribution $P_q(R_q^i, \mathbf{R}_q^{<i}, \cdot)$ of R_q^i conditioned on the history $\mathbf{R}_q^{<i}$. However, one must remember that R_q^i no longer has a probability distribution of the form $P_q(R_q^i, \mathbf{R}_q^{<i}, \cdot)$ once one averages over all possible histories. Since R_q^i does not have a Poisson distribution, one must resist the urge to estimate the variance $E_0((R_q^i)^2) - E_0(R_q^i)E_0(R_q^i)$ as being equal to the mean $E_0(R_q^i)$. Instead, one must calculate $E_0((R_q^i)^2)$ in the same manner as that described above for calculating $E_0(R_q^i)$. Since we have an analytic formula for the second moment of R_q^i conditioned on this history $\mathbf{R}_q^{<i}$ (for the Poisson distribution, it is $\lambda_q^2 + \lambda_q$), we can estimate $E_0((R_q^i)^2)$ as the average of such conditioned second moments. To estimate $E_0(R_q^i R_q^{i-j})$ (for $j > 0$), we take our analytic expression for the average of R_q^i conditioned the history $\mathbf{R}_q^{<i}$, multiply it by the sampled value of R_q^{i-j}, and average over all samples.

For the derivative term, $E_0(R_q^i(\partial P_q^{i-j}/\partial w)/P_q^{i-j})$, we first look at the $j = 0$ case. We can rewrite it as

$$E_0\left(R_q^i \frac{\partial P_q^i}{\partial w}\frac{1}{P_q^i}\right) = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^i}{\partial w}\frac{1}{P_q^i}\prod_{\tilde{\imath}\le i} P_q^{\tilde{\imath}} = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^i}{\partial w}\prod_{\tilde{\imath}<i} P_q^{\tilde{\imath}}, \tag{C.1}$$

where the sum is over all possible values of the r_q^k for $k \le i$. At least for the Poisson distribution, we can calculate an analytic expression[11] for $\sum_{r_q^i} r_q^i \partial P_q^i/\partial w$, and we take the average of that quantity over all samples. For $j > 0$, the term is

$$E_0\left(R_q^i \frac{\partial P_q^{i-j}}{\partial w}\frac{1}{P_q^{i-j}}\right) = \sum_{\mathbf{r}_q^{<i+1}} r_q^i \frac{\partial P_q^{i-j}}{\partial w}\frac{1}{P_q^{i-j}}\prod_{\tilde{\imath}\le i} P_q^{\tilde{\imath}}. \tag{C.2}$$

[10] Hence, the computational cost would increase exponentially in length of the history that could affect the activity.

[11] For the Poisson distribution, $\sum_{r_q^i} r_q^i \partial P_q^i/\partial w = \partial_w \lambda_q(\mathbf{r}_q^{<i}, \mathbf{x}, 0; \theta_q^i)$.

In this case, we take the average value of R_q^i conditioned on the sampled history and multiply it by $(\partial P_q^{i-j}/\partial w)/P_q^{i-j}$. We average this quantity over all samples.[12]

Acknowledgments. The author thanks Liam Paninski, Jayant Kulkarni, Jonathan Pillow, and an anonymous reviewer for helpful input.

REFERENCES

[1] A. M. H. J. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm, *Dynamics of neuronal firing correlation: Modulation of "effective connectivity"*, J. Neurophysiol., 61 (1989), pp. 900–917.

[2] E. N. Brown, R. E. Kass, and P. P. Mitra, *Multiple neural spike train data analysis: State-of-the-art and future challenges*, Nat. Neurosci., 7 (2004), pp. 456–461.

[3] E. S. Chornoboy, L. P. Schramm, and A. F. Karr, *Maximum likelihood identification of neural point process systems*, Biol. Cybern., 59 (1988), pp. 265–275.

[4] D. R. Cox and V. Isham, *Point Processes*, Chapman and Hall, New York, 1980.

[5] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 1988.

[6] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi, *GNU Scientific Library Reference Manual*, 2nd ed., Network Theory Ltd., Bristol, United Kingdom, 2003; also available online from http://www.gnu.org/software/gsl/.

[7] K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsáki, *Organization of cell assemblies in the hippocampus*, Nature, 424 (2003), pp. 552–556.

[8] J. Keat, P. Reinagel, R. C. Reid, and M. Meister, *Predicting every spike: A model for the responses of visual neurons*, Neuron, 30 (2001), pp. 803–817.

[9] J. E. Kulkarni and L. Paninski, *Common-Input Models for Multiple Neural Spike-Train Data*, Network: Comput. Neural Syst., to appear.

[10] S. Marcelja, *Mathematical description of the responses of simple cortical cells*, J. Opt. Soc. Amer., 70 (1980), pp. 1297–1300.

[11] L. Martignon, G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia, *Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies*, Neural Comp., 12 (2000), pp. 2621–2653.

[12] D. Q. Nykamp, *Reconstructing Stimulus-Driven Neural Networks from Spike Times*, in Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun, and K. Obermayer, eds., MIT Press, Cambridge, MA, 2003, pp. 309–316.

[13] D. Q. Nykamp, *Revealing pairwise coupling in linear-nonlinear networks*, SIAM J. Appl. Math., 65 (2005), pp. 2005–2032.

[14] D. Q. Nykamp, *A mathematical framework for inferring connectivity in probabilistic neuronal networks*, Math. Biosci., 205 (2007), pp. 204–251.

[15] M. Okatan, M. A. Wilson, and E. N. Brown, *Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity*, Neural Comp., 17 (2005), pp. 1927–1961.

[16] G. Palm, A. M. H. J. Aertsen, and G. L. Gerstein, *On the significance of correlations among neuronal spike trains*, Biol. Cybern., 59 (1988), pp. 1–11.

[17] L. Paninski, *Maximum likelihood estimation of cascade point-process neural encoding models*, Network: Comput. Neural Syst., 15 (2004), pp. 243–262.

[18] L. Paninski, J. W. Pillow, and E. P. Simoncelli, *Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model*, Neural Comp., 16 (2004), pp. 2533–2561.

[19] D. H. Perkel, G. L. Gerstein, and G. P. Moore, *Neuronal spike trains and stochastic point processes*. II. *Simultaneous spike trains*, Biophys. J., 7 (1967), pp. 419–440.

[20] J. R. Rosenberg, A. M. Amjad, P. Breeze, D. R. Brillinger, and D. M. Halliday, *The Fourier approach to the identification of functional coupling between neuronal spike trains*, Prog. Biophys. Mol. Biol., 53 (1989), pp. 1–31.

[12] Here, we couldn't avoid explicitly dividing by P_q^{i-j} because the conditioned expected value of R_q^i depended on the particular value of R_q^{i-j} that we randomly generated. Note that P_q^{i-j} must be greater than zero for this value of R_q^{i-j} because R_q^{i-j} was randomly generated with probabilty P_q^{i-j}.

[21] M. N. Shadlen and W. T. Newsome, *The variable discharge of cortical neurons: implications for connectivity, computation, and information coding*, J. Neurosci., 18 (1998), pp. 3870–3896.

[22] D. Snyder and M. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, Berlin, 1991.

[23] C. F. Stevens and A. M. Zador, *Input synchrony and the irregular firing of cortical neurons*, Nat. Neurosci., 1 (1998), pp. 210–217.

[24] L. Stuart, M. Walter, and R. Borisyuk, *The correlation grid: Analysis of synchronous spiking in multi-dimensional spike train data and identification of feasible connection architectures.*, Biosystems, 79 (2005), pp. 223–234.

[25] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, *A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects*, J. Neurophysiol., 93 (2005), pp. 1074–1089.

SIAM J. APPL. MATH.
Vol. 68, No. 2, pp. 392–412

INVERSION OF SPHERICAL MEANS AND THE WAVE EQUATION IN EVEN DIMENSIONS[*]

DAVID FINCH[†], MARKUS HALTMEIER[‡], AND RAKESH[§]

Abstract. We establish inversion formulas of the so-called filtered back-projection type to recover a function supported in the ball in even dimensions from its spherical means over spheres centered on the boundary of the ball. We also find several formulas to recover initial data of the form $(f, 0)$ (or $(0, g)$) for the free space wave equation in even dimensions from the trace of the solution on the boundary of the ball, provided that the initial data has support in the ball.

Key words. spherical means, thermoacoustic tomography, wave equation

AMS subject classifications. 35R30, 35L05, 35Q05, 92C55, 65R32

DOI. 10.1137/070682137

1. Introduction and statement of results.

The problem of determining a function from a subset of its spherical means has a rich history in pure and applied mathematics. Our interest in the subject was provoked by the new medical imaging technologies called thermoacoustic and photoacoustic tomography. The idea behind these [10, 16] is to illuminate an object by a short burst of radiofrequency or optical energy which causes rapid (though small in magnitude) thermal expansion which generates an acoustic wave. The acoustic wave can be measured on the periphery or in the exterior of the object. The inverse problem we consider is to find the distribution of the absorbed energy throughout the body. This is of interest, since the amount of energy absorbed at different points may be diagnostic of disease or indicative of uptake of probes tagged to metabolic processes or gene expression [9]. For a more thorough discussion of the modeling and biomedical applications, the reader is referred to the recent survey [17]. If the illuminating energy is impulsive in time, the propagation may be modeled as an initial value problem for the wave equation. The problem of recovering the initial data of a solution of the wave equation from the value of the solution on the boundary of a domain is of mathematical interest in every dimension, but for the application to thermo-/photoacoustic tomography it would appear that the three dimensional case is the only one of interest, since sound propagation is not confined to a lower dimensional submanifold. However, there exist methods of measuring the generated wave field which do not rely on point measurements of the sort that would be generated by an (idealized) acoustic transducer. In particular, integrating line detectors, which have been studied in [3, 14], in effect compute the integral of the acoustic wave field along a specified line. In this paper, we work under the assumption that the speed of sound, c, is constant throughout the body, and since the x-ray transform in a given direction of a solution of the three dimensional wave equation is a solution of the two dimensional wave equation, the problem is

[*] Received by the editors February 7, 2007; accepted for publication (in revised form) June 25, 2007; published electronically November 16, 2007.

http://www.siam.org/journals/siap/68-2/68213.html

[†] Department of Mathematics, Oregon State University, Corvallis, OR 97331 (finch@math.oregonstate.edu).

[‡] Department of Computer Science, Universität Innsbruck, Technikerstraße 21a, A-6020 Innsbruck, Austria (Markus.Haltmeier@uibk.ac.at).

[§] Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (rakesh@math.udel.edu).

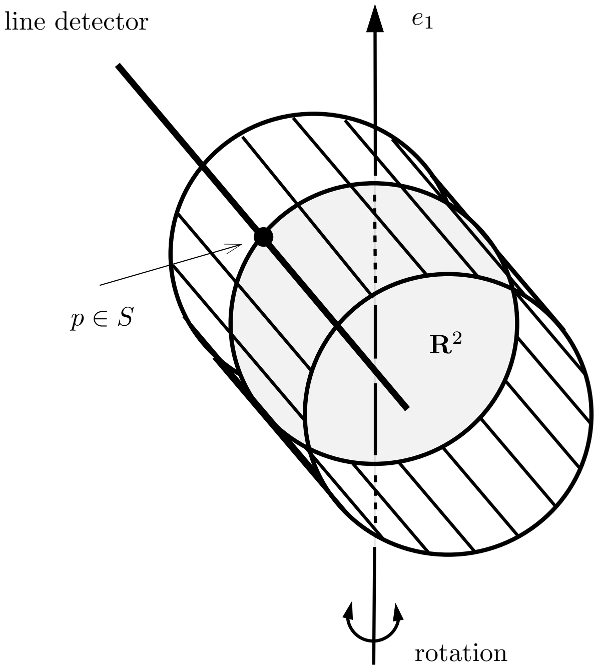


FIG. 1.1. *Principle of thermoacoustic tomography with integrating line detectors. A cylindrical array of line detectors records the acoustic field and is rotated around the axis e_1. For each fixed rotation angle the array outputs the x-ray transform (projection along straight lines) of the trace of the solution of the wave equation restricted to the boundary S of the disk. The initial condition is given by the x-ray transform of the initially induced pressure restricted to lines orthogonal to the base of the cylinder.*

transformed. If a circular array of line detectors is rotated around an axis orthogonal to the direction of the line detectors [7, 14], then for each fixed rotation angle the measurement provides the trace of the solution of the two dimensional wave equation on the circle corresponding to the array. The initial data of this two dimensional problem is the x-ray transform of the three dimensional initial data. If the initial data can be recovered in the disk bounded by the detector array and assuming that the projection of the object to be imaged lies in this disk, then the problem of recovering the three dimensional initial data is reduced to the inversion of the x-ray transform in each plane orthogonal to the axis of rotation. One such two dimensional problem is illustrated in Figure 1.1.

To our knowledge, the first work to tackle the problem of recovering a function from its circular means with centers on a circle was [13], whose author was interested in ultrasound reflectivity tomography. He found an inversion method based on harmonic decomposition and, for each harmonic, the inversion of a Hankel transform. This method has been the basis for most subsequent work on exact inversion of circular means. The inversion of the Hankel transform involves a quotient of a Hankel transform of a harmonic component of the data and a Bessel function. That this quotient be well defined turns out to be a condition on the range of the circular mean transform [2]. See also [1] for range results on the spherical mean transform on functions supported in a ball in all dimensions and [6] for range results for the wave trace map for functions supported in the ball in odd dimensions.

In Finch, Patch, and Rakesh [5], several formulas were found to recover a smooth function f with support in the closure $\overline{B}$ of the open ball $B \subseteq \mathbf{R}^n$ from the trace of the solution of the wave equation on the product $\partial B \times [0, \operatorname{diam}(B)]$, provided that the space dimension is odd. Specifically, if u is the solution of the initial value problem

$$u_{tt} - \Delta u = 0 \quad \text{in } \mathbf{R}^n \times [0, \infty), \tag{1.1}$$

$$u(., t = 0) = f(.), \quad u_t(., t = 0) = 0, \tag{1.2}$$

where f is smooth and has support in the $\overline{B}$, then several formulas were found to recover f from $u(p, t)$ for $p \in S := \partial B$ and $t \in \mathbf{R}^+$.

Finch and Rakesh tried, at that time, to extend the method to even dimensions but did not see a way. Recently, Haltmeier tried numerical experiments using a two dimensional analogue of one of the inversion formulas and found that it gave excellent reconstructions. This prompted our reexamination of the problem. Among the results of this paper is a proof of the validity of this formula.

To describe our results, we introduce some notation. The spherical mean transform $\mathcal{M}$ is defined by

$$(\mathcal{M} f)(x, r) = \frac{1}{|S^{n-1}|} \int_{S^{n-1}} f(x + r\theta) \, dS(\theta) \tag{1.3}$$

for $f \in C^\infty(\mathbf{R}^n)$ and $(x, r) \in \mathbf{R}^n \times [0, \infty)$. In this expression, $|S^{n-1}|$ denotes the area of the unit sphere S^{n-1} in $\mathbf{R}^n$ and $dS(\theta)$ denotes the area measure on the sphere. In general, we write the area measure on a sphere of any radius as dS, except when $n = 2$ when we write ds. We will denote the (partial) derivative of a function q with respect to a variable r by $\partial_r q$, except in a few formulas where the subscript notation q_r is used. At several points we use D_r to denote the operator

$$(D_r u)(r) := \frac{(\partial_r u)(r)}{2r}$$

acting on smooth (even) functions u with compact support. Moreover, r will be used to denote the operator that multiplies a function $u(r)$ by r.

Our first set of results is a pair of inversion formulas for the spherical mean transform in even dimensions. We state and prove these first in dimension two, that is, for the circular mean transform.

Theorem 1.1. *Let $D \subset \mathbf{R}^2$ be the disk of radius R_0 centered at the origin, let $S := \partial D$ denote the boundary circle, and let $f \in C^\infty(\mathbf{R}^2)$ with $\operatorname{supp} f \subset \overline{D}$. Then, for $x \in D$,*

$$f(x) = \frac{1}{2\pi R_0} \Delta_x \int_S \int_0^{2R_0} r \, (\mathcal{M} f)(p, r) \log \left| r^2 - |x - p|^2 \right| dr \, ds(p) \tag{1.4}$$

and

$$f(x) = \frac{1}{2\pi R_0} \int_S \int_0^{2R_0} (\partial_r r \partial_r \mathcal{M} f)(p, r) \log \left| r^2 - |x - p|^2 \right| dr \, ds(p). \tag{1.5}$$

In Theorem 1.1, $\partial_r r \partial_r \mathcal{M} f$ denotes the composition of ∂_r, r, ∂_r, and $\mathcal{M}$ applied to f. The same convention will be used throughout the article to denote the composition of any operators.

While $\mathcal{M} f$ has a natural extension to the negative reals as an even function, we instead take the odd extension in the second variable. In view of the support hypothesis on f, this extension is also smooth. Then formula (1.5) has the following corollary.

COROLLARY 1.2. *With the same hypotheses as in Theorem* 1.1*, and $\mathcal{M} f$ extended as an odd function in the second variable r, f can be recovered for $x \in D$ by*

$$f(x) = \frac{1}{2\pi R_0} \int_S \int_{-2R_0}^{2R_0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{|x-p|-r}\, dr\, ds(p) \tag{1.6}$$

and

$$f(x) = \frac{1}{2\pi R_0} \int_S |x-p| \int_{-2R_0}^{2R_0} \frac{(\partial_r \mathcal{M} f)(p,r)}{|x-p|-r}\, dr\, ds(p), \tag{1.7}$$

where the inner integrals are taken in the principal value sense.

These forms are very close to the standard inversion formula for the Radon transform in the plane [12, eq. (2.5)].

In higher even dimensions we prove a similar pair of results.

THEOREM 1.3. *Let $B \subset \mathbf{R}^n$, $n > 2$ even, be the ball of radius R_0 centered at the origin, let $S := \partial B$ be the boundary of the ball, set*

$$c_n = (-1)^{(n-2)/2} 2((n-2)/2)!\pi^{n/2} = (-1)^{(n-2)/2}[((n-2)/2)!]^2 |S^{n-1}|,$$

and let $f \in C^\infty(\mathbf{R}^n)$ have support in $\overline{B}$. Then, for $x \in B$,

$$f(x) = \frac{1}{c_n R_0} \Delta_x \int_S \int_0^{2R_0} \log\left|r^2 - |x-p|^2\right| (rD_r^{n-2} r^{n-2} \mathcal{M} f)(p,r)\, dr\, dS(p), \tag{1.8}$$

$$f(x) = \frac{2}{c_n R_0} \int_S \int_0^{2R_0} \log\left|r^2 - |x-p|^2\right| (rD_r^{n-1} r^{n-1} \partial_r \mathcal{M} f)(p,r)\, dr\, dS(p). \tag{1.9}$$

Recently, Kunyansky [11] has also established inversion formulas of the filtered back-projection type for the spherical mean transform. His method and results appear to be very different from ours.

For some results, it will be more convenient to use the wave equation (1.1) with initial condition

$$u(.,t=0) = 0, \quad u_t(.,t=0) = f(.). \tag{1.10}$$

It is obvious that the solution of (1.1) with initial values (1.2) is the time derivative of the solution of (1.1) with initial values (1.10). We denote by $\mathcal{P}$ the operator which takes smooth initial data with support in $\overline{B}$ to the solution of (1.1), (1.10) restricted to $S \times [0,\infty)$ and by $\mathcal{W}$ the operator taking f to the solution of (1.1), (1.2) restricted to $S \times [0,\infty)$. These operators are simply related by $\mathcal{W} = \partial_t \mathcal{P}$. An explicit representation for $\mathcal{P}$ comes from the well-known formula [4]

$$u(p,t) = \frac{1}{(n-2)!} \partial_t^{n-2} \int_0^t r(t^2-r^2)^{(n-3)/2} (\mathcal{M} f)(p,r)\, dr, \tag{1.11}$$

giving the solution of the initial value problem (1.1), (1.10), in dimension $n \geq 2$. We denote by $\mathcal{P}^*$ and $\mathcal{W}^* = -\mathcal{P}^* \partial_t$ the formal L^2 adjoints of $\mathcal{P}$ and $\mathcal{W}$ mapping from smooth functions $u \in C^\infty(S \times [0,\infty))$ with sufficient decay in the second variable. An explicit expression for $\mathcal{P}^* u$ will be given in section 3.

We have two types of inversion results for the wave equation. The first type is based on the inversion results for the spherical mean transform, since the spherical mean transform itself can be recovered from the solution of the wave equation by solving an Abel-type equation. In dimension two, this approach yields the following result.

Theorem 1.4. *Let $D \subset \mathbf{R}^2$ be the open disc with radius R_0, and let $S := \partial D$ denote the boundary circle. Then there exists a kernel function $K : [0, 2R_0]^2 \to \mathbf{R}$ such that for any $f \in C^\infty(\mathbf{R}^2)$ with support in $\overline{D}$ and any $x \in D$*

$$f(x) = \frac{1}{R_0 \pi^2} \Delta_x \int_S \int_0^{2R_0} (\mathcal{W} f)(p,t) K(t, |x-p|)\, dt\, ds(p). \tag{1.12}$$

An analytic expression for K will be given in section 3.

Theorem 1.4 provides inversion formulas of the filtered back-projection type for reconstruction of f from $(\mathcal{W} f)(p,t) = (\partial_t \mathcal{P} f)(p,t)$ using only data with $t \in [0, 2R_0]$, despite the unbounded support of $\mathcal{W} f$ and $\mathcal{P} f$ in t.

The second type of inversion results holds in all even dimensions and takes the following form.

Theorem 1.5. *Let f be smooth and supported in closure of the ball B of radius R_0 in $\mathbf{R}^{2m}$, and let $\mathcal{P} f$ and $\mathcal{W} f$ be as above. Then, for $x \in B$,*

$$f(x) = -\frac{2}{R_0} \left(\mathcal{P}^* t \partial_t^2 \mathcal{P} f\right)(x), \tag{1.13}$$

$$f(x) = \frac{2}{R_0} \left(\mathcal{W}^* t \mathcal{W} f\right)(x) = -\frac{2}{R_0} \left(\mathcal{P}^* \partial_t t \partial_t \mathcal{P} f\right)(x). \tag{1.14}$$

We will prove (1.13) in dimension $n = 2m = 2$ directly. The higher dimensional case of (1.13), and (1.14) in all dimensions, are consequences of the following trace identities, relating the L^2 inner product of the initial data to the weighted L^2 inner product of the traces of the solutions of the wave equation.

Theorem 1.6. *Let f, g be smooth and supported in the ball B of radius R_0, in $\mathbf{R}^{2m}$ with $m \geq 1$, let $S := \partial B$, and let u (resp., v) be the solution of the initial value problem* (1.1), (1.10) *with initial value f (resp., g). Then*

$$\int_B f(x) g(x)\, dx = -\frac{2}{R_0} \int_S \int_0^\infty t u_{tt}(p,t) v(p,t)\, dt\, dS(p), \tag{1.15}$$

$$\int_B f(x) g(x)\, dx = \frac{2}{R_0} \int_S \int_0^\infty t u_t(p,t) v_t(p,t)\, dt\, dS(p). \tag{1.16}$$

In the proof of this theorem, (1.15) for $n = 2$ follows from (1.13) for $n = 2$, while (1.15) in higher even dimensions is derived from the $n = 2$ case; (1.16) is a consequence of (1.15) in all dimensions. We remark that these identities were already proved in [5] for odd dimensions, and so they hold for all dimensions.

Section 2 is devoted to the proof of the inversion formulas for the spherical mean transform, that is, Theorems 1.1 and 1.3 and Corollary 1.2. Section 3 treats the wave equation and contains the proofs of Theorems 1.4, 1.5, and 1.6. This is followed by a section reporting on the implementation of the various reconstruction formulas of the preceding sections and results of numerical tests in dimension two.

2. Spherical means. In this section we prove the theorems related to the inversion from spherical means and Corollary 1.2. We begin by establishing an elementary integral identity, which is the key to the results in this paper.

PROPOSITION 2.1. *Let $D \subset \mathbf{R}^2$ be the disk of radius R_0, and let $S = \partial D$ be the boundary circle. Then, for $x, y \in D$ with $x \neq y$,*

$$\int_S \log \left| |x-p|^2 - |y-p|^2 \right| ds(p) = 2\pi R_0 \log|x-y| + 2\pi R_0 \log R_0. \tag{2.1}$$

Proof. Let $x \neq y$ both lie in D, and let I denote the integral on the left-hand side of (2.1). Expanding the argument of the logarithm as

$$\left| |x-p|^2 - |y-p|^2 \right| = 2R_0|x-y| \left| \left(\frac{x+y}{2R_0} - \frac{p}{R_0} \right) \cdot \frac{x-y}{|x-y|} \right|,$$

setting $e := \frac{x-y}{|x-y|}$, and writing $p = R_0\theta$ for $\theta \in S^1$, we have

$$I = 2\pi R_0 \log\left(2R_0|x-y|\right) + R_0 \int_{S^1} \log|e\cdot\theta - a|\, d\theta, \tag{2.2}$$

where

$$a = \frac{x+y}{2R_0} \cdot e = \frac{|x|^2 - |y|^2}{2R_0|x-y|}.$$

We note that $|a| < 1$.

Using the parameterization $\theta = \cos(\phi)e + \sin(\phi)e^{\perp}$, the integral term on the right-hand side of (2.2) has the form

$$R_0 \int_0^{2\pi} \log|\cos\phi - a|\, d\phi.$$

Writing $a = \cos\alpha$ and using the *sum to product* trigonometric identity $\cos\phi - \cos\alpha = -2\sin((\phi+\alpha)/2)\sin((\phi-\alpha)/2)$, this is equal to

$$R_0 \int_0^{2\pi} \left(\log 2 + \log|\sin((\phi+\alpha)/2)| + \log|\sin((\phi-\alpha)/2)|\right) d\phi.$$

By periodicity and two linear changes of variable, this reduces to

$$R_0 \int_0^{2\pi} \left(\log 2 + 2\log|\sin(\phi/2)|\right) d\phi = 2R_0\pi \log 2 + 4R_0 \int_0^{\pi} \log\sin u\, du,$$

which is independent of α and hence of x and y. The latter integral can be found in tables and is equal to $-R_0\pi\log 2$, and so the sum is $-2\pi R_0 \log 2$. Substituting in (2.2) gives the desired result. □

Proposition 2.1 is already enough to establish Theorem 1.1.

Proof of Theorem 1.1. Let $f \in C^\infty(\mathbf{R}^2)$ be supported in $\overline{D}$, and let p be any point in $S = \partial B$. Using the definition of $\mathcal{M} f$ and Fubini's theorem, we have that

$$\int_0^{2R_0} (r\,\mathcal{M} f)(p,r) q(r)\, dr = \frac{1}{2\pi} \int_{\mathbf{R}^2} f(p+z) q(|z|)\, dz \tag{2.3}$$

for any measurable function q, provided that the product of functions on the right-hand side is absolutely integrable. Applying this with $q(r) = \log\left|r^2 - |x-p|^2\right|$ and making the change of variables $y = p + z$ gives

$$\int_S \int_0^{2R_0} (r\,\mathcal{M})(f)(p,r) \log\left|r^2 - |x-p|^2\right| dr\, ds(p)$$
$$= \frac{1}{2\pi} \int_S \int_{\mathbf{R}^2} f(y) \log\left||y-p|^2 - |x-p|^2\right| dy\, ds(p).$$

Fubini's theorem again justifies the change of order of integration in the iterated integral on the right-hand side, and so

$$\frac{1}{2\pi} \int_{\mathbf{R}^2} f(y) \int_S \log\left||y-p|^2 - |x-p|^2\right| ds(p)\, dy$$
$$= \frac{2\pi R_0}{2\pi} \int_{\mathbf{R}^2} f(y)(\log|x-y| + \log R_0)\, dy$$

upon application of (2.1). Recalling that for any constant c, $1/(2\pi)\log|x-y| + c$ is a fundamental solution of the Laplacian in $\mathbf{R}^2$, we have

$$f(x) = \frac{1}{2\pi R_0} \Delta_x \int_S \int_0^{2R_0} (r\,\mathcal{M}\,f)(p,r) \log|r^2 - |x-p|^2|\, dr\, ds(p),$$

which proves (1.4).

The second formula, (1.5), has a similar proof. In this case, we use that the spherical means satisfy the Euler–Poisson–Darboux equation [4]

$$(\partial_r^2\,\mathcal{M}\,f)(x,r) + \frac{1}{r}(\partial_r\,\mathcal{M}\,f)(x,r) = (\Delta\,\mathcal{M}\,f)(x,r) = (\mathcal{M}\,\Delta f)(x,r).$$

The left-hand side of the Darboux equation may be written as $(1/r)(\partial_r r \partial_r\,\mathcal{M}\,f)(x,r)$, and so the expression on the right-hand side of (1.5) may be rewritten as

$$\frac{1}{2\pi R_0} \int_S \int_0^{2R_0} (r\,\mathcal{M}\,\Delta f)(p,r) \log\left|r^2 - |x-p|^2\right| dr\, ds(p). \tag{2.4}$$

Again applying (2.3), now with the function $q(r) = r\log\left|r^2 - |x-p|^2\right|$ and Δf instead of f, interchanging the order of integration and using (2.1) shows that the expression (2.4) is equal to

$$\frac{1}{2\pi} \int_{\mathbf{R}^2} \Delta_y f(y)(\log|x-y| + \log R_0)\, dy = f(x),$$

since no boundary terms arise in view of the support hypothesis on f. □

Proof of Corollary 1.2. Let $x \in D$, and let

$$U(p,x) := \int_0^{2R_0} \left(\partial_r r \partial_r\,\mathcal{M}\,f\right)(p,r) \log\left|r^2 - |x-p|^2\right| dr$$

denote the inner integral in (1.5). Taking the support of f into account, writing the logarithm as

$$\log\left|r^2 - |x-p|^2\right| = \log|r - |x-p|| + \log|r + |x-p||\,,$$

and integrating (1.5) by parts leads to

$$U(p,x) = -P.V.\int_0^\infty \frac{(r\partial_r \mathcal{M} f)(p,r)}{r-|x-p|}\,dr - \int_0^\infty \frac{(r\partial_r \mathcal{M} f)(p,r)}{r+|x-p|}\,dr.$$

Here we have used that the distributional derivative of $\log|r|$ is $P.V.\frac{1}{r}$ as well as an ordinary integration by parts. Therefore (1.5) implies that

$$\begin{aligned} f(x) &= \frac{1}{2\pi R_0}\int_S U(p,x)\,ds(p) \\ &= \frac{-1}{2\pi R_0}\int_S\int_0^{2R_0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{r-|x-p|}\,dr\,ds(p) \\ &\quad - \frac{1}{2\pi R_0}\int_S\int_0^{2R_0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{r+|x-p|}\,dr\,ds(p), \end{aligned} \tag{2.5}$$

where the inner integral of the first term on the right-hand side is taken in the principal value sense. The odd extension of $\mathcal{M} f$, $\mathcal{M} f(p,-r) := -\mathcal{M} f(p,r)$, is smooth on $\mathbf{R}$ since $\mathcal{M} f$ vanishes to infinite order at $r=0$ by the support hypothesis on f and $(r\partial_r \mathcal{M} f)(p,r)$ is an odd function in r. Substituting $r=-r$ into the last integral in (2.5) gives

$$\begin{aligned} f(x) = \frac{-1}{2\pi R_0}&\int_S\int_0^{2R_0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{r-|x-p|}\,dr\,ds(p) \\ - \frac{1}{2\pi R_0}&\int_S\int_{-2R_0}^{0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{r-|x-p|}\,dr\,ds(p) \end{aligned}$$

and hence

$$f(x) = \frac{1}{2\pi R_0}\int_S\int_{-2R_0}^{2R_0} \frac{(r\partial_r \mathcal{M} f)(p,r)}{|x-p|-r}\,dr\,ds(p).$$

This is (1.6). To prove (1.7), it suffices to write

$$\frac{r}{|x-p|-r} = -1 + \frac{|x-p|}{|x-p|-r}$$

in (1.6) and to note that $\int_{-2R_0}^{2R_0}(\partial_r \mathcal{M} f)(p,r)\,dr = 0$ by the support hypothesis on f. □

2.1. Proof of Theorem 1.3. We have found several proofs of Theorem 1.3, the extension of Theorem 1.1 to higher even dimensions. The one we present is based on reduction of the higher dimensional problem to the two dimensional case already established. Another, which is not presented in this article, is based on an extension of (2.1) to higher dimensions.

We first observe that by a dilation, we may reduce the problem to the case when f is supported in the unit ball. Tracing through the formulas (1.8) and (1.9) it is routine to verify that scaling from the unit ball to the ball of radius R_0 introduces a factor of R_0. To simplify notation, we shall now suppose that f is supported in the unit ball B. Let Q and N denote the operators

$$(Qf)(x) = \Delta_x \int_S\int_0^2 (rD_r^{n-2}r^{n-2}\mathcal{M} f)(p,r)\log\left|r^2-|x-p|^2\right|\,dr\,dS(p), \tag{2.6}$$

$$(Nf)(x) = \int_S\int_0^2 (rD_r^{n-1}r^{n-1}\partial_r\mathcal{M} f)(p,r)\log\left|r^2-|x-p|^2\right|\,dr\,dS(p), \tag{2.7}$$

which map $f \in C^\infty(\mathbf{R}^n)$ supported in $\overline{B}$ to constant multiples of the right-hand sides of (1.8) and (1.9). Moreover, $\langle f, g\rangle$ denotes the L^2 product of two functions supported in $\overline{B}$. To establish $Qf = c_n f$ and $Nf = (c_n/2)f$ we will use the following auxiliary results.

Proposition 2.2. *Let f, g be smooth and supported in $\overline{B}$. Then*

$$\int_{\mathbf{R}^n} (Qf)(x)g(x)\,dx = \langle Qf, g\rangle = 2\langle f, Ng\rangle = 2\int_{\mathbf{R}^n} f(x)(Ng)(x)\,dx. \tag{2.8}$$

Proof. Let $F = \mathcal{M}f$ and $G = \mathcal{M}g$. We introduce the temporary notation $\tilde{F}(p,r) = rD_r^{n-2}r^{n-2}F(p,r)$. Using the self-adjointness of Δ, applying Fubini's theorem and an n-dimensional analogue of (2.3), we obtain

$$\begin{aligned}
\langle Qf, g\rangle &= \int_B \left(\int_S \int_0^2 (rD_r^{n-2}r^{n-2}F)(p,r)\log\left|r^2 - |x-p|^2\right| dr\,dS(p)\right)(\Delta_x g)(x)\,dx \\
&= |S^{n-1}| \int_S \int_0^2 \left(\int_0^2 \tilde{F}(p,r)\log\left|r^2 - \bar{r}^2\right| (\mathcal{M}\,\Delta_x g)(p,\bar{r})\bar{r}^{n-1}\,d\bar{r}\right) dr\,dS(p) \\
&= |S^{n-1}| \int_S \int_0^2 \left(\int_0^2 \tilde{F}(p,r)\log\left|r^2 - \bar{r}^2\right| dr\right) (\mathcal{M}\,\Delta_x g)(p,\bar{r})\bar{r}^{n-1}\,d\bar{r}\,dS(p) \\
&= |S^{n-1}| \int_S \int_0^2 \left(\int_0^2 \tilde{F}(p,r)\log\left|r^2 - \bar{r}^2\right| dr\right) \partial_{\bar{r}}\bar{r}^{n-1}\partial_{\bar{r}}G(p,\bar{r})\,d\bar{r}\,dS(p). \qquad (2.9)
\end{aligned}$$

To justify the last equation it is used that G satisfies the Euler–Poisson–Darboux equation and the identity $\bar{r}^{n-1}(\partial_{\bar{r}}^2 + \frac{n-1}{\bar{r}}\partial_{\bar{r}}) = \partial_{\bar{r}}(\bar{r}^{n-1}\partial_{\bar{r}})$. Applying the identities $(D_r^{n-2})^* r\log|r^2 - \bar{r}^2| = (-1)^{n-2}rD_r^{n-2}\log|r^2-\bar{r}^2| = rD_{\bar{r}}^{n-2}\log|r^2-\bar{r}^2|$ in two stages to the last expression, this becomes

$$\begin{aligned}
&|S^{n-1}| \int_S \int_0^2 \left(\int_0^2 r^{n-1}F(p,r)D_{\bar{r}}^{n-2}\log\left|r^2-\bar{r}^2\right| dr\right)\left(\partial_{\bar{r}}\bar{r}^{n-1}\partial_{\bar{r}}G(p,\bar{r})\right) d\bar{r}\,dS(p) \\
&\quad= \int_S \int_0^2 \left(\int_B f(y)D_{\bar{r}}^{n-2}\log\left||y-p|^2 - \bar{r}^2\right| dy\right)\left(\partial_{\bar{r}}\bar{r}^{n-1}\partial_{\bar{r}}G(p,\bar{r})\right) d\bar{r}\,dS(p) \\
&\quad= \int_B \left(\int_S \int_0^2 \log\left||y-p|^2 - \bar{r}^2\right| ((D_{\bar{r}}^*)^{n-2}\partial_{\bar{r}}\bar{r}^{n-1}\partial_{\bar{r}}G)(p,\bar{r})\,d\bar{r}\,dS(p)\right) f(y)\,dy
\end{aligned}$$

after applying Fubini's theorem. This is finally seen to be equal to $\langle f, 2Ng\rangle$ since $(D_{\bar{r}}^{n-2})^*\partial_{\bar{r}} = 2\bar{r}(-1)^{n-2}D_{\bar{r}}^{n-1}$. □

We now look at the spherical means of products

$$f(x) = \rho^k\alpha(\rho)\Phi(\theta), \tag{2.10}$$

where $x = \rho\theta$ with $\rho \geq 0$, $\theta \in S^{n-1}$, Φ is a spherical harmonic of degree k, and $\alpha : \mathbf{R} \to \mathbf{R}$ is an even smooth function supported in $[-1,1]$. Let $F := \mathcal{M}f$ be extended to an even function in the second component, and let $\nu = n + 2k$. Then F satisfies the initial value problem (IVP) for the Euler–Poisson–Darboux equation

$$\left(\partial_r^2 F + \frac{n-1}{r}\partial_r F\right)(x,r) = \Delta_x F(x,r), \qquad (x,r) \in \mathbf{R}^n \times \mathbf{R}, \tag{2.11}$$

$$F(x,0) = \alpha(\rho)\rho^k\Phi(\theta), \quad \partial_r F(x,0) = 0, \qquad x \in \mathbf{R}^n, \tag{2.12}$$

and, conversely, any solution of (2.11), (2.12) is the spherical mean of the initial values. The unique solution of (2.11), (2.12) has the form $F(x,r) = \rho^k A(\rho,r)\Phi(\theta)$, where $A(\rho,r)$ is the solution of the IVP

$$(L_n A)(\rho,r) = \left(\partial_\rho^2 A + \frac{\nu-1}{\rho}\partial_\rho A\right)(\rho,r), \qquad (\rho,r)\in \mathbf{R}^2, \tag{2.13}$$

$$A(\rho,0) = \alpha(\rho), \quad \partial_\rho A(\rho,0) = 0, \qquad \rho \in \mathbf{R}. \tag{2.14}$$

Here $(L_n A)(\rho,r) := (\partial_r^2 A + \frac{n-1}{r}\partial_r A)(\rho,r)$.

We recall that the operator D_r satisfies $L_n D_r = D_r L_{n-2}$ and for any $\mu \in \mathbf{N}$

$$\left(\partial_r^2 + \frac{1-\mu}{r}\partial_r\right)(r^\mu w) = r^\mu \left(\partial_r^2 + \frac{1+\mu}{r}\partial_r\right) w,$$

that is, $L_{2-\mu} r^\mu = r^\mu L_{\mu+2}$. So

$$(L_{2-\mu+2\sigma} D_r^\sigma r^\mu w)(r) = (D_r^\sigma L_{2-\mu} r^\mu w)(r) = (D_r^\sigma r^\mu L_{\mu+2} w)(r). \tag{2.15}$$

If we set $\mu = n-2$ and $\sigma = (n-2)/2$ in (2.15), then $\mu+2 = n$ and $2-\mu+2\sigma = 2$. Therefore

$$(L_2 D_r^{(n-2)/2} r^{n-2} w)(r) = (D_r^{(n-2)/2} r^{n-2} L_n w)(r). \tag{2.16}$$

Now we set

$$H(\rho,r) := \frac{1}{((n-2)/2)!}(D_r^{(n-2)/2} r^{n-2} A)(\rho,r). \tag{2.17}$$

Since $A(\rho,r)$ is even in r and D_r corresponds to differentiation with respect to r^2, $H(\rho,r)$ is even in r. Moreover, by (2.14), $H(\rho,0) = \frac{1}{((n-2)/2)!}A(\rho,0)(D_r^{(n-2)/2} r^{n-2}) = \alpha(\rho)$, and therefore from (2.13) and (2.16) it follows that H is the solution of the IVP

$$\left(\partial_r^2 H + \frac{1}{r}\partial_r H\right)(\rho,r) = \left(\partial_\rho^2 H + \frac{\nu-1}{\rho}\partial_\rho H\right)(\rho,r), \qquad (\rho,r)\in\mathbf{R}^2, \tag{2.18}$$

$$H(\rho,0) = \alpha(\rho), \quad \partial_r H(\rho,0) = 0, \qquad \rho\in\mathbf{R}. \tag{2.19}$$

Proposition 2.3. *Let $A_i(\rho,r)$, $i=1,2$, solve* (2.13) *with $n=2$, subject to initial conditions $A_i(\rho,0) = \alpha_i(\rho)$, $\partial_r A_i(\rho,0) = 0$, where α_i are smooth even functions with support in $[-1,1]$ and $\nu \geq 2$ is even. Then*

$$\int_0^1 \rho^{\nu-1}\alpha_1(\rho)\alpha_2(\rho)\,d\rho = -\int_0^2\int_0^2 rA_1(1,r)\partial_{\bar r}\log|r^2-\bar r^2|\,\bar r(\partial_{\bar r}A_2)(1,\bar r)\,d\bar r\,dr. \tag{2.20}$$

Proof. Let $k = (\nu-2)/2$, and let $\Phi(\theta)$ be a nontrivial real circular harmonic of degree k. Then $F_i(x,r) := A_i(\rho,r)\rho^k\Phi(\theta)$ satisfies (2.11), (2.12) for $n=2$, and so is the circular mean of its initial value, $f_i(x) = \alpha_i(\rho)\rho^k\Phi(\theta)$. By (1.4), $f_1 = \frac{1}{2\pi}Qf_1$, and using (2.9) gives

$$\begin{aligned}\langle f_1, f_2\rangle &= \frac{1}{2\pi}\langle Qf_1, f_2\rangle \\ &= \int_S\int_0^2\int_0^2 rF_1(p,r)\log|r^2-\bar r^2|(\partial_{\bar r}\bar r\partial_{\bar r}F_2)(p,\bar r)\,dr\,d\bar r\,ds(p).\end{aligned}$$

Taking into account the form of F_i and that $\rho = 1$ on S, this may be rewritten as

$$\langle f_1, f_2 \rangle = \int_S \Phi^2(p)\, ds(p) \int_0^2 \int_0^2 rA_1(1,r) \log|r^2 - \bar{r}^2| (\partial_{\bar{r}} \bar{r} \partial_{\bar{r}} A_2)(1,\bar{r})\, dr\, d\bar{r}. \tag{2.21}$$

Appealing to the form of $f_i = F(x,0)$,

$$\begin{aligned} \langle f_1, f_2 \rangle &= \int_0^1 \rho(\rho^k \alpha_1)(\rho^k \alpha_2)\, d\rho \int_S \Phi^2(p)\, ds(p) \\ &= \int_0^1 \rho^{\nu-1} \alpha_1(\rho) \alpha_2(\rho)\, d\rho \int_S \Phi^2(p)\, ds(p). \end{aligned} \tag{2.22}$$

Since $\int_S \Phi^2(p)\, ds(p) \neq 0$, a comparison of (2.21) and (2.22) and an integration by parts on the right-hand side of (2.21) establishes (2.20), which completes the proof. □

Proof of Theorem 1.3. Let $\{\Phi_j\}$ be an orthonormal basis for the spherical harmonics on S^{n-1}, and consider f_i, $i = 1, 2$, of the form (2.10) with $\alpha = \alpha_i$ and $\Phi = \Phi_{j_i}$ of possibly different degrees. Let F_i be the even extensions of $\mathcal{M} f_i$ as above. Then by orthogonality, $\langle f_1, f_2 \rangle = 0$ unless $j_1 = j_2$, in which case

$$\langle f_1, f_2 \rangle = \int_0^1 \rho^{\nu-1} \alpha_1(\rho) \alpha_2(\rho)\, d\rho, \tag{2.23}$$

with $\nu = n + 2k$, where k is the degree of Φ_{j_1}. Evaluating $\langle Qf_1, f_2 \rangle$ by (2.9) and using that $F_i = \rho^{k_i} A_i(\rho, r) \Phi_{j_i}$, we see that it is also zero unless $j_1 = j_2$. In this case we have

$$\begin{aligned} \langle Qf_1, f_2 \rangle &= |S^{n-1}| \int_0^2 \int_0^2 (rD_r^{n-2} r^{n-2} A_1)(1,r) \log|r^2 - \bar{r}^2| \tilde{A}_2(1,\bar{r})\, dr\, d\bar{r} \\ &= |S^{n-1}| \int_0^2 \int_0^2 (D_r^{\frac{n-2}{2}} r^{n-2} A_1)(1,r) (D_r^*)^{\frac{n-2}{2}} (r \log|r^2 - \bar{r}^2|) \tilde{A}_2(1,\bar{r})\, dr\, d\bar{r} \\ &= |S^{n-1}| \int_0^2 \int_0^2 r(D_r^{\frac{n-2}{2}} r^{n-2} A_1)(1,r) D_{\bar{r}}^{\frac{n-2}{2}} \log|r^2 - \bar{r}^2| \tilde{A}_2(1,\bar{r})\, dr\, d\bar{r}, \end{aligned} \tag{2.24}$$

where we have abbreviated $\tilde{A}_2 = \partial_{\bar{r}} \bar{r}^{n-1} \partial_{\bar{r}} A_2$ and used

$$(D_r^{(n-2)/2})^* r \log|r^2 - \bar{r}^2| = (-1)^{(n-2)/2} r D_r^{(n-2)/2} \log|r^2 - \bar{r}^2| = r D_{\bar{r}}^{(n-2)/2} \log|r^2 - \bar{r}^2|.$$

Applying the adjoint (distributional derivative) again in (2.24) and using $\hat{A}_1(r)$ to abbreviate $r(D_r^{\frac{n-2}{2}} r^{n-2} A_1)(1,r)$, which depends only on r,

$$\begin{aligned} \langle Qf_1, f_2 \rangle &= |S^{n-1}| \int_0^2 \int_0^2 \hat{A}_1(r) \log|r^2 - \bar{r}^2| (D_{\bar{r}}^{\frac{n-2}{2}})^* (\partial_{\bar{r}} \bar{r}^{n-1} \partial_{\bar{r}} A_2)(1,\bar{r})\, dr\, d\bar{r} \\ &= |S^{n-1}| \int_0^2 \int_0^2 \hat{A}_1(r) \log|r^2 - \bar{r}^2| (-1)^{\frac{n-2}{2}} (\partial_{\bar{r}} D_{\bar{r}}^{\frac{n-2}{2}} \bar{r}^{n-1} \partial_{\bar{r}} A_2)(1,\bar{r})\, dr\, d\bar{r} \\ &= |S^{n-1}| (-1)^{n/2} \int_0^2 \int_0^2 \hat{A}_1(r) \partial_{\bar{r}} \log|r^2 - \bar{r}^2| (D_{\bar{r}}^{\frac{n-2}{2}} \bar{r}^{n-1} \partial_{\bar{r}} A_2)(1,\bar{r})\, dr\, d\bar{r}. \end{aligned}$$

We now use the following identity, which is readily proved by induction:

$$D_{\bar{r}}^{(n-2)/2} \bar{r}^{n-1} \partial_{\bar{r}} q = \bar{r} \partial_{\bar{r}} D_{\bar{r}}^{(n-2)/2} \bar{r}^{n-2} q$$

taking $q = A_2$, and observe that defining H_i by (2.17) with $A = A_i$, it then holds that

$$\langle Qf_1, f_2\rangle = \gamma_n \int_0^2 \int_0^2 rH_1(1, r)\partial_{\bar{r}} \log |r^2 - \bar{r}^2|\, \bar{r}(\partial_{\bar{r}} H_2)(1, \bar{r})\, dr\, d\bar{r}$$

for $\gamma_n = |S^{n-1}|(-1)^{n/2}[((n-2)/2)!]^2$. By (2.18) and (2.19) the H_i satisfy the hypotheses of Proposition 2.3 with initial data α_i, and so by (2.20) the expression on the right-hand side is equal to

$$-\gamma_n \int_0^1 \rho^{\nu-1}\alpha_1(\rho)\alpha_2(\rho)\, d\rho.$$

Thus we have proved that for f_i of the form above

$$\langle Qf_1, f_2\rangle = (-1)^{(n-2)/2}|S^{n-1}|[((n-2)/2)!]^2\langle f_1, f_2\rangle. \tag{2.25}$$

We note that the constant on the right-hand side is c_n of Theorem 1.3. By linearity and orthogonality of spherical harmonics, this still holds when either f_1 or f_2 is replaced by a finite linear combination of such functions. The set of finite linear combinations of functions of form (2.10) is dense in L^2, and so we have $Qf = c_n f$ in L^2 when f is a finite linear combination of functions of the form (2.10). Now let g be smooth with support in the unit ball. Applying Proposition 2.2, it follows that

$$\langle f, Ng\rangle = (1/2)\langle Qf, g\rangle = (c_n/2)\langle f, g\rangle \tag{2.26}$$

for all f as above. Since (2.26) holds for a dense subset of functions f in $L^2(B)$, it implies that $Ng = (c_n/2)g$ almost everywhere in B. However, Ng is easily seen to be a continuous function, and so $Ng = (c_n/2)g$ holds pointwise in B, which is (1.9). But if N is a multiple of the identity, then so is Q, and the proof is complete. □

3. The wave equation. We begin the analysis of recovery of initial data from the trace of the solution of the wave equation on the lateral boundary of the cylinder. As mentioned in the introduction, we have two types of inversion results. The first, Theorem 1.4, is really a corollary of one of the inversion formulas for circular means from the previous section.

Proof of Theorem 1.4. Let $u(x, t)$ be the solution of the IVP (1.1), (1.2) in dimension two. Then by (1.11),

$$u(p, t) = \partial_t \int_0^t \frac{(r\,\mathcal{M} f)(p, r)}{\sqrt{t^2 - r^2}}\, dr$$

for $p \in S$. We can recover the circular means from u by the standard method of inverting an Abel-type equation. The details are not hard and may be found, for example, in [12]. The result is

$$(\mathcal{M} f)(p, r) = \frac{2}{\pi} \int_0^r \frac{u(p, t)}{\sqrt{r^2 - t^2}}\, dt. \tag{3.1}$$

Inserting (3.1) into the inversion formula (1.4) for $\mathcal{M}$ and applying Fubini's theorem gives, for $x \in D$,

$$\begin{aligned}
f(x) &= \frac{1}{2\pi R_0}\,\Delta \int_S \int_0^{2R_0} (r\,\mathcal{M}\,f)(p,r) \log\left|r^2 - |x-p|^2\right| dr\, ds(p) \\
&= \frac{1}{R_0\pi^2}\,\Delta \int_S \int_0^{2R_0} r \int_0^r \frac{u(p,t)}{\sqrt{r^2-t^2}} \log\left|r^2 - |x-p|^2\right| dt\, dr\, ds(p) \\
&= \frac{1}{R_0\pi^2}\,\Delta \int_S \int_0^{2R_0} u(p,t) \int_t^{2R_0} \frac{r}{\sqrt{r^2-t^2}} \log\left|r^2 - |x-p|^2\right| dr\, dt\, ds(p) \\
&= \frac{1}{R_0\pi^2}\,\Delta \int_S \int_0^{2R_0} u(p,t) K(t, |x-p|)\, dt\, ds(p).
\end{aligned}$$

Since $u(p,t) = (\mathcal{W}\,f)(p,t)$, this is (1.12), with

$$K(t,\bar{r}) := \int_t^{2R_0} \frac{r}{\sqrt{r^2-t^2}} \log\left|r^2 - \bar{r}^2\right| dr. \tag{3.2}$$

The integral in (3.2) can be evaluated exactly. For the sake of completeness, we give the analytic expression. If we substitute $r = \sqrt{t^2+\xi^2}$ into (3.2), then $dr = (\xi/r)\,d\xi$ and thus

$$\begin{aligned}
K(t,\bar{r}) &= \int_0^{\sqrt{4R_0^2-t^2}} \log\left|\xi^2 + (t^2 - \bar{r}^2)\right| d\xi \\
&= \sqrt{4R_0^2 - t^2}\left(-2 + \log|4R_0^2 - \bar{r}^2|\right) + \Gamma(t,\bar{r}),
\end{aligned}$$

where

$$\Gamma(t,\bar{r}) = \begin{cases} \sqrt{\bar{r}^2 - t^2}\log \frac{\sqrt{4R_0^2-t^2}+\sqrt{\bar{r}^2-t^2}}{\sqrt{4R_0^2-t^2}-\sqrt{\bar{r}^2-t^2}}, & t < \bar{r}, \\ 2\sqrt{t^2-\bar{r}^2}\arctan\sqrt{\frac{4R_0^2-t^2}{t^2-\bar{r}^2}}, & t > \bar{r}. \end{cases} \qquad \square$$

For the second type of inversion formula, we start by deriving a representation of the formal adjoint $\mathcal{P}^*$ for $n=2$. For any continuous function $G(p,t)$ on $S \times [0,\infty)$ that has a small amount of decay as $t \to \infty$, by Fubini's theorem, we have

$$\begin{aligned}
\langle \mathcal{P}\,f,\, G\rangle &= \int_S \int_0^\infty (\mathcal{P}\,f)(p,t)\, G(p,t)\, dt\, ds(p) \\
&= \frac{1}{2\pi}\int_S \int_0^\infty G(p,t)\left(\int_0^t \frac{r}{\sqrt{t^2-r^2}} \int_{S^1} f(p + r\omega)\, ds(\omega)\, dr\right) dt\, ds(p) \\
&= \frac{1}{2\pi}\int_S \int_0^\infty \left(\int_0^t \int_{S^1} \frac{f(p+r\omega)}{\sqrt{t^2-r^2}}\, r\, dr\, dS(\omega)\right) G(p,t)\, dt\, ds(p) \\
&= \frac{1}{2\pi}\int_S \int_0^\infty \left(\int_{\mathbf{R}^2} \frac{f(y)}{\sqrt{t^2-|y-p|^2}} \chi(\{|y-p|<t\})\, dy\right) G(p,t)\, dt\, ds(p) \\
&= \frac{1}{2\pi}\int_{\mathbf{R}^2} f(y) \left(\int_S \int_{|y-p|}^\infty \frac{G(p,t)}{\sqrt{t^2-|y-p|^2}}\, dt\, ds(p)\right) dy \\
&= \langle f,\, \mathcal{P}^*\, G\rangle,
\end{aligned}$$

where

$$\left(\mathcal{P}^*\, G\right)(y) := \frac{1}{2\pi}\int_S \int_{|y-p|}^\infty \frac{G(p,t)}{\sqrt{t^2-|y-p|^2}}\, dt\, ds(p). \tag{3.3}$$

The integral in (3.3) will be absolutely convergent for continuous G, provided that G has a small amount of decay as $t \to \infty$, for example, if $G(p,t) = \mathcal{O}(1/t^\alpha)$, as $t \to \infty$, for some $\alpha > 0$.

Next, we note a differentiation formula for the fractional integral appearing in (1.11).

PROPOSITION 3.1. *Let h be differentiable on $[0,\infty)$. Then, for $t > 0$,*

$$\partial_t \int_0^t \frac{r\,h(r)}{\sqrt{t^2-r^2}}\,dr = \frac{1}{t}\int_0^t \frac{r\,(\partial_r r h)(r)}{\sqrt{t^2-r^2}}\,dr. \tag{3.4}$$

Proof. Making the change of variable $r = t\xi$ in the integral on the left-hand side we have to evaluate

$$\partial_t \int_0^1 \frac{\xi}{\sqrt{1-\xi^2}}\, th(t\xi)\,d\xi.$$

Here differentiation under the integral yields $\int_0^1 \frac{\xi}{\sqrt{1-\xi^2}}\,(t\xi h'(t\xi) + h(t\xi))\,d\xi$, which is equal to the expression on the right-hand side after changing back to integration with respect to $r = t\xi$. □

Proof of (1.13) *in Theorem* 1.5 *for* $n = 2$. We compute $(\mathcal{P}^*\, t\partial_t^2\, \mathcal{P}\, f)(x)$ for smooth f supported in $\overline{B}$ and $x \in B$. The function $t\partial_t^2\,\mathcal{P}\, f$ has decay of order $1/t^2$ as $t \to \infty$ and so lies in the domain of $\mathcal{P}^*$. Using the definitions of $\mathcal{P}$ and $\mathcal{P}^*$ and relation (3.4),

$$\begin{aligned}
&(\mathcal{P}^*\, t\partial_t^2\, \mathcal{P}\, f)(x) \\
&\quad = \frac{1}{2\pi}\int_S \int_{|x-p|}^\infty \partial_t^2 \left(\int_0^t \frac{r\,(\mathcal{M}\, f)(p,r)}{\sqrt{t^2-r^2}}\,dr\right) \frac{t\,dt\,ds(p)}{\sqrt{t^2-|x-p|^2}} \\
&\quad = \frac{1}{2\pi}\int_S \int_{|x-p|}^\infty \partial_t \left(\frac{1}{t}\int_0^t \frac{r\,(\partial_r r\,\mathcal{M}\, f)(p,r)}{\sqrt{t^2-r^2}}\,dr\right) \frac{t\,dt\,ds(p)}{\sqrt{t^2-|x-p|^2}}.
\end{aligned}$$

Carrying out the differentiation in t using the chain rule, again using (3.4), and combining terms, the last integral can be rewritten as

$$\frac{1}{2\pi}\int_S \int_{|x-p|}^\infty \left(\int_0^r \frac{r\,(\partial_r r \partial_r r\,\mathcal{M}\, f)(p,r) - r\,(\partial_r r\,\mathcal{M}\, f)(p,r)}{t\sqrt{t^2-|x-p|^2}\sqrt{t^2-r^2}}\,dr\right) dt\,ds(p).$$

Using the identity

$$\partial_r r \partial_r r h - \partial_r r h = \partial_r r(\partial_r r h - h) = \partial_r r r \partial_r h = \partial_r r^2 \partial_r h$$

and applying Fubini's theorem $(\mathcal{P}^*\, t\partial_t^2\, \mathcal{P}\, f)(x)$ is in turn equal to

$$\frac{1}{2\pi}\int_S \int_0^\infty r\,(\partial_r r^2 \partial_r\,\mathcal{M}\, f)(p,r) \left(\int_{\max(|x-p|,r)}^\infty \frac{dt}{t\sqrt{t^2-|x-p|^2}\sqrt{t^2-r^2}}\right) dr\,ds(p).$$

The inner integral evaluates to

$$\frac{1}{2r|x-p|}\log\frac{r+|x-p|}{|r-|x-p||},$$

giving

$$(\mathcal{P}^*\, t\partial_t^2\, \mathcal{P}\, f)(x) = \frac{1}{4\pi}\int_S \left(\int_0^\infty (\partial_r r^2 \partial_r\,\mathcal{M}\, f)(p,r)\log\frac{r+|x-p|}{|r-|x-p||}\,dr\right)\frac{ds(p)}{|x-p|}. \tag{3.5}$$

Treating the inner integral in the principal value sense and integrating by parts, it is equal to the limit as $\varepsilon \to 0$ of boundary terms

$$\left[(r^2\partial_r \mathcal{M} f)(p,r)\log\frac{r+|x-p|}{|x-p|-r}\right]_0^{|x-p|-\varepsilon} + \left[(r^2\partial_r \mathcal{M} f)(p,r)\log\frac{r+|x-p|}{r-|x-p|}\right]_{|x-y|+\varepsilon}^{\infty}$$

plus the term

$$I_\epsilon := -\int_{\mathbf{R}^+\setminus[|x-p|-\varepsilon,|x-p|+\varepsilon]} (r\partial_r \mathcal{M} f)(p,r)\, r\partial_r \log\frac{r+|x-y|}{|r-|x-p||}\,dr.$$

Using that $\mathcal{M} f$ is smooth, flat at $r=0$, and of bounded support in $(0,\infty)$, the limit of the boundary terms is zero. Using the identity

$$r\partial_r \log\frac{r+|x-p|}{|r-|x-p||} = -|x-p|\partial_r \log|r^2-|x-p|^2|,$$

followed by another integration by parts, yields the sum of another pair of boundary terms and

$$I_\epsilon = -|x-p|\int_{R^+\setminus[|x-p|-\varepsilon,|x-p|+\varepsilon]} \left(\partial_r r\partial_r \mathcal{M} f\right)(p,r)\log\left|r^2-|x-p|^2\right|dr.$$

The boundary terms again evaluate to zero as $\varepsilon \to 0$, while the integral I_ϵ converges to

$$-|x-p|\int_0^\infty \left(\partial_r r\partial_r \mathcal{M} f\right)(p,r)\log\left|r^2-|x-p|^2\right|dr.$$

Inserting this into (3.5) and taking into account the support of $\mathcal{M} f$ gives

$$(\mathcal{P}^*\, t\partial_t^2\, \mathcal{P} f)(x) = -\frac{1}{4\pi}\int_S\int_0^{2R_0} \left(\partial_r r\partial_r \mathcal{M} f\right)(p,r)\log\left|r^2-|x-p|^2\right|dr\,ds(p).$$

In view of (1.5) of Theorem 1.1, (1.13) in Theorem 1.5 is proved for $n=2$. □

Proof of Theorem 1.6. Formula (1.15), for $n=2$, is an easy corollary of the result just established. Indeed, for f, g smooth with compact support in the closed disk of radius R_0, then

$$\langle f, g\rangle = -\frac{2}{R_0}\left\langle \mathcal{P}^*\, t\partial_t^2\, \mathcal{P} f,\, g\right\rangle = -\frac{2}{R_0}\left\langle t\partial_t^2\, \mathcal{P} f,\, \mathcal{P} g\right\rangle, \tag{3.6}$$

which is (1.15) for $n=2$, due to the definition of the operator $\mathcal{P}$.

In (1.15), the left-hand side is symmetric in f and g, while the right-hand side is not. Thus there is a companion identity, reversing the roles of u and v on the right-hand side. Taking the difference gives the equation

$$0 = \int_S\int_0^\infty t(u_{tt}v - uv_{tt})\,dt\,ds(p).$$

Integrating by parts (the boundary terms vanish) yields

$$0 = \int_S\int_0^\infty (u_t v - uv_t)\,dt\,ds(p),$$

and another integration by parts proves

$$0 = \int_S \int_0^\infty u_t v \, dt \, ds(p) = \int_S \int_0^\infty u v_t \, dt \, ds(p).$$

Using this and one integration by parts in (1.15) establishes (1.16), which completes the proof of Theorem 1.6 for $n = 2$. The extension to higher (even) dimensions follows almost word for word the proof from [5, section 4.2], where the trace identities in odd dimensions greater than three were proved from the three dimensional case. $\square$

Proof of Theorem 1.5 *for* $n > 2$. Reversing the chain of reasoning in (3.6) proves (1.13) in the L^2 sense from (1.15). Similarly, (1.14) follows from (1.16). However, as both sides are continuous functions when f is smooth, the formulas hold pointwise as well. $\square$

4. Numerical results. In the previous sections, we have established several exact inversion formulas to recover a function f supported in a closed disc $\overline{D}$ from either its spherical means $\mathcal{M} f$ or the trace $\mathcal{W} f$ of the solution of the wave equation with initial data $(f, 0)$. However, those formulas require continuous data, whereas in practical applications only a discrete data set is available. For example, in thermoacoustic tomography (see Figure 1.1) only a finite number of positions of the line detectors and a finite number of samples in time are feasible. In this section we derive discrete *filtered back-projection* (FBP) algorithms with linear interpolation in dimension two and present some numerical results.

The derived FBP algorithms are numerical implementations of discretized versions of (1.4)–(1.7) and (1.12)–(1.14), and the derivation of any of them follows the same line. We shall focus on the implementation of (1.5), assuming uniformly sampled discrete data

$$F^{k,m} := (\mathcal{M} f)(p^k, r^m), \qquad (k, m) \in \{0, \dots, N_\varphi\} \times \{0, \dots, N_r\}\,, \tag{4.1}$$

where $p^k := R_0 \, (\cos(kh_\varphi), \sin(kh_\varphi))$, $r^m := mh_r$, $h_\varphi := 2\pi/(N_\varphi + 1)$, and $h_r := 2R_0/N_r$. In order to motivate the derivation of a discrete FBP algorithm based on (1.5), we introduce the differential operator $\mathcal{D} := \partial_r r \partial_r$ and the integral operator

$$\begin{aligned} &\mathcal{I} : C_0^\infty(S \times [0, 2R_0)) \to C^\infty(S \times [0, 2R_0)), \\ &(\mathcal{I}\, G)(p, \bar{r}) := \int_0^{2R_0} G(p, r) \log |r^2 - \bar{r}^2| \, dr, \end{aligned} \tag{4.2}$$

which both act in the second component, and the so-called *back-projection operator*

$$\begin{aligned} \mathcal{B} : C^\infty(S \times [0, 2R_0)) &\to C^\infty(\overline{D}), \\ (\mathcal{B}\, G)(x) &:= \frac{1}{2\pi R_0} \int_S G(p, |x - p|) \, ds(p) \\ &= \frac{1}{2\pi} \int_0^{2\pi} G(p(\varphi), |x - p(\varphi)|) \, d\varphi, \end{aligned} \tag{4.3}$$

where $p(\varphi) := R_0(\cos\varphi, \sin\varphi)$. Therefore, we can rewrite (1.5) as

$$f = (\mathcal{B} \mathcal{I} \mathcal{D})(\mathcal{M} f). \tag{4.4}$$

In the numerical implementation the operators $\mathcal{B}$, $\mathcal{I}$, and $\mathcal{D}$ in (4.4) are replaced by finite dimensional approximations $\mathbf{B}$, $\mathbf{I}$, and $\mathbf{D}$ (as described below) and (4.4) is approximated by

$$f(x^i) \approx f^i := (\mathbf{B}\,\mathbf{I}\,\mathbf{D}\,\mathbf{F})^i, \quad i \in \{0, \dots, N\}^2\,. \tag{4.5}$$

Here $\mathbf{F} := (F^{k,m})_{k,m}$ with $F^{k,m}$ defined by (4.1), $x^i := -(R_0, R_0) + ih_x$ with $i = (i_1, i_2) \in \{0, N\}^2$, and $h_x := 2R_0/N$. In the following $\mathbf{S}_{\varphi,r}$ and $\mathbf{S}_x$ denote the sampling operators that map $G \in C^\infty(S \times [0, 2R_0])$ and $f \in C^\infty(\overline{D})$ onto their samples, $\mathbf{S}_{\varphi,r}\, G := (G(p^k, r^m))_{k,m}$ and $\mathbf{S}_x\, f := \mathbf{f} := (f(x^i))_i$, where we set $f(x^i) := 0$ if $x^i \notin \overline{D}$. Moreover, $|\cdot|_\infty$ denotes the maximum norm on either $\mathbf{R}^{(N_\varphi+1)\times(N_r+1)}$ or $\mathbf{R}^{(N+1)\times(N+1)}$.

1. The operator $\mathcal{D}$ can be written as $\partial_r + r\partial_r^2$. We approximate $\partial_r G$ with symmetric finite differences $\left(G^{k,m+1} - G^{k,m-1}\right)/(2h_r)$, $\partial_r^2 G$ by $\left(G^{k,m+1} + G^{k,m-1} - 2G^{k,m}\right)/h_r^2$ and the multiplication operator $G \mapsto rG$ by pointwise discrete multiplication $(G^{k,m})_{k,m} \mapsto (r^m G^{k,m})_{k,m}$. This leads to the discrete approximation

   (4.6)
   $$\mathbf{D} : \mathbf{R}^{(N_\varphi+1)\times(N_r+1)} \to \mathbf{R}^{(N_\varphi+1)\times(N_r+1)} : \qquad \mathbf{G} \mapsto \left((\mathbf{D}\,\mathbf{G})^{k,m}\right)_{k,m},$$
   $$(\mathbf{D}\,\mathbf{G})^{k,m} := \frac{1}{h_r}\left(\left(m + \frac{1}{2}\right) G^{k,m+1} + \left(m - \frac{1}{2}\right) G^{k,m-1} - 2mG^{k,m}\right),$$

   where we set $G^{k,-1} := G^{k,N_r+1} := 0$. The approximation of ∂_r with symmetric finite differences is of second order, and therefore $|(\mathbf{S}_{\varphi,r}\,\mathcal{D} - \mathbf{D}\,\mathbf{S}_{\varphi,r})G|_\infty \le C_1 h_r^2$ for some constant C_1, which does not depend on h_r.
2. Next, we define a second order approximation to the integral operator $\mathcal{I}$. This is done by replacing $G(p^k, \cdot)$ in (4.2) by the piecewise linear spline $T^k[G] : [0, 2R_0] \to \mathbf{R}$ interpolating G at the nodes r^m. More precisely,

   $$\mathbf{I} : \mathbf{R}^{(N_\varphi+1)\times(N_r+1)} \to \mathbf{R}^{(N_\varphi+1)\times(N_r+1)} : \quad \mathbf{G} \mapsto \left((\mathbf{I}\,\mathbf{G})^{k,m}\right)_{k,m}$$

   is defined by

   $$(4.7) \qquad T^k[\mathbf{G}](r) := G^{k,m} + \frac{r - r^m}{h_r}(G^{k,m+1} - G^{k,m}), \quad r \in [r^m, r^{m+1}],$$

   and

   (4.8)
   $$\begin{aligned}(\mathbf{I}\,\mathbf{G})^{k,m} &:= \int_0^{2R_0} T^k[\mathbf{G}](r) \log|r^2 - (r^m)^2|\, dr \\ &= \sum_{m'=0}^{N_r-1} G^{k,m'} \left(\int_{r^{m'}}^{r^{m'+1}} \log|r^2 - (r^m)^2|\, dr\right) \\ &\quad + \sum_{m'=0}^{N_r-1} \frac{G^{k,m'+1} - G^{k,m'}}{h_r} \left(\int_{r^{m'}}^{r^{m'+1}} (r - r^{m'}) \log|r^2 - (r^m)^2|\, dr\right).\end{aligned}$$

   For an efficient and accurate numerical implementation it is crucial that the integrals in (4.8) are evaluated analytically. In fact, by straightforward computation it can be verified that

   $$(\mathbf{I}\,\mathbf{G})^{k,m} = \sum_{m'=0}^{N_r-1} a_{m'}^m G^{k,m'} + \frac{1}{h_r}\sum_{m'=0}^{N_r-1} b_{m'}^m \left(G^{k,m'+1} - G^{k,m'}\right),$$

   (4.9)
   $$a_{m'}^m := \Big[(r - r^m)\log|r - r^m| + (r + r^m)\log|r + r^m| - 2r\Big]_{r=r^{m'}}^{r^{m'+1}},$$
   $$b_{m'}^m := -r^{m'} a_{m'}^m + \frac{1}{2}\Big[(r^2 - (r^m)^2)\log|r^2 - (r^m)^2| - r^2\Big]_{r=r^{m'}}^{r^{m'+1}}.$$

Moreover, using the fact that piecewise linear interpolation is of second order [15] and that $r \mapsto \log|r^2 - (r^m)^2|$ is integrable, it can be readily verified that the approximation error satisfies $|(\mathbf{S}_{\varphi,r}\,\mathcal{I} - \mathbf{I}\,\mathbf{S}_{\varphi,r})G|_\infty \le C_2 h_r^2$ with some constant C_2 independent of h_r.

3. Finally, we define a second order approximation to the back-projection (4.3). The discrete back-projection operator $\mathbf{B} : \mathbf{R}^{(N_\varphi+1)\times(N_r+1)} \to \mathbf{R}^{(N+1)\times(N+1)}$ is obtained by approximating (4.3) with the trapezoidal rule and piecewise linear interpolation (4.7) in the second variable,

$$(\mathbf{B}\,\mathbf{G})^i := \frac{1}{N_\varphi + 1} \sum_{k=0}^{N_\varphi} T^k[\mathbf{G}](|x^i - p^k|), \quad x^i \in D, \tag{4.10}$$

and setting $(\mathbf{B}\,\mathbf{G})^i := 0$ for $x^i \notin D$. It is well known [15] that both linear interpolation in r and the trapezoidal rule in φ are second order approximations and therefore $|(\mathbf{S}_x\,\mathcal{B} - \mathbf{B}\,\mathbf{S}_{\varphi,r})G|_\infty \le C_3 \max\left\{h_r^2, h_\varphi^2\right\}$ for some constant C_3.

The discrete FBP algorithm is given by (4.5) with $\mathbf{D}$, $\mathbf{I}$, $\mathbf{B}$ defined in (4.6), (4.9), (4.10) and is summarized in Algorithm 1. Using $f(x^i) = (\mathbf{S}_x\,\mathcal{B}\mathcal{I}\mathcal{D}\,F)^i = (\mathbf{S}_x\,f)^i$ and $f^i = (\mathbf{B}\,\mathbf{I}\,\mathbf{D}\,\mathbf{S}_{\varphi,r}\,F)^i$, the discretization error $|f(x^i) - f^i|$ can be estimated as

$$\begin{aligned} |(\mathbf{S}_x\,\mathcal{B}\mathcal{I}\mathcal{D} - \mathbf{B}\,\mathbf{I}\,\mathbf{D}\,\mathbf{S}_{\varphi,r})F|_\infty &\le |(\mathbf{S}_x\,\mathcal{B} - \mathbf{B}\,\mathbf{S}_{\varphi,r})(\mathcal{I}\mathcal{D}\,F)|_\infty \\ &\quad + |\mathbf{B}(\mathbf{S}_{\varphi,r}\,\mathcal{I} - \mathbf{I}\,\mathbf{S}_{\varphi,r})(\mathcal{D}\,F)|_\infty \\ &\quad + |\mathbf{B}\,\mathbf{I}(\mathbf{S}_{\varphi,r}\,\mathcal{D} - \mathbf{D}\,\mathbf{S}_{\varphi,r})(F)|_\infty. \end{aligned} \tag{4.11}$$

Using the facts that $\mathbf{B}$ and $\mathbf{I}$ are bounded by some constant independent of h_r and that the approximations of $\mathcal{D}$, $\mathcal{I}$, $\mathcal{B}$ with $\mathbf{D}$, $\mathbf{I}$, $\mathbf{B}$ are of second order implies that

$$|\mathbf{S}_x\,f - \mathbf{B}\,\mathbf{I}\,\mathbf{D}\,\mathbf{F}|_\infty \le C \max\left\{h_r^2, h_\varphi^2\right\} \tag{4.12}$$

for some constant C independent of h_r, h_φ. This shows that the derived FBP algorithm has second order accuracy (for exact data).

In the numerical implementation, the coefficients in (4.9) are precomputed and stored. Therefore the numerical effort of evaluating (4.9) is $\mathcal{O}(N_r^2 N_\varphi)$. Moreover, (4.6) requires $\mathcal{O}(N_r N_\varphi)$ operations and the discrete FBP $\mathcal{O}(N^2 N_\varphi)$, since for all $(N+1)^2$ reconstruction points x^i we have to sum over $N_\varphi + 1$ center locations on S. Hence, assuming $N \sim N_r$ and $N \sim N_\varphi$, Algorithm 1 requires $\mathcal{O}(N^3)$ operations and therefore has the same numerical effort as the classical FBP algorithm used in x-ray CT [12]. Analogous to the procedure described above, discrete FBP algorithms were derived using (1.4), (1.6) for inverting $\mathcal{M}$ and (1.12) for inverting $\mathcal{W}$.

In the following we present numerical results of our FBP algorithms for reconstruction of the phantom shown in the left picture in Figure 4.1, consisting of a superposition of characteristic functions and one Gaussian kernel. We calculated the data $\mathcal{M}\,f$ via numerical integration and the operator $\mathcal{W}\,f = \partial_t\,\mathcal{P}\,f$ using (1.11). Subsequently we added 5% uniformly distributed noise to $\mathcal{M}\,f$ and 10% uniformly distributed noise to $\mathcal{W}\,f$. The results for $N = N_\varphi = N_r = 300$ using the algorithms based on (1.4), (1.5), (1.6), and (1.12) are depicted in Figures 4.2, 4.3, and 4.4. All implementations show good results, although no explicit regularization strategy is incorporated in order to regularize the involved (mildly) ill-posed numerical differentiation. In particular, (1.6) and (1.12) appear to be most insensitive to noise. However, for noisy data, the accuracy of FBP algorithms can be further improved by incorporating a regularizing strategy similar to that used in [8]. The derived identities in this article provide the

Algorithm 1 Discrete FBP algorithm with linear interpolation for reconstruction of $\mathbf{f}$ using data $\mathbf{F}$.

```
$h_\varphi \leftarrow 2\pi/(N_\varphi + 1)$
$h_r \leftarrow 2R_0/N_r$                                              ▷ initialization
**for** $m, m' = 0, \dots, N_r$ **do**                                 ▷ precompute kernel
    Calculate $a^m_{m'}$, $b^m_{m'}$ according to (4.8)
**end for**

**for** $k = 0, \dots, N_\varphi$ **do**                               ▷ filtering
    **for** $m = 0, \dots, N_r$ **do**
        $F^{k,m} \leftarrow \big(m + 1/2\big)F^{k,m+1} + \big(m - 1/2\big)F^{k,m-1} - 2mF^{k,m}$     ▷ (4.6)
    **end for**
    **for** $m = 0, \dots, N_r$ **do**
        $F^{k,m} \leftarrow \sum_{m'=0}^{N_r-1} a^m_{m'} F^{k,m'} + \sum_{m'=0}^{N_r-1} b^m_{m'} \Big(F^{k,m'+1} - F^{k,m'}\Big)/h_r$     ▷ (4.9)
    **end for**
**end for**

**for** $i_1, i_2 = 0, \dots, N$ **do**                                ▷ BP with linear interpolation
    $i \leftarrow (i_1, i_2)$
    $f^i \leftarrow 0$
    **for** $k = 0, \dots, N_\varphi$ **do**
        Find $m \in \{0, \dots, N_r - 1\}$ with $r^m \leq |p^k - x^i| < r^{m+1}$
        $T \leftarrow F^{k,m} + (r - r^m)(F^{k,m+1} - F^{k,m})/h_r$     ▷ interpolation (4.7)
        $f^i \leftarrow f^i + T/(N_\varphi + 1)$                        ▷ discrete back-projection (4.10)
    **end for**
**end for**
```

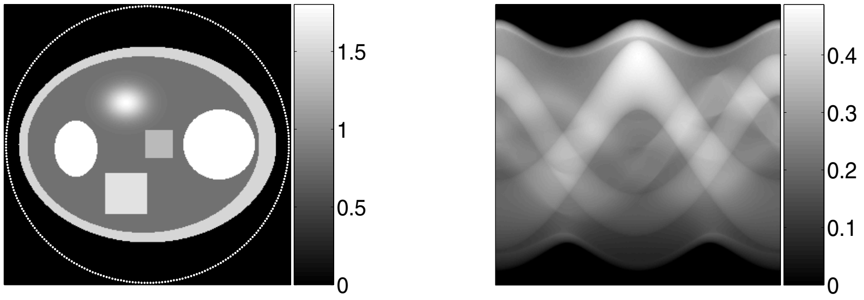


Fig. 4.1. *Imaging phantom and data. Left: Imaging phantom f consisting of several characteristic functions and one Gaussian kernel. Right: Simulated data $F = \mathcal{M} f$.*

mathematical foundation for further development of FBP algorithms for the inversion from spherical means and the inversion of the wave equation.

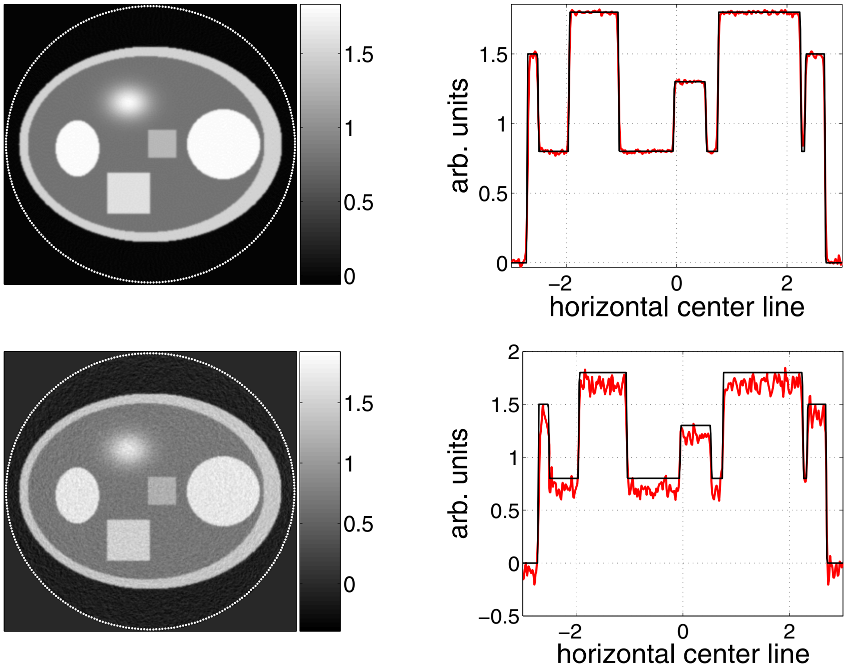


FIG. 4.2. *Numerical reconstruction with Algorithm* 1. *Top: Reconstructions from simulated data. Bottom: Reconstructions from simulated data after adding* 5% *uniformly distributed noise.*

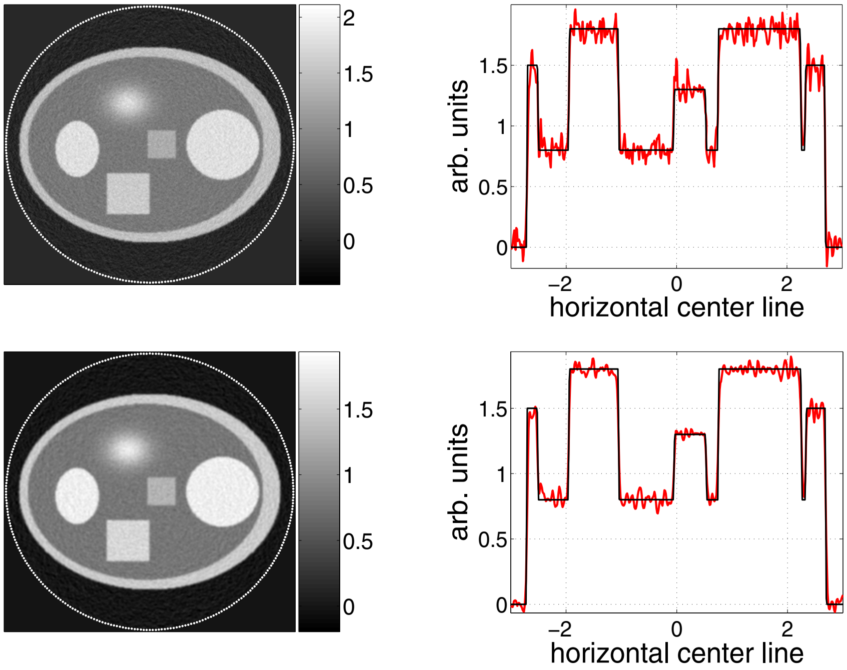


FIG. 4.3. *Numerical reconstruction from spherical means with* 5% *noise added. Top: Reconstruction using* (1.4). *Bottom: Reconstruction using* (1.6).

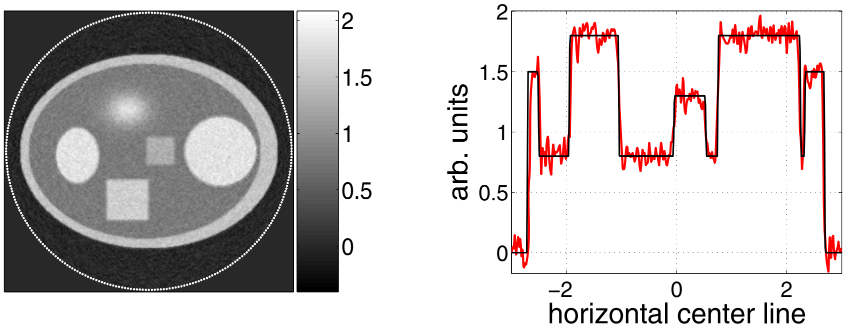


FIG. 4.4. *Numerical reconstruction using* (1.12) *from trace* $\mathcal{W}f$ *of the solution of the wave equation with* 10% *noise added.*

REFERENCES

[1] M. AGRANOVSKY, P. KUCHMENT, AND E. T. QUINTO, *Range descriptions for the spherical mean Radon transform*, J. Funct. Anal., 248 (2007), pp. 344–386.

[2] G. AMBARTSOUMIAN AND P. KUCHMENT, *A range description for the planar circular Radon transform*, SIAM J. Math. Anal., 38 (2006), pp. 681–692.

[3] P. BURGHOLZER, C. HOFER, G. PALTAUF, M. HALTMEIER, AND O. SCHERZER, *Thermoacoustic tomography with integrating area and line detectors*, IEEE Trans. Ultrason. Ferroelec. Freq. Contr., 52 (2005), pp. 1577–1583.

[4] R. COURANT AND D. HILBERT, *Methoden der Mathematischen Physik*, 4th ed., Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[5] D. FINCH, S. K. PATCH, AND RAKESH, *Determining a function from its mean values over a family of spheres*, SIAM J. Math. Anal., 35 (2004), pp. 1213–1240.

[6] D. FINCH AND RAKESH, *The range of the spherical mean value operator for functions supported in a ball*, Inverse Problems, 22 (2006), pp. 923–938.

[7] M. HALTMEIER AND T. FIDLER, *Mathematical Challenges Arising in Thermoacoustic Computed Tomography with Line Detectors*, http://arxiv.org/abs/math/0610155 (2006).

[8] M. HALTMEIER, T. SCHUSTER, AND O. SCHERZER, *Filtered backprojection for thermoacoustic computed tomography in spherical geometry*, Math. Methods Appl. Sci., 28 (2005), pp. 1919–1937.

[9] R. A. KRUGER, J. L. KISER, D. R. REINECKE, G. A. KRUGER, AND K. D. MILLER, *Thermoacoustic optical molecular imaging of small animals*, Molecular Imaging, 2 (2003), pp. 113–123.

[10] R. A. KRUGER, K. D. MILLER, H. E. REYNOLDS, W. L. KISER, D. R. REINECKE, AND G. A. KRUGER, *Breast cancer in vivo: Contrast enhancement with thermoacoustic CT at* 434 *MHz-feasibility study*, Radiology, 216 (2000), pp. 279–283.

[11] L. KUNYANSKY, *Explicit inversion formulas for the spherical mean transform*, Inverse Problems, 23 (2007), pp. 373–383.

[12] F. NATTERER, *The Mathematics of Computerized Tomography*, Wiley, Chichester, UK, 1986.

[13] S. J. NORTON, *Reconstruction of a two-dimensional reflecting medium over a circular domain: Exact solution*, J. Acoust. Soc. Amer., 67 (1980), pp. 1266–1273.

[14] G. PALTAUF, R. NUSTER, M. HALTMEIER, AND P. BURGHOLZER, *Thermoacoustic computed tomography using a Mach-Zehnder interferometer as acoustic line detector*, Appl. Opt., 46 (2007), pp. 3352–3358.

[15] A. QUARTERONI, R. SACCO, AND F. SALERI, *Numerical Mathematics*, Springer-Verlag, New York, 2000.

[16] X. D. WANG, G. PANG, Y. J. KU, X. Y. XIE, G. STOICA, AND L.-H. V. WANG, *Noninvasive laser-induced photoacoustic tomography for structural and functional in vivo imaging of the brain*, Nat. Biotechnol., 21 (2003), pp. 803–806.

[17] M. XU AND L.-H. V. WANG, *Photoacoustic imaging in biomedicine*, Rev. Sci. Instrum., 77 (2006), 0411011.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 413–436

A HYBRID LAGRANGIAN MODEL BASED ON THE AW–RASCLE TRAFFIC FLOW MODEL[*]

S. MOUTARI[†] AND M. RASCLE[†]

Abstract. In this paper, we propose a simple fully discrete hybrid model for vehicular traffic flow, for which both the macroscopic and the microscopic models are based on a Lagrangian discretization of the Aw–Rascle (AR) model [A. Aw and M. Rascle, *SIAM J. Appl. Math.*, 60 (2000), pp. 916–938]. This hybridization makes use of the relation between the AR macroscopic model and a *follow-the-leader*-type model [D. C. Gazis, R. Herman, and R. W. Rothery, *Oper. Res.*, 9 (1961), pp. 545–567; R. Herman and I. Prigogine, *Kinetic Theory of Vehicular Traffic*, American Elsevier, New York, 1971], established in [A. Aw, A. Klar, M. Materne, and M. Rascle, *SIAM J. Appl. Math.*, 63 (2002), pp. 259–278]. Moreover, in the hybrid model, the total variation in space of the velocity v is nonincreasing, the total variation in space of the specific volume τ is bounded, and the total variations in time of v and τ are bounded. Finally, we present some numerical simulations which confirm that the models' synchronization processes do not affect the waves propagation.

Key words. traffic flow, hybrid model, Lagrangian discretization, macroscopic model, microscopic model, total variation

AMS subject classifications. 35L, 35L65

DOI. 10.1137/060678415

1. Introduction.

Most of the vehicular traffic models are either macroscopic [3, 12, 18, 25, 30, 33, 34, 36, 37] or microscopic [15, 22]. When following a macroscopic approach, one focuses on global parameters such as traffic density or traffic flow. In general, from a macroscopic perspective vehicular traffic is viewed as a compressible fluid flow, whereas a microscopic approach describes the behavior of each individual vehicle. Macroscopic models allow one to simulate traffic on large networks but with a poor description of the details. On the other hand, microscopic models can cover such details, but they are intractable on a large network.

However, a typical road transport system or a road network includes obstacles, different road geometries and configurations (intersections, roundabouts, multiple lanes, etc.), as well as control features, such as traffic lights and crossings, which have a nonnegligible impact on traffic in the whole network. Therefore, neither of the two approaches is separately able to capture real traffic dynamics. A natural strategy is therefore to combine macroscopic and microscopic models, depending on the number of details that we need. This *hybrid* approach has recently received a considerable interest in traffic modeling [1, 6, 8, 19, 21, 27, 31]. Indeed, such models enable one to take into account the most important details of the traffic but still allow for descriptions of the traffic on a large network. However, they require strong consistency and compatibility between macroscopic and microscopic models to be coupled [20, 31].

Here, our macroscopic description is based on the Aw–Rascle (AR) model [3], whereas the microscopic model is a *follow-the-leader* (FLM)-type model [15, 22].

[*] Received by the editors December 22, 2006; accepted for publication (in revised form) July 30, 2007; published electronically November 16, 2007. This work was partially supported by the former EU Network HYKE HPRN-CT-2002-00282 and by the French ACI-NIM (Nouvelles Interactions des Mathématiques) 193 (2004).

http://www.siam.org/journals/siap/68-2/67841.html

[†] Department de Mathématiques, Laboratoire J. A. Dieudonné, UMR CNRS 6621, Université de Nice-Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France (salissou@math.unice.fr, rascle@math.unice.fr).

In [2], Aw et al. established a connection between the two classes of models. More precisely, the macroscopic model can be viewed as the limit of the time discretization of a microscopic FLM-type model when the number of vehicles increases. This can be done via a (hyperbolic) scaling in space and time (zoom) for which the density and the velocity are invariant.

Our aim in the current work is to propose a simple and fully discrete hybrid model for which both the macroscopic and the microscopic parts are based on the Lagrangian discretization of the AR second order model of traffic flow.

The outline of this paper is as follows: Sections 2 and 3 provide, respectively, some details on the discretizations of the macroscopic and the microscopic models. Section 4 describes the relations between the two models. Section 5 is devoted to the presentation of the hybrid model. In section 6, we establish estimates on the total variation both in space and time for the velocity v and the specific volume τ in Lagrangian coordinates. These estimates are of course the main ingredient for studying the convergence of our hybrid scheme to a suitable initial boundary value problem, which we will investigate in a forthcoming work. Finally, in section 7, some numerical simulations of the hybrid model confirm that this micro-macro description allows for a very nice description in *both* regimes.

2. The AR macroscopic model. We are concerned with the AR macroscopic model of traffic flow. It consists of the conservative form (in Eulerian coordinates) of the two following equations:

$$\begin{cases}\partial_t\rho+\partial_x(\rho v)=0,\\ \partial_t(\rho w)+\partial_x(\rho v w)=0,\end{cases} \tag{2.1}$$

where ρ denotes the fraction of space occupied by cars (a dimensionless local density), v is the macroscopic velocity of cars, and, for instance, $w=v+p(\rho)$. Many other choices could be considered as well. In what follows, we will assume for concreteness that

$$p(\rho)=\begin{cases}\frac{v_{ref}}{\gamma}\left(\frac{\rho}{\rho_m}\right)^{\gamma}, & \gamma>0,\\ -v_{ref}ln\left(\frac{\rho}{\rho_m}\right), & \gamma=0,\end{cases} \tag{2.2}$$

with v_{ref} a given reference velocity and $\rho_m:=\rho_{max}=1$ the maximal density.

Let $\tau=1/\rho$ be the specific volume and denote by (X,T) the Lagrangian "mass" coordinates. We have

$$\partial_x X=\rho,\qquad \partial_t X=-\rho v,\qquad T=t.$$

We recall that ρ is dimensionless; thus $X=\int^x\rho(y,t)dy$ describes the total length occupied by cars up to the point x if they were packed "nose to tail."

The system (2.1) can be rewritten in Lagrangian "mass" coordinates (X,T) as

$$\begin{cases}\partial_T\tau-\partial_X v=0,\\ \partial_T w=0,\end{cases} \tag{2.3}$$

now with $w=v+P(\tau):=v+p\left(\frac{1}{\tau}\right)$ (we set $\tau_m:=\tau_{min}:=\frac{1}{\rho_m}:=\frac{1}{\rho_{max}}=1$).

Away from the vacuum, the system (2.3) is strictly hyperbolic and is equivalent to the system (2.1). Its eigenvalues are

$$\lambda_1=P'(\tau)<0\quad\text{and}\quad\lambda_2=0.$$

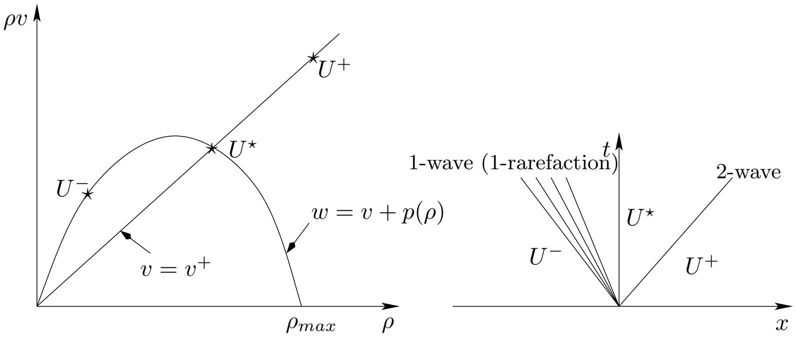


FIG. 2.1. *Riemann problem in $(\rho, \rho v)$ and (x, t) planes.*

Moreover, λ_1 is genuinely nonlinear and λ_2 is linearly degenerate. The Riemann invariants associated with the two eigenvalues λ_1 and λ_2 are v and w, respectively.

2.1. The Riemann solver. Let us consider the following Riemann problem:

$$\begin{cases} \partial_t \tau - \partial_X v = 0, \\ \partial_t w = 0, \end{cases} \tag{2.4}$$

with the initial data

$$\begin{cases} U^+(X,0) = (\tau^+, w^+) & \text{if } X > 0, \\ U^-(X,0) = (\tau^-, w^-) & \text{if } X < 0. \end{cases} \tag{2.5}$$

The natural solution $U(X, t)$ to the Riemann problem (2.4)–(2.5) involves two waves: a rarefaction or a shock wave associated with the first characteristic field λ_1 followed by a contact discontinuity associated with the second one λ_2.

PROPOSITION 2.1. *The solution of the Riemann problem* (2.4)–(2.5) *is constructed as follows. First, we connect U^- with an intermediate state $U^* = (\tau^*, w^*)$ (such that $v^* = w^* - P(\tau^*) = v^+$ and $w^* = w^-$) by a 1-shock wave (if $v^+ < v^-$) or a 1-rarefaction (if $v^+ > v^-$). Then, U^* is connected with U^+ by a 2-contact discontinuity (see Figure* 2.1*).*

Through each wave, the specific volume τ and the velocity v are monotonous functions of X/t. Therefore, away from the vacuum, the solution $U(X, t)$ remains in the bounded invariant region $\mathcal{R}$ defined in (2.6)*, i.e.,*

$$U(X,t) = (\tau, w), \quad \textit{where } \tau = P^{-1}(w - v)$$

and

$$(v, w) \in \mathcal{R} = \{[v^{min}, v^{max}] \times [w^{min}, w^{max}]\} \cap \{w \geq v\}, \tag{2.6}$$

where $v^{min}, w^{min} \geq 0$ and $v^{max}, w^{max} < +\infty$ (see Figure 2.2*).*

In the (w, v) coordinates (see Figure 2.2) we have

$$U^{\pm} = (w^{\pm}, v^{\pm}) \quad \text{and} \quad U^* = (w^-, v^+). \tag{2.7}$$

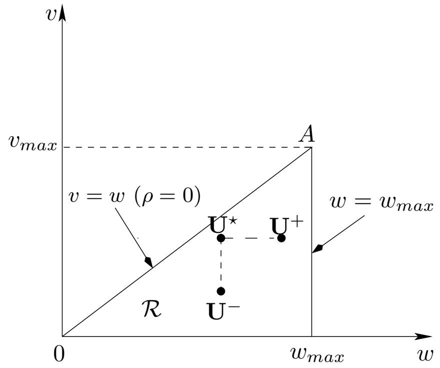


FIG. 2.2. *Riemann problem. The above triangle* $(0, A, w_{max})$ *is an invariant region in the* (w, v) *plane.*

2.2. Lagrangian discretization of the AR macroscopic model. Many approximate methods for (2.3) are based on solutions to the Riemann problem. Here, we are particularly interested in the Godunov scheme. In order to define the Godunov scheme associated with the above Riemann solver, we introduce grid points in space $X_j := j\Delta X$, $j \in \mathbb{Z}$, and in time $t_n = n\Delta t$, $n \in \mathbb{N}$. Let $h := (\Delta X, \Delta t)$ tend to $(0, 0)$, with $r := \frac{\Delta t}{\Delta X} =$ *constant*, and assume that for all $(\Delta X, \Delta t)$ the CFL condition is satisfied:

$$r \sup_{U \in \mathcal{R}} \left(\max_{i=1,2} \{|\lambda_i(U)|\} \right) \leq 1, \tag{2.8}$$

where $\mathcal{R}$ is the invariant region defined in (2.6), containing the initial data $U(x)$ for all $x \in \mathbb{R}$. Moreover, we assume that $\mathcal{R}$ does not touch the vacuum, i.e.,

$$\inf\{w - v,\ (v, w) \in \mathcal{R}\} > 0.$$

Then, the Lagrangian Godunov discretization of the AR macroscopic model (2.3) (see [2] for more details) is given by

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta X} \left(v_{j+1}^n - v_j^n \right), \\ w_j^{n+1} = w_j^n, \end{cases} \tag{2.9}$$

with initial data

$$\begin{cases} \tau_j(0) = \tau_j^0 \geq \frac{1}{\rho_m} = \tau_m = 1, \\ 0 \leq v_j(0) = v_j^0 \leq w_j - P(\tau_m). \end{cases} \tag{2.10}$$

PROPOSITION 2.2. *Starting with arbitrary initial data for the Godunov scheme*

$$U_h(X, 0) = (\tau_h^0, w_h^0),$$

with $(v_h^0 = w_h^0 - P(\tau_h^0), w_h^0) \in \mathcal{R}$ *(defined in* (2.6)*), the solutions*

$$U_h(X, t_n) = (\tau_h^n, w_h^n)$$

constructed by the Godunov scheme satisfy

$$(v_h^n = w_h^n - P(\tau_h^n), w_h^n) \in \mathcal{R}.$$

Therefore, the region $\mathcal{R}$ is also invariant for the Godunov scheme.

Proof. In each cell $I_j = [X_{j-1/2}, X_{j+1/2}]$, the variable w is constant, whereas the velocity and the specific volume τ are monotonous, by Proposition 2.1. Therefore, in each cell, $U(X,t) = (\tau, w)(X,t)$ remains in the same bounded region $\mathcal{R}$, which is thus invariant for the Godunov scheme. □

We are going to use the following results; see [4, Theorem 3.1] and [2, Theorem 1].

PROPOSITION 2.3 (see [4]). *Assume that the sequence v_h^0 is in $BV(\mathbb{R})$; that is, the total variation in space is bounded: there exists a constant $C_0 < +\infty$ such that*

$$TV_X(v_h^0; \mathbb{R}) = \sum_{j\in\mathbb{Z}} \left| v_{j+1}^0 - v_j^0 \right| \leq C_0.$$

Let

$$\tilde{v}_h(X,t) := v_j^n + (t - t_n)(v_j^{n+1} - v_j^n)/\Delta t \tag{2.11}$$

be the linear interpolation in time of v_h on I_j between t_n and t_{n+1} and similarly for $\tilde{\tau}_h$.

Then, for all $n \in \mathbb{N}$, for all $h = (\Delta X, \Delta t)$, we have the following:

(i) *The total variation in X of $v_h(.,t)$ is nonincreasing in time, and the total variation in t of $\tilde{v}_h(.,.)$ is bounded on $\mathbb{R} \times [0,T]$:*

$$\begin{aligned} \sup_{t\geq 0} TV_X(v_h(.,t) &= \sup_{n\in\mathbb{N}} TV_X(v_h^n; \mathbb{R}) \\ &= \sup_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} \left| v_{j+1}^n - v_j^n \right| \leq TV_X(v_h^0, \mathbb{R}) \leq C_0, \end{aligned} \tag{2.12}$$

$$TV_t(\tilde{v}_h(.,.); \mathbb{R} \times [t,t']) \leq C \max(|t'-t|, \Delta t) C_0. \tag{2.13}$$

(ii) *The total variation in X of $\tau_h(.,t)$ on $\cup_{j\in\mathbb{Z}} I_j$ and the total variation in t of $\tilde{\tau}_h(.,.)$ on $\mathbb{R} \times [0,\infty]$ are bounded uniformly in h; i.e., there exists a constant C' independent of h such that*

$$\sup_h \sup_{t\geq 0} \sum_{j\in\mathbb{Z}} TV_X(\tau_h(.,t); I_j) \leq C' C_0, \tag{2.14a}$$

$$\forall t \in [0,t'],\ \sup_h TV_t(\tilde{\tau}_h(.,.); \mathbb{R} \times [t,t']) \leq C' \max(|t'-t|, \Delta t) C_0. \tag{2.14b}$$

We recall that I_j is the open interval $\left(X_{j-1/2}, X_{j+1/2}\right)$. Moreover, if we assume that w_h and τ_h are initially in $BV(\mathbb{R})$, then (2.14a) *can be replaced by the stronger result*

$$\sup_h \sup_{t\geq 0} TV_X(\tau_h(.,t); \mathbb{R}) \leq C' C_0. \tag{2.15}$$

3. The microscopic (FLM) model.

The microscopic model that we consider is an FLM-type model [15, 22]. In such a model, the basic idea is that the acceleration at time t depends on the relative speeds of the vehicle and its leading vehicle at time t as well as the distance between the vehicles. Therefore, the dynamics of a vehicle j is given by the two equations

$$\begin{cases} \frac{dx_j}{dt} = v_j, \\ \frac{dv_j}{dt} = P'\left(\frac{x_{j+1}-x_j}{\Delta X}\right)\left(\frac{v_{j+1}-v_j}{\Delta X}\right), \end{cases} \tag{3.1}$$

where $x_j(t)$ and $v_j(t)$ are, respectively, the position and the velocity of the vehicle j at time t, and ΔX is its length. Here, $\rho_j := \frac{\Delta X}{(x_{j+1}-x_j)}$ is the normalized local density (more precisely the fraction of space occupied by a vehicle) of this vehicle. Therefore, the specific volume is $\tau_j = \frac{1}{\rho_j} = \frac{(x_{j+1}-x_j)}{\Delta X}$ and the maximal density $\rho_m = \rho_{max} = \frac{1}{\tau_m} = 1 := \frac{1}{\tau_{min}}$. In this section, contrarily to sections 5 and 6, the car $(j+1)$ is the leader of car j. A prototype is the case where $P(\tau) = p\left(\frac{1}{\tau}\right) = \frac{v_{ref}}{\gamma}\left(\frac{\rho}{\rho_m}\right)^\gamma$ (with v_{ref} a given reference velocity and $\gamma > 0$ a given parameter) and $w_j = v_j + P(\tau_j)$. Then system (3.1) writes

$$\begin{cases} \frac{d\tau_j}{dt} = \frac{(v_{j+1}-v_j)}{\Delta X}, \\ \frac{dw_j}{dt} = 0. \end{cases} \tag{3.2}$$

The explicit first order Euler time discretization (with step Δt of system (3.2) is then

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta X}\left(v_{j+1}^n - v_j^n\right), \\ w_j^{n+1} = w_j^n, \end{cases} \tag{3.3}$$

with

$$v_j^{n+1} = w_j^{n+1} - P(\tau_j^{n+1})$$

and initial conditions

$$\begin{cases} \tau_j(0) = \tau_j^0 \geq \frac{1}{\rho_m} = \tau_m = 1, \\ 0 \leq v_j(0) = v_j^0 \leq w_j - P(\tau_m). \end{cases} \tag{3.4}$$

As in the macroscopic scheme, we can define $(\tilde{\tau}_h, \tilde{v}_h)$ by (2.11). Since system (3.3) is *exactly* the same as (2.9), $\tilde{\tau}_h$ and $\tilde{v}_h$ satisfy the same BV estimates as in Proposition 2.3.

4. Link between the macroscopic and microscopic model: The scaling. We now consider a large number of vehicles on a long stretch of road. Let us now introduce in the AR macroscopic model (2.3) a scaling (zoom) such that the size of the considered domain and the number of vehicles tend to infinity, whereas the vehicle length tends to 0. Let ϵ be the scaling parameter. For some given Eulerian or Lagrangian coordinates (x,t) or (X,t), we consider the rescaled coordinates

$$(x', t') = (\epsilon x, \epsilon t); \qquad (X', t') = (\epsilon X, \epsilon t).$$

Consequently, the length of a vehicle is now $\Delta X' = \epsilon \Delta X$. The parameter ϵ is proportional to the inverse of maximal possible number of vehicles per new unit length. However, in the new coordinates (X', t'), the variable τ (resp., ρ) and the Riemann invariant (v, w) remain unchanged, that is,

$$\tau' = \tau \text{ (resp., } \rho' = \rho), \quad v' = v, \quad w' = w.$$

Thus the system (2.3) becomes

$$\begin{cases} \frac{\partial \tau}{\partial t} = \frac{\partial v}{\partial X'}, \\ \frac{\partial w}{\partial t'} = 0. \end{cases} \tag{4.1}$$

Using the same scaling for the microscopic model, (3.2) now writes

$$\begin{cases} \frac{d\tau_j}{dt'} = \frac{1}{\Delta X'}\left(v_{j+1} - v_j\right), \\ \frac{dw_j}{dt'} = 0. \end{cases} \tag{4.2}$$

Both in the original and in the rescaled coordinates, the standard first order explicit Euler discretization of the microscopic model (3.3) is equivalent to the Godunov discretization (2.9) of the macroscopic model; see (7.1), (7.2), (7.3), and (7.4).

5. Hybrid Lagrangian model.

In order to construct our hybrid Lagrangian model, we want to combine a macroscopic description away from the junctions, traffic lights, etc. and a microscopic view near these obstacles as shown in Figure 5.1.

Thanks to the equivalence established above between the two models, we expect to get rid of the usual compatibility problems encountered when developing a hybrid model.

5.1. Description of the two Lagrangian models.

Since the macroscopic model is shown to be the limit of a large number of vehicles on a long stretch of road (away from the vacuum), we may consider that a Lagrangian macroscopic (moving) cell (or the corresponding moving Eulerian cell) contains a "*long vehicle*" made by juxtaposition of *several (say N) ordinary vehicles*, whereas in the microscopic model, a Lagrangian cell contains a *single ordinary vehicle*, as depicted in Figure 5.1. Obviously macroscopic Lagrangian cells are much larger than microscopic ones. In Eulerian hybrid models (see, e.g., [6, 8, 19, 21, 27, 31]), the "microscopic region" is fixed in Eulerian coordinates. In contrast, here, this "actual microscopic region (AMR)" (see Figure 5.1) is piecewise constant in Lagrangian coordinates. It is moving in Eulerian coordinates and is periodically refreshed in order to *always* contain a *fixed* Eulerian region: the "minimal microscopic region (MMR)" (see Figure 5.1) around the junction, traffic light, etc., in which our description will *always* be microscopic.

5.2. Model synchronization.

In this section we show in detail how to pass from the macroscopic to the microscopic description and vice versa. Here and in the next sections, we order the cells and the cars by calling $(i-1)$ the leader and i the follower.

5.2.1. From the macroscopic to the microscopic model.

When a macroscopic Lagrangian cell enters the MMR, we split the "*long vehicle*" (which is nothing but a juxtaposition of N cars) into N different microscopic cells and uniformly distribute these cars over the length L of the macroscopic cell, as shown in Figure 5.2. Here, all our vehicles are supposed to have the same length. In principle, we could also cover the case of vehicles with different lengths; see Remark 5.1 below.

This splitting does not modify the specific volume τ. Indeed, in the macroscopic

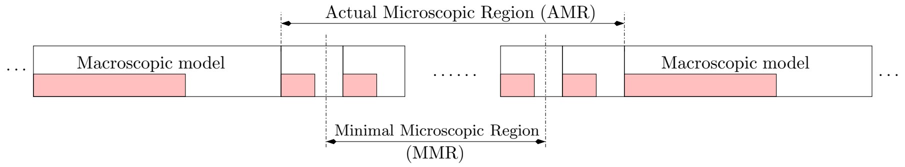


FIG. 5.1. *Hybrid Lagrangian model.*

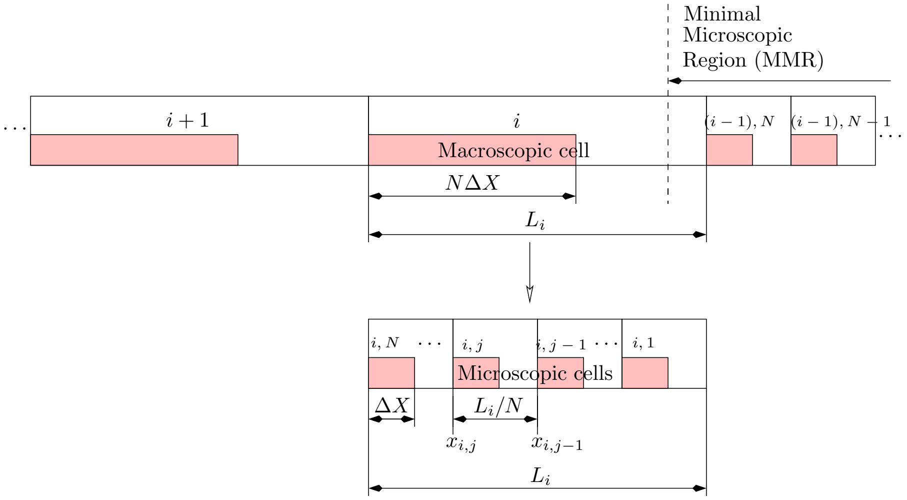


FIG. 5.2. *From the macroscopic to the microscopic model: before (above) and after (below) the synchronization.*

cell i, we have

$$\tau_i = \frac{L_i}{N\Delta X} = \tau_{mac}. \tag{5.1}$$

When this cell i becomes microscopic, the distance between two successive cars (i,j) (the follower) and $(i,j-1)$ (the leader) is

$$(x_{i,j-1} - x_{i,j}) = \frac{L_i}{N}. \tag{5.2}$$

Therefore, the microscopic specific volume in each of these microscopic cells is now

$$\tau_{i,j} = \tau_{mic} = \frac{L_i/N}{\Delta X} = \tau_{mac}. \tag{5.3}$$

So the specific volume does not change when passing from the macroscopic to the microscopic model. Notice that in the Godunov scheme the Lagrangian variable w does not change in time inside a cell i: $w_{i,j}^{n+1} = w_{i,j}^{n}$. Therefore, in the microscopic cells $w_{i,j}$ will be the same as in the macroscopic cell, i.e., $w_{i,j} = w_i$. Consequently, the velocity also does not change:

$$v_{i,j} = w_{i,j} - P(\tau_{i,j}) = w_i - P(\tau_i) = v_i \quad \text{for all microscopic cars } j \text{ in this cell } i.$$

REMARK 5.1. *In the case of vehicles with different lengths, we need to keep the order of vehicles and their respective lengths in each macroscopic cell. Since there is no possibility of overtaking (our model is in principle a one lane model), this order remains constant in time. Let us consider a macroscopic cell i of total length L_i and let $\Delta X_{i,j}$ be the length of the microscopic vehicle j in this cell. When this cell becomes microscopic, we need to distribute uniformly the cars over the length L_i. In the synchronization macro-micro, we distribute uniformly the specific volume among all the vehicles of this cell:*

$$\tau_{i,j} = \tau_{mic} = \frac{x_{i,j-1} - x_{i,j}}{\Delta X_{i,j}} = \tau_i = \frac{L_i}{\sum_{j=1}^{N} \Delta X_{i,j}} = \tau_{mac} \quad \forall\ j = 1, \ldots, N. \tag{5.4}$$

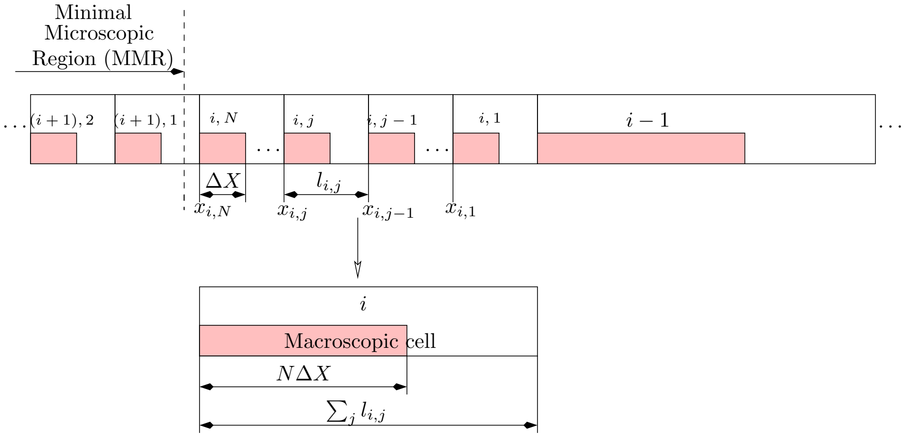


FIG. 5.3. *From the microscopic to the macroscopic model: before (above) and after (below) the synchronization.*

Therefore, we compute the distance $l_{i,j} = (x_{i,j-1} - x_{i,j})$ *for all* $j = 2, \ldots, N$ *by solving the following system:*

$$(5.5) \qquad \begin{cases} \frac{l_{i,j}}{\Delta X_{i,j}} = \frac{l_{i,j-1}}{\Delta X_{i,j-1}} \quad \forall\ j = 1, \ldots, N, \\ \displaystyle\sum_{j=1}^{n} l_{ij} = L_i. \end{cases}$$

5.2.2. From the microscopic to the macroscopic model. When the last (say the (i,N)th) vehicle has completely left the MMR, we do exactly the converse; i.e., we aggregate the N vehicles to form a new macroscopic "*vehicle*." We set $l_{i,j} = x_{ij-1} - x_{i,j}$, with $x_{i,j}$ the position of vehicle (i,j) as indicated in Figure 5.3.

The macroscopic specific volume will be

$$(5.6) \qquad \bar{\tau}_i = \frac{\sum_{j=1}^{N} l_{i,j}}{N\Delta X} = \frac{1}{N}\sum_{j=1}^{N} \frac{l_{i,j}}{\Delta X} = \frac{1}{N}\sum_{j=1}^{N} \tau_{i,j}.$$

The Lagrangian variable $w_{i,j}$ is conserved, since according to subsection 5.2.1 the N vehicles have the same Lagrangian variable $w_{i,j} = w_i$. Thus, averaging in Lagrangian coordinates, we have

$$(5.7) \qquad \frac{1}{N}\sum_{j=1}^{N} w_{i,j} = \frac{1}{N}\sum_{j=1}^{N} w_i = w_i.$$

Therefore, the corresponding macroscopic velocity is $\bar{v}_i = w_i - P(\bar{\tau}_i)$.

In this case, the macroscopic model does not inherit exactly the microscopic parameters but only the average values for τ and w and the above corresponding velocity. In spite of this change of parameters, we will prove in the following section that in the hybrid model the total variation in space of v is nonincreasing, the total variation in space of τ is bounded, and total variations in time of v and τ are bounded.

6. Estimates on the total variation in the hybrid model. In this section, with the above synchronizations between the macroscopic and the microscopic models, we show that in the hybrid model the total variation in space of the velocity v is nonincreasing, the total variation in space of the specific volume τ is bounded, and the total variations in time of v and τ are bounded.

First, the following type of results is classical. Its proof is recalled below for the convenience of the reader.

Lemma 6.1. *Let $U = (u_1, u_2, \dots, u_n) \in \mathbb{R}^n$ and $\bar{u} \in \mathbb{R}$ such that*

$$m = \min_i(u_i) \leq \bar{u} \leq \max_i(u_i) = M.$$

Then,

$$\forall\ \alpha, \beta \in \mathbb{R}, \quad |\alpha - \bar{u}| + |\bar{u} - \beta| \leq |\alpha - u_1| + \sum_{i=1}^{n-1} |u_i - u_{i+1}| + |u_n - \beta|\,. \tag{6.1}$$

Proof.

$$\begin{aligned} |\alpha - \bar{u}| + |\bar{u} - \beta| &= |\alpha - u_1 + u_1 - \bar{u}| + |\bar{u} - u_n + u_n - \beta| \\ &\leq |\alpha - u_1| + |u_1 - \bar{u}| + |\bar{u} - u_n| + |u_n - \beta|\,. \end{aligned} \tag{6.2}$$

Now let us prove that $|u_1 - \bar{u}| + |\bar{u} - u_n| \leq \sum_{i=1}^{n-1} |u_i - u_{i+1}| = TV(u_i)_{i=1,\dots,n}$.

Since the function $\{\bar{u} \longmapsto |u_1 - \bar{u}| + |\bar{u} - u_n|\}$ is convex, its maximum is attained at an extremum point u_k equal to m or M; therefore,

$$\begin{aligned} |u_1 - \bar{u}| + |\bar{u} - u_n| &\leq \max\left(|u_1 - M| + |M - u_n|\,, |u_1 - m| + |m - u_n|\right) \\ &\leq |u_1 - u_k| + |u_k - u_n| \\ &\leq |u_1 - u_2| + \dots + |u_{k-1} - u_k| \\ &\quad + |u_k - u_{k+1}| + \dots + |u_{n-1} - u_n| \\ &= \sum_{i=1}^{n-1} |u_i - u_{i+1}| = TV(u_i)_{i=1,\dots,n}. \qquad \square \end{aligned} \tag{6.3}$$

Let us denote by v_i^n (resp., τ_i^n) and $v_{i,j}^n$ (resp., $\tau_{i,j}^n$), respectively, the macroscopic and the microscopic velocities (resp., specific volumes) at time t_n. At time $t = 0$, the velocities are in $BV(\mathbb{R})$ in both models and therefore in the hybrid one.

According to sections 2 and 3, the results of Proposition 2.3 hold, in both the macroscopic and the microscopic models, away from the synchronized cells. Therefore, we focus on the synchronization process, i.e., when passing from one representation to another.

Passing from a macroscopic cell to several microscopic cells does not change the velocity and the specific volume and therefore does not change their total variation.

On the other hand, when we convert N microscopic cells into a macroscopic one, the total variation of the hybrid model can change. We are going to show that nevertheless the total variation in space of the velocity v is nonincreasing, the total variation in space of the specific volume τ is bounded, and the total variations in time of v and τ are bounded.

Let us denote by TV_X and TV'_X the total variation in space, respectively, before and after this *micro-macro* synchronization. Since inside the synchronized cell (say $\bar{I}_{sync} = \bar{I}_i$), $\bar{\tau} = \frac{1}{N}\sum_j^N \tau_j$, $\bar{v} = w - P(\bar{\tau})$, and P is nonincreasing, we have, by

Lemma 6.1,

$$\begin{aligned} TV'_X(\tau_h(.,t_n),\bar{I}_{sync}) &= \left|\tau^n_{i-1} - \bar{\tau}^n_i\right| + \left|\bar{\tau}^n_i - \tau^n_{i+1,1}\right| \\ &\leq \left|\tau^n_{i-1} - \tau^n_{i,1}\right| + \sum_{j=1}^{N-1} \left|\tau^n_{i,j} - \tau^n_{i,j+1}\right| + \left|\tau^n_{i,N} - \tau^n_{i+1,1}\right| \\ &= TV_X(\tau_h(.,t_n),\bar{I}_{sync}), \end{aligned} \tag{6.4}$$

and similarly

$$\begin{aligned} TV'_X(v_h(.,t_n),\bar{I}_{sync}) &= \left|v^n_{i-1} - \bar{v}^n_i\right| + \left|\bar{v}^n_i - v^n_{i+1,1}\right| \\ &\leq \left|v^n_{i-1} - v^n_{i,1}\right| + \sum_{j=1}^{N-1} \left|v^n_{i,j} - v^n_{i,j+1}\right| + \left|v^n_{i,N} - v^n_{i+1,1}\right| \\ &= TV_X(v_h(.,t_n),\bar{I}_{sync}). \end{aligned} \tag{6.5}$$

Therefore, after any (micro-macro or conversely) synchronization process, the total variation *in space* of v is nonincreasing and the total variation *in space* of τ is bounded:

$$TV'_X(v_h(.,t_n),\bar{I}_{sync}) \leq TV_X(v_h(.,t_n),\bar{I}_{sync}). \tag{6.6}$$

We recall that in each macroscopic cell i, v_h and τ_h are monotonous in time in $[t_n, t_{n+1})$, that is,

$$\min(v^n_i, v^n_{i+1}) \leq v^{n+1}_i \leq \max(v^n_i, v^n_{i+1}), \tag{6.7}$$

and similarly for τ. Therefore, v_h, τ_h and their linear interpolations in time $\tilde{v}_h, \tilde{\tau}_h$ defined in (2.11) satisfy (2.14a) (or (2.15)) and (2.14b); see Proposition 2.3.

Moreover, as we said in sections 3 and 4, the macroscopic and microscopic models are essentially the same. In particular, the same BV estimates hold in microscopic cells for $(\tilde{v}_h, \tilde{\tau}_h)$.

According to the monotonicity property of v (see (6.7)), we have

$$\left|v^{n+1}_i - v^n_{i+1}\right| + \left|v^n_i - v^{n+1}_i\right| = \left|v^n_{i+1} - v^n_i\right|. \tag{6.8}$$

Let us denote, respectively, by $\mathcal{M}^n$ and μ^n the sets of indices of macroscopic cells and discretized macroscopic cells (i.e., split into microscopic cells) at time t_n. Therefore, $\mathcal{M}^n \cup \mu^n = \mathbb{Z}$.

From (6.6) and (6.8) we have

$$TV_X(v_h(.,t_n),\mathbb{R}) \leq TV_X(v_h(.,t_{n-1}),\mathbb{R}) \leq \cdots \leq TV_X(v^0_h,\mathbb{R}) = C_0. \tag{6.9}$$

In each cell I_k, $\tau_h(X,t) = P^{-1}(w_k - v_h(X,t))$, with P^{-1} Lipschitz-continuous. Therefore,

$$TV_X(\tau_h(.,t_n),\{\cup I_k; k \in \mathcal{M}^n \cup \mu^n\}) = \sum_{k\in\mathcal{M}^n\cup\mu^n} TV_X(\tau_h(.,t_n),I_k) \leq C_1 C_0, \tag{6.10}$$

where $C_1 := \left\|(P^{-1})'\right\|_{L^\infty}$.

Now let us study how the total variation in time evolves in our hybrid model:

$$TV_t(\tilde{\tau}_h(.,.);\mathbb{R}\times[t,t']) \leq \sum_{t-\Delta t\leq n\Delta t\leq t'+\Delta t} \left(A_n + B_n + C_n + D_n\right), \tag{6.11}$$

with

$$A_n = \sum_{i \in \mathcal{M}^n \cap \mathcal{M}^{n-1}} \left|\tau_i^n - \tau_i^{n-1}\right| N\Delta X,$$

$$B_n = \sum_{i \in \mu^n \cap \mu^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^n - \tau_{ij}^{n-1}\right| \Delta X,$$

$$C_n = \sum_{i \in \mu^n \cap \mathcal{M}^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^n - \tau_i^{n-1}\right| \Delta X,$$

$$D_n = \sum_{i \in \mathcal{M}^n \cap \mu^{n-1}} \sum_{j=1}^{N} \left|\bar{\tau}_i^n - \tau_{ij}^{n-1}\right| \Delta X.$$

Note that in (6.11) the summation involves all the times $t = n\Delta t$ between times t and t'.

We recall that we have, respectively, in the macroscopic model (see (2.9)) and the microscopic model (see (3.3)) the following:

$$\tau_i^n = \tau_i^{n-1} - \frac{\Delta t}{N\Delta X}\left(v_{i+1}^{n-1} - v_i^n\right) \text{ and} \tag{6.12a}$$

$$\tau_{ij}^n = \tau_{ij}^{n-1} - \frac{\Delta t}{\Delta X}\left(v_{ij+1}^{n-1} - v_{ij}^n\right), \tag{6.12b}$$

$$\begin{aligned} A_n &= \Delta t \sum_{i \in \mathcal{M}^n \cap \mathcal{M}^{n-1}} \left|v_{i+1}^{n-1} - v_i^n\right| \text{ by (6.12a)} \\ &\leq \Delta t \sum_{i \in \mathcal{M}^n \cap \mathcal{M}^{n-1}} \left|v_{i+1}^{n-1} - v_i^{n-1}\right| \text{ due to (6.7)} \\ &\leq \Delta t \sum_{i \in \mathbb{Z}} \left|v_{i+1}^{n-1} - v_i^{n-1}\right| \\ &= \Delta t\ TV_X(v_h(.,t_{n-1}),\mathbb{R}) \leq \Delta t\ TV_X(v_h^0,\mathbb{R}) = \Delta t C_0, \end{aligned}$$

$$\begin{aligned} B_n &= \Delta t \sum_{i \in \mu^n \cap \mu^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^n\right| \text{ by (6.12b)} \\ &\leq \Delta t \sum_{i \in \mu^n \cap \mu^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right| \text{ due to (6.7)} \\ &\leq \Delta t \sum_{i \in \mathbb{Z}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right| \\ &= \Delta t\ TV_X(v_h(.,t_{n-1}),\mathbb{R}) \leq \Delta t\ TV_X(v_h^0,\mathbb{R}) = \Delta t C_0, \end{aligned}$$

$$\begin{aligned} C_n &= \Delta X \sum_{i \in \mu^n \cap \mathcal{M}^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^n - \tau_i^{n-1}\right| \\ &\leq \Delta X \sum_{i \in \mu^n \cap \mathcal{M}^{n-1}} \sum_{j=1}^{N} \left(\left|\tau_{ij}^n - \tau_{ij}^{n-1}\right| + \left|\tau_{ij}^{n-1} - \tau_i^{n-1}\right|\right) \end{aligned}$$

$$
\begin{aligned}
&= \Delta X \sum_{i\in\mu^n\cap\mathcal{M}^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^{n} - \tau_{ij}^{n-1}\right| \text{ (since } \tau_{ij}^{n-1} = \tau_{i}^{n-1}, \\
&\quad \text{according to the macro-micro synchronization process)} \\
&= \Delta t \sum_{i\in\mu^n\cap\mathcal{M}^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n}\right| \\
&\leq \Delta t \sum_{i\in\mu^n\cap\mathcal{M}^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right| \text{ due to (6.7)} \\
&\leq \Delta t \sum_{i\in\mathbb{Z}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right| \\
&= \Delta t \; TV_X(v_h(.,t_{n-1}),\mathbb{R}) \leq \Delta t \; TV_X(v_h^0,\mathbb{R}) = \Delta t C_0,
\end{aligned}
$$

$$
\begin{aligned}
D_n &= \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|\bar{\tau}_{i}^{n} - \tau_{ij}^{n-1}\right| \\
&\leq \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|\bar{\tau}_{i}^{n} - \tau_{ij}^{n}\right| + \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^{n} - \tau_{ij}^{n-1}\right| \\
&=: D_n^1 + D_n^2,
\end{aligned}
$$

$$
\begin{aligned}
D_n^1 &= \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|\bar{\tau}_{i}^{n} - \tau_{ij}^{n}\right| \\
&\leq \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} N \max_{j} \left|\bar{\tau}_{i}^{n} - \tau_{ij}^{n}\right| \\
&\leq N\Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \left|\tau_{ir}^{n} - \tau_{is}^{n}\right|, \\
&\qquad \text{(since } \min_{j}(\tau_{ij}^{n}) = \tau_{is}^{n} \leq \bar{\tau}_{i}^{n} \leq \max_{j}(\tau_{ij}^{n}) = \tau_{ir}^{n}\text{)} \\
&\leq N\Delta X \; TV_X(\tau_h(.,t_n), \cup_{i\in\mathbb{Z}} I_i) \text{ by the same argument as in} \\
&\qquad \text{the proof of Lemma 6.1} \\
&\leq N\Delta X \; TV_X(v_h^0,\mathbb{R}) \left\|(P^{-1})'\right\|_{L^\infty} = N\Delta X C_1 C_0 \text{ due to (6.10),}
\end{aligned}
$$

$$
\begin{aligned}
D_n^2 &= \Delta X \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|\tau_{ij}^{n} - \tau_{ij}^{n-1}\right| \\
&= \Delta t \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n}\right| \\
&\leq \Delta t \sum_{i\in\mathcal{M}^n\cap\mu^{n-1}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right| \text{ due to (6.7)}
\end{aligned}
$$

$$\leq \Delta t \sum_{i\in\mathbb{Z}} \sum_{j=1}^{N} \left|v_{ij+1}^{n-1} - v_{ij}^{n-1}\right|$$
$$= \Delta t \; TV_X(v_h(.,t_{n-1}),\mathbb{R}) \leq \Delta t \; TV_X(v_h^0,\mathbb{R}) = \Delta t C_0.$$

So, $D_n \leq N\Delta X C_1 C_0 + \Delta t C_0$.

Finally, we get

$$TV_t(\tilde{\tau}_h(.,.);\mathbb{R}\times[t,t']) \leq (|t'-t| + 2\Delta t)(N\Delta X C_1 + 4\Delta t)C_0. \tag{6.13}$$

We recall that we study the limit when $\frac{\Delta t}{\Delta X}$ is constant and satisfies the CFL condition (2.8).

In each cell, the specific volume τ_h is monotonous, w_h is constant, and the velocity is

$$v_h = w_h - P(\tau_h),$$

where P is a monotonous Lipschitz continuous function. Therefore, there exists a $C_2 < +\infty$ such that

$$TV_t(\tilde{v}_h(.,.);\mathbb{R}\times[t,t']) \leq (|t'-t| + 2\Delta t)C_2 C_0. \tag{6.14}$$

We summarize the above results in the following.

THEOREM 6.2. *Assume that the sequences* (v_h^0, τ_h^0)*, respectively, the initial data for* v *and* τ *(and therefore for* w*), are in* $BV(\mathbb{R})$*; i.e., there exist some constants* $c_v < +\infty$ *and* $c_\tau < +\infty$ *such that*

$$TV_X(v_h^0;\mathbb{R}) = \sum_{i\in\mathcal{M}^0} \left|v_i^0 - v_{i+1}^0\right| + \sum_{i\in\mu^0}\sum_{j=1}^{N} \left|v_{ij}^0 - v_{ij+1}^0\right| + b_v \leq c_v, \tag{6.15}$$

$$TV_X(\tau_h^0;\mathbb{R}) = \sum_{i\in\mathcal{M}^0} \left|\tau_i^0 - \tau_{i+1}^0\right| + \sum_{i\in\mu^0}\sum_{j=1}^{N} \left|\tau_{ij}^0 - \tau_{ij+1}^0\right| + b_\tau \leq c_\tau, \tag{6.16}$$

where b_v *and* b_τ *are the boundary terms (for instance* $b_v = \sum \left|v_k^0 - v_{lj}^0\right|$ *with* $k \in \mathcal{M}^0$ *and* $l = k \pm 1 \in \mu^0$*).*

Moreover, assume that the CFL condition (2.8) *is satisfied and let* Δt, ΔX*, and* $N\Delta X$ *be constant with* $\frac{\Delta t}{\Delta X}$ = *constant. Then, the following hold:*

(i) *In the macroscopic region* (2.9) *(resp., the microscopic region* (3.3)*), we have the following (see* [2, 4]*):*
  (a) *the total variation in* X *of* $v_h(.,t)$ *(resp., in* t *of* $\tilde{v}_h(.,.)$ *and therefore of* $v_h(.,.)$*) is nonincreasing in time (resp., is bounded on* $\mathbb{R}\times[t,t']$*);*
  (b) *the total variation in* X *of* $\tau_h(.,t)$ *on* $\cup_{j\in\mathbb{Z}} I_j$ *(resp., in* t *of* $\tilde{\tau}_h(.,.)$ *and therefore of* $\tau_h(.,.)$*, on* $\mathbb{R}\times[t,t')$*) is bounded (resp., is bounded).*

(ii) *During the synchronization process, at each time* t_n*, the total variations in* X *of* v_h *and* τ_h *do not increase and their total variations in time are controlled from above.*

(iii) *Therefore, in the whole hybrid model we have the following:*
  (a) *the total variation in* X *of* $v_h(.,t)$ *(resp., in* t *of* $\tilde{v}_h(.,.)$ *and therefore of* $v_h(.,.)$*) is nonincreasing in time thanks to* (6.9) *(resp., is bounded on* $\mathbb{R}\times[t,t']$ *thanks to* (6.14)*);*

(b) *the total variation in X of $\tau_h(.,t)$ on $\cup_{j\in\mathbb{Z}} I_j$ (resp., in t of $\tilde{\tau}_h(.,.)$ and therefore of $\tau_h(.,.)$ on $\mathbb{R}\times[t,t'])$ is bounded thanks to* (6.10) *(resp., is bounded thanks to* (6.13)*).*

We are now in position to establish the convergence of the discrete solution constructed by our hybrid scheme. Let us first recall the following.

Definition 6.3 (see [2, 4]).

(i) *An L^∞ function*

$$U := (\tau, w) : \mathbb{R}\times\mathbb{R}_+ \longrightarrow \mathbb{R}^2$$

is called a weak entropy solution to the Lagrangian system (2.3) *if it is a weak solution (see* (ii)*) and if for any entropy-flux pair $(\eta(\tau,w), q(\tau,w))$ with $\eta(\tau,w)$ convex in τ, and for any $\phi(X,t)\in C_0^\infty(\mathbb{R}\times\mathbb{R}_+)$, $\phi\geq 0$, we have*

$$\int_0^\infty\int_{\mathbb{R}}(\eta(U)\partial_t\phi + q(U)\partial_X\phi)dXdt + \int_{\mathbb{R}}\eta(U_0(X))\phi(X,0)dX \geq 0. \tag{6.17}$$

(ii) *U is called a weak solution to* (2.3) *if the above inequality holds (and therefore is an equality) for the trivial entropy flux pairs $(\eta,q) := (\pm\tau, \mp v)$.*

Adapting the results of [4], it is easy to show that any entropy $\eta(\tau,w)$ is convex in τ if and only if the associated entropy flux $q\equiv q(v)$ is concave (in v), and that there is an associated L^1 contraction principle "à la Kružkov," which implies the uniqueness. Consequently, we obtain the following.

Theorem 6.4.

(i) *The sequence $(U_h = (\tau_h, w_h))$ in the hybrid model (given by the Godunov scheme* (2.9) *of the AR model and the Euler discretization* (3.3) *of the FLM model) is therefore a sequence of approximate weak entropy solutions to* (2.3)*, associated with the initial data $U_h^0 = (\tau_h^0, w_h^0)$.*

(ii) *Consequently, a subsequence (U_h)—in fact, by uniqueness, the whole sequence —converges to the unique weak entropy solution to* (2.3)*, with the same initial data.*

Proof. First, since $\partial_t\tau_h - \partial_X v_h = 0$, for $t_n < t < t_{n+1}$, for any smooth function $\phi(X,t)$ with compact support, we have

$$\begin{aligned}
&\int_0^\infty\int_{\mathbb{R}}(\tau_h\partial_t\phi - v_h\partial_X\phi)(X,t)dXdt + \int_{\mathbb{R}}\tau_h^0(X)\phi(X,0)dX\\
&= \sum_{n\geq 1}\sum_{i\in\mathbb{Z}}\left[\int_{I_i}\tau_h(X,t)\phi(X,t)dX\right]_{t_n^-}^{t_n^+} + \sum_i\int_{I_i}\tau_h^0(X)\phi(X,0)dX\\
&= \sum_{n\geq 1}\sum_{i\in\mathcal{M}^n}\int_{I_i}(\tau_h(X,t_n^-) - \tau_i^n)\phi(X,t_n)dX + \sum_{n\geq 1}\sum_{i\in\mu^n}\sum_{j=1}^N\int_{I_{ij}}(\tau_h(X,t_n^-) - \tau_{ij}^n)\phi(X,t_n)dX\\
&+ \sum_{i\in\mathcal{M}^n\cup\mu^n}\int_{I_i}\tau_h^0(X)\phi(X,0)dX =: A_h.
\end{aligned}$$

Now, since the test function ϕ is C^1 and compactly supported, using a Taylor expansion of ϕ on each interval, we get

$$|A_h| \leq (\Delta X)^2\,\|\partial_X\phi\|_{L^\infty}\sum_{n\geq 0}\sum_{i\in\mathcal{M}^n\cup\mu^n} TV_X(\tau_h(.,t);I_i)$$

$$\leq (\Delta X)^2 \frac{T}{\Delta t} \left\| \partial_X \phi \right\|_{L^\infty} \sup_{n \leq \frac{T}{\Delta t}} \sum_{i \in \mathcal{M}^n \cup \mu^n} TV_X(\tau_h(.,t); I_i)$$
$$\leq \left(\frac{\Delta X}{\Delta t}\right) \Delta X T \left\| \partial_X \phi \right\|_{L^\infty} \sup_{n \leq \frac{T}{\Delta t}} \sum_{i \in \mathcal{M}^n \cup \mu^n} TV_X(\tau_h(.,t); I_i)$$
$$= \left(\frac{\Delta X}{\Delta t}\right)^2 \Delta t T \left\| \partial_X \phi \right\|_{L^\infty} \sup_{n \leq \frac{T}{\Delta t}} \sum_{i \in \mathcal{M}^n \cup \mu^n} TV_X(\tau_h(.,t); I_i).$$

Therefore, thanks to the BV estimates in Theorem 6.2, $|A_h| \longrightarrow 0$ as $\Delta t \longrightarrow 0$ (or $\Delta X \longrightarrow 0$), since $\frac{\Delta X}{\Delta t}$ is constant.

We proceed similarly with the entropy production term. On the one hand, the approximate solutions U_h are weak entropy solutions on any $[t_n, t_{n+1})$, and on the other hand, by Jensen's inequality, any convex entropy does not increase at time t_n in the Godunov averaging step as well as in the *macro-micro* or *micro-macro* synchronization. Consequently, for any entropy $\eta(\tau, w)$ convex with respect to τ, associated with the flux q, for all $\phi \in C_0^\infty(\mathbb{R} \times \mathbb{R}_+)$, $\phi \geq 0$, we have

$$\int_0^\infty \int_{\mathbb{R}} (\eta(U_h)\partial_t \phi + q(U_h)\partial_X \phi)(X,t) dX dt + \int_{\mathbb{R}} \eta(U_h(X,0))\phi(X,0) dX$$
$$\geq \sum_{n \geq 1} \sum_{i \in \mathcal{M}^n} \int_{I_i} (\eta(U_h(X, t_n^-)) - \eta(U_i^n))\phi(X, t_n) dX$$
$$+ \sum_{n \geq 1} \sum_{i \in \mu^n} \sum_{j=1}^{N} \int_{I_{ij}} (\eta(u_h(X, t_n^-)) - \eta(u_{ij}^n))\phi(X, t_n) dX \geq 0. \qquad \square$$

Naturally, the above BV estimates are the crucial ingredient for showing the convergence of such a hybrid scheme to the unique globally defined weak entropy solution of a class of initial boundary value problems for a road with traffic light(s) or a road network in the spirit of [24] or [23]; see also [10]. We will investigate this in a forthcoming work.

7. Numerical simulations.

In this section, we are concerned with the numerical investigation of the hybrid model. We consider the equations given by the explicit Euler time discretization of the microscopic model or equivalently the Godunov discretization of the macroscopic model in Lagrangian mass coordinates, which is itself equivalent to the Godunov discretization in Eulerian coordinates, in Lagrangian (moving) cells. We set the number of vehicles in the macroscopic cell to $N := 10$ and the length of a vehicle to $\Delta X := 5$ m. In order to show how information travels through the different parts of the hybrid model, we have fixed the MMR to a distance of 200 m (i.e., 100 m on both sides of the obstacle), which corresponds to the grey rectangle in Figures 7.1, 7.2, and 7.3. In order to get a better insight into the use of the hybrid model, we consider two scenarios and in each case we compare the densities given by the fully macroscopic model, the hybrid model, and the fully microscopic model.

The vertical line at $x = 0$ in Figures 7.1, 7.2, 7.3, and 7.5 is thickened to emphasize the state of the traffic light.

7.1. Case 1: The same time step in the whole hybrid scheme.

This case corresponds exactly to the assumptions of the theoretical analysis of sections 5 and 6. We consider the same time step for the macroscopic and microscopic parts of the

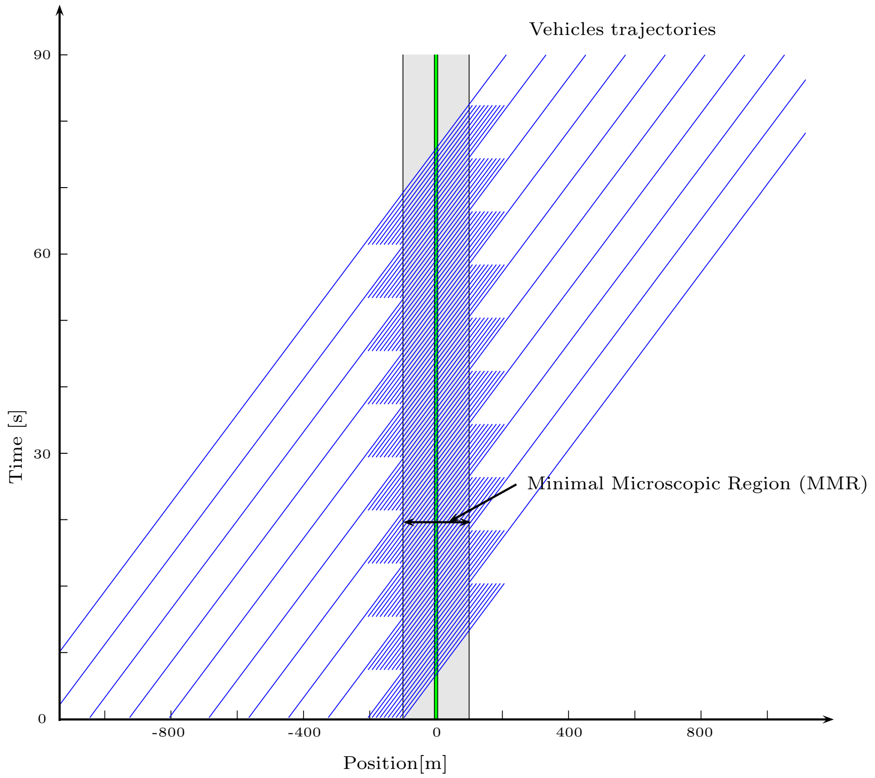


FIG. 7.1. *Case* 1. *The same time step in the microscopic and macroscopic regime: free flow traffic.*

hybrid model. Therefore, at time t_{n+1}, we have

$$\begin{cases} \tau_{ij}^{n+1} = \tau_{ij}^{n} + \frac{\Delta t}{\Delta X}\left(v_{ij-1}^{n} - v_{ij}^{n}\right), \\ w_{ij}^{n+1} = w_{ij}^{n} \end{cases} \tag{7.1}$$

in the microscopic part and

$$\begin{cases} \tau_{i}^{n+1} = \tau_{i}^{n} + \frac{\Delta t}{N\Delta X}\left(v_{i-1}^{n} - v_{ij}^{n}\right), \\ w_{i}^{n+1} = w_{i}^{n} \end{cases} \tag{7.2}$$

in the macroscopic part.

The time step Δt is chosen such that the CFL condition (2.8) is satisfied in both models. Obviously, the more drastic condition is the microscopic one, and thus we also expect too much numerical diffusion in the macroscopic part. For the simulations below the time step Δt is such that the Courant number is equal to 1 in the microscopic part.

The position of the vehicles is computed as follows:

$$x_{ij}^{n+1} = x_{ij}^{n} + \Delta t v_{ij}^{n}, \quad \text{with} \quad v_{ij}^{n} = w_{ij}^{n} - P(\tau_{ij}^{n}), \tag{7.3}$$

$$x_{i}^{n+1} = x_{i}^{n} + \Delta t v_{i}^{n}, \quad \text{with} \quad v_{i}^{n} = w_{i}^{n} - P(\tau_{i}^{n}). \tag{7.4}$$

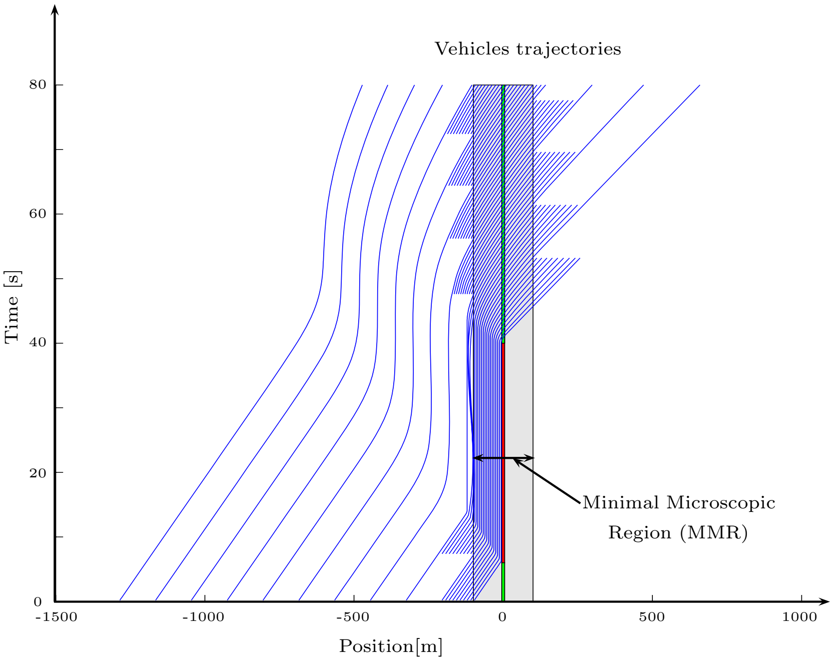


FIG. 7.2. *Case* 1. *The same time step in the microscopic and macroscopic regime: first a shock wave and then a rarefaction fan, both produced in the microscopic region, propagating backward in the macroscopic region.*

First, we simulate the hybrid model in a free flow situation. Trajectories of the "vehicles" in space and time for this case are plotted in Figure 7.1. In Figure 7.2, we plotted the "vehicles" trajectories when a "shock wave" appears at $t = 5$ s, followed by a "rarefaction fan" at $t = 40$ s in the minimal microscopic region (e.g., at a traffic light). Finally, Figure 7.3 shows the vehicles trajectories when a shock wave appears at $t = 40$ s followed by a rarefaction fan at $t = 70$ s, both downstream in the macroscopic regime (typically an incident on a highway). These "shocks" or "rarefaction waves" are produced numerically by forcing the leading vehicle to brake or accelerate. The simulations of these three situations show that the model synchronization does not perturb the wave propagation and allows for a nice description in each regime. In Figure 7.4, for the same situation (first a shock wave and then a rarefaction fan, both downstream of the minimal microscopic region), we plot the density for the fully macroscopic model, the hybrid model, and the fully microscopic model, respectively. The fully microscopic model gives a more precise description (but would be intractable for a large road network). In contrast, the description given by the macroscopic model is rather coarse to describe precisely the details in the minimal microscopic region. We also note that the shock is completely smoothed out by the numerical diffusion since the macroscopic CFL condition is $\frac{1}{10}$. Finally, the hybrid scheme gives an intermediate description, which is precise only in the region in which the details are important, in particular, here in the MMR [−100 m; +100 m].

7.2. Case 2: Different time steps in the hybrid scheme. This case is not exactly covered by the assumptions of sections 5 and 6, even though the same ideas could in principle be extended to this situation. We consider different time steps in the two parts (microscopic part and macroscopic part) of the hybrid scheme, in order

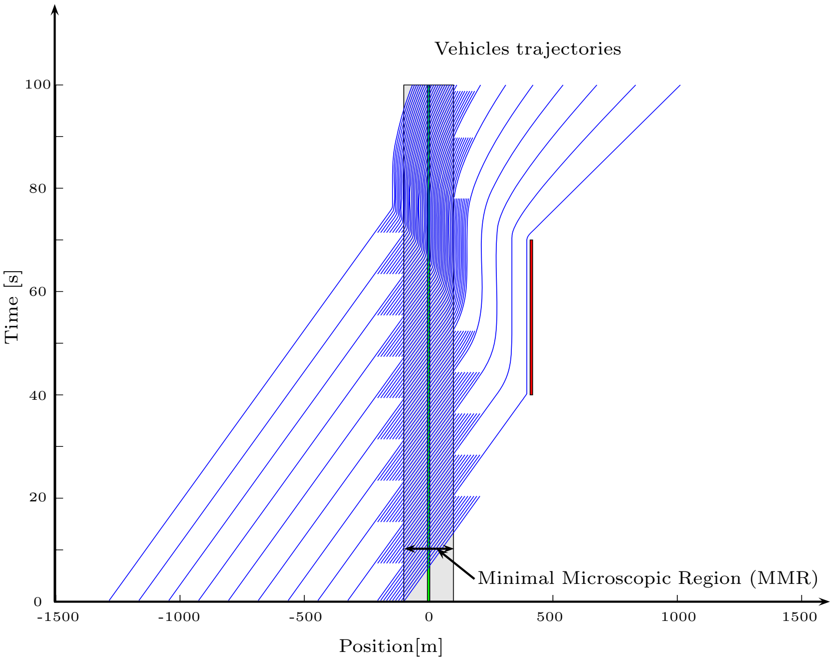


FIG. 7.3. *Case* 1. *The same time step in the microscopic and macroscopic regime: first a shock wave and then a rarefaction fan, both produced in the macroscopic region, propagating backward in the microscopic region.*

to have a Courant number (CFL condition) smaller than 1, but as large as possible in each region. Let $\Delta t_{\rm mic}$ and $\Delta t_{\rm mac}$ be, respectively, the time step in the microscopic part and the macroscopic part. In the microscopic region the updating takes place at each microscopic time $t_{n+1} = t_n + \Delta t_{\rm mic}$, i.e.,

$$\begin{cases} \tau_{ij}^{n+1} = \tau_{ij}^n + \frac{\Delta t_{\rm mic}}{\Delta X}\left(v_{ij-1}^n - v_{ij}^n\right), \\ w_{ij}^{n+1} = w_{ij}^n, \\ x_{ij}^{n+1} = x_{ij}^n + \Delta t_{\rm mic} v_{ij}^n, \\ v_{ij}^{n+1} = w_{ij}^{n+1} - P(\tau_{ij}^{n+1}), \end{cases} \tag{7.5}$$

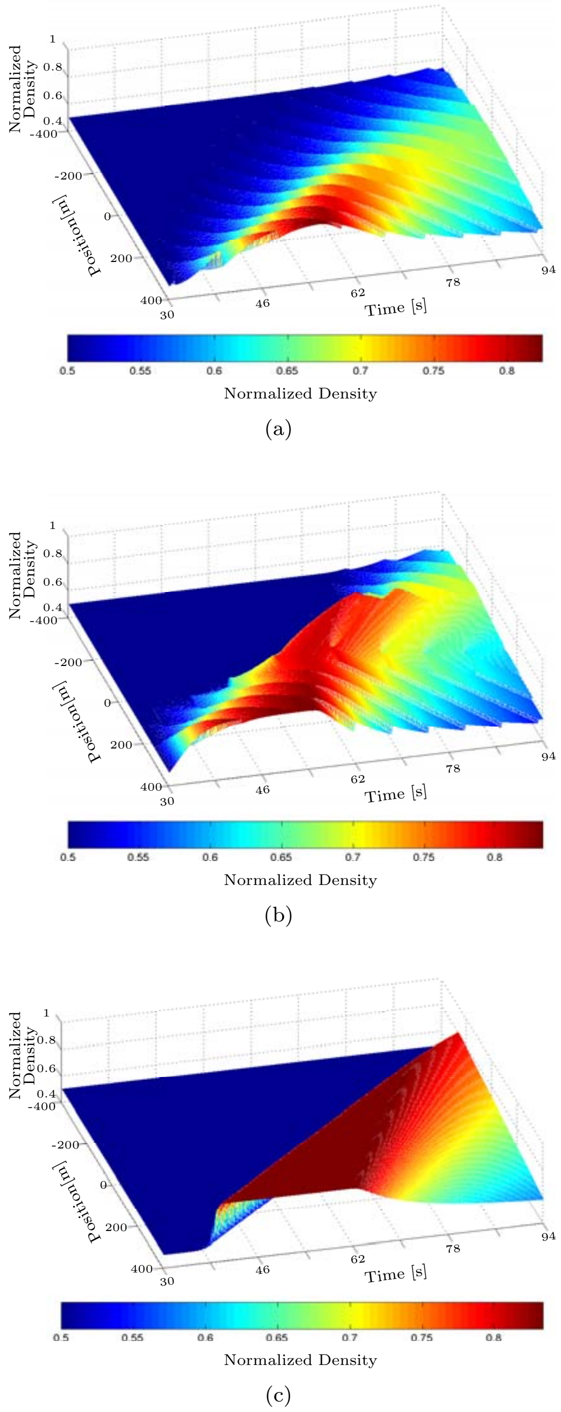
with $\Delta t_{\rm mic}$ chosen in such a way that the CFL condition (2.8) is satisfied in the microscopic model.

In contrast, in the macroscopic region, the updating takes place at each macroscopic time $\bar{t}_{n+1} = \bar{t}_n + \Delta t_{\rm mac}$, i.e.,

$$\begin{cases} \tau_{i}^{n+1} = \tau_{i}^n + \frac{\Delta t_{\rm mac}}{N\Delta X}\left(v_{i-1}^n - v_{i}^n\right), \\ w_{i}^{n+1} = w_{i}^n, \\ x_{i}^{n+1} = x_{i}^n + \Delta t_{\rm mac} v_{i}^n, \\ v_{i}^{n+1} = w_{i}^{n+1} - P(\tau_{i}^{n+1}), \end{cases} \tag{7.6}$$

with $\Delta t_{\rm mac}$ chosen such that the CFL condition (2.8) is satisfied in the macroscopic model. Moreover, $\Delta t_{\rm mac} = k\Delta t_{\rm mic}$, with $1 < k \leq N$ and in general $k = N$.

One of the severe test cases is to handle the propagation of a strong shock, produced, e.g., at a (red) traffic light. In order to better track the propagation of this



(a)

(b)

(c)

FIG. 7.4. *Case* 1. *The same time step in the microscopic and macroscopic regime: first a shock wave and then a rarefaction fan, both downstream from the minimal microscopic region* ([−100 m, 100 m]): (a) *the density in the macroscopic model,* (b) *the density in the hybrid model,* (c) *the density in the microscopic model.*

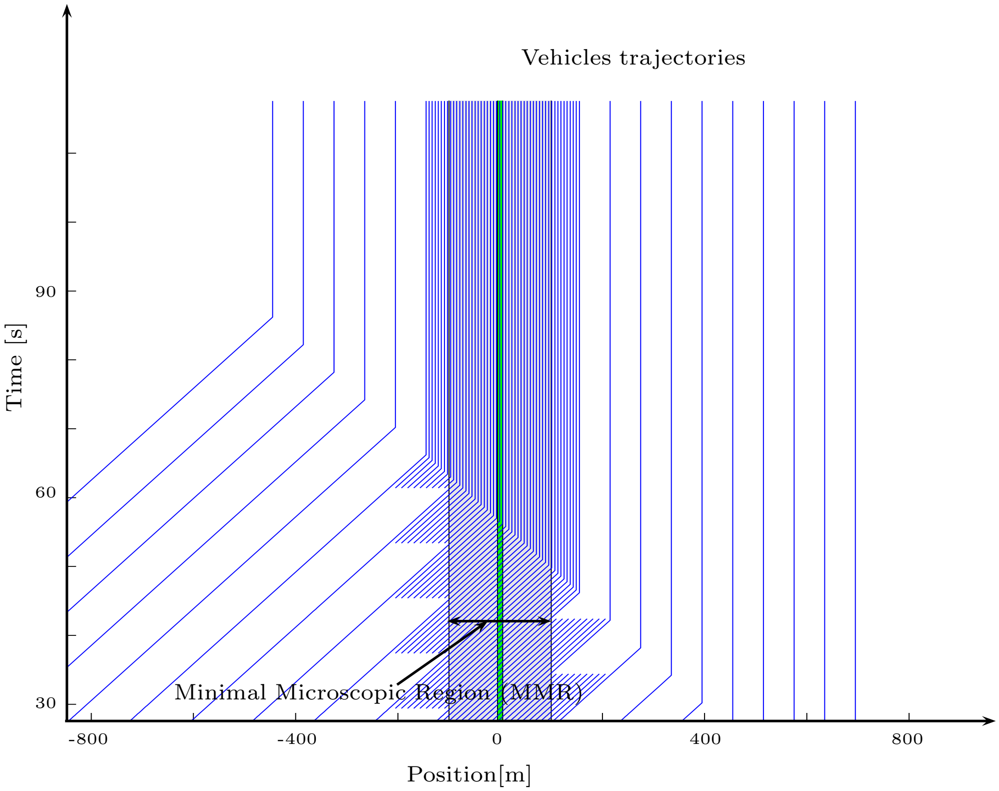
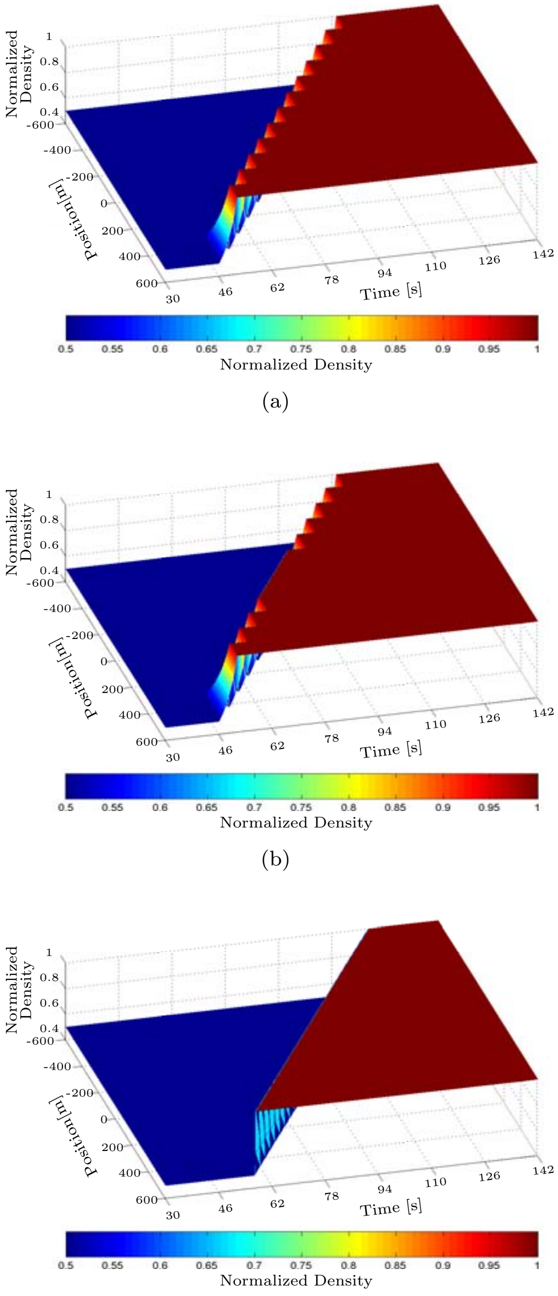


FIG. 7.5. *Case* 2. *Different time steps in the microscopic and macroscopic regime: a shock wave produced in the macroscopic region, propagating backward in the microscopic region.*

strong shock, we could use a front tracking method, which is a completely different numerical strategy. Here, we have used the following numerical procedure: At the macroscopic-microscopic interface, we include the following procedure to treat the case of a strong shock, as it happens at a traffic light. As soon as the last microscopic vehicle of a microscopic cell $(i-1)$ stops, the macroscopic cell i just behind this vehicle is tracked at each microscopic time step, until the time at which the "macroscopic vehicle" in cell i gets nose to tail with the microscopic vehicle ahead. Let t_s be this time. If t_s does not correspond to a macroscopic time, then we shift the macroscopic time step, not only for the macroscopic cell i but, exceptionally, for all the macroscopic cells behind cell i, which are updated at this time t_s using (once) the time step $\tilde{\Delta} t_{\text{mac}} = (t_s - \bar{t}_n)$, where $\bar{t}_n$ is the last macroscopic time before t_s. Thereafter, the time t_s becomes the beginning of the macroscopic time step for the cells behind cell i. We also recall that $\bar{t}_n \leq t_s \leq \bar{t}_n + \Delta t_{\text{mac}} = \bar{t}_{n+1}$.

For the numerical simulations plotted in Figures 7.5 and 7.6, we consider the CFL condition (2.8) with equality in both the microscopic and the macroscopic parts, and we set $\Delta t_{\text{mac}} = N \Delta t_{\text{mic}}$. As in the previous case, the fully microscopic model gives a more precise description (still too heavy for a large road network). In contrast, the description given by the macroscopic scheme is still a bit coarse to describe precisely the details in the MMR, but on the other hand, the shock is now much sharper even on this relatively coarse grid, thanks to the much better macroscopic CFL condition. Finally, the hybrid model gives an intermediate description which is very precise in the



(a)

(b)

(c)

FIG. 7.6. *Case* 2. *Different time steps in the microscopic and macroscopic regime: a shock travelling from downstream to upstream coming from downstream of the minimal microscopic region (*[−100 m, 100 m]*):* (a) *the density in the macroscopic model,* (b) *the density in the hybrid model,* (c) *the density in the microscopic model.*

MMR and still tractable elsewhere. Note in particular that the numerical propagation speed of the shock is quite well preserved in each region.

8. Conclusion.

Traffic investigation has simultaneously progressed both on the macroscopic and microscopic front. However, each description has its own limits. The recent success of hybrid models is due to their ability to capture traffic dynamics in a large domain with enough details near the obstacles.

In this paper, we propose a simple hybrid model based solely on a Lagrangian discretization of both the macroscopic and the microscopic models, coupled through Lagrangian interfaces periodically refreshed in order to *always* contain a fixed *Eulerian* region near an obstacle. As shown in the above numerical examples, the waves produced in either region nicely propagate through the other region. In particular, by this construction, the mass is automatically conserved through the interfaces.

This approach, which establishes a link between microscopic and macroscopic models and allows one to carry out simultaneous macro-micro simulations, could be hopefully a promising way to treat road intersections, in particular, on urban road networks.

REFERENCES

[1] B. Argall, E. Cheleshkin, J. M. Greenberg, C. Hinde, and P.-J. Lin, *A rigorous treatment of a follow-the-leader traffic model with traffic lights present*, SIAM J. Appl. Math., 63 (2002), pp. 149–168.

[2] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[3] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[4] P. Bagnerini and M. Rascle, *A multiclass homogenized hyperbolic model of traffic flow*, SIAM J. Math. Anal., 35 (2003), pp. 949–973.

[5] C. Bardos, A. LeRoux, and J. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1043.

[6] E. Bourrel and J. B. Lessort, *Mixing microscopic and macroscopic representations of traffic flow: Hybrid model based on the Lighthill-Whitham–Richards theory*, Transportation Research Record, 1852 (2003), pp. 193–200.

[7] A. Bressan and H. K. Jenssen, *On the convergence of Godunov scheme for nonlinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 269–284.

[8] W. Burghout, H. Koutsopoulos, and I. Andréasson, *Hybrid mesoscopic-microscopic traffic simulation*, Transportation Research Record, 1934 (2005), pp. 218–225.

[9] G.-Q. Chen, *Propagation and cancellation of oscillations for hyperbolic systems of conservation laws*, Comm. Pure Appl. Math., 44 (1991), pp. 121–140.

[10] G. M. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[11] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.

[12] C. F. Daganzo, *The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory*, Trans. Res. B, 28 (1994), pp. 269–287.

[13] C. F. Daganzo, *Requiem for second order fluid approximations of traffic flow*, Trans. Res. B, 29 (1995), pp. 277–286.

[14] R. J. Diperna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[15] D. C. Gazis, R. Herman, and R. W. Rothery, *Nonlinear follow-the-leader models of traffic flow*, Oper. Res., 9 (1961), pp. 545–567.

[16] E. Godlewsky and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.

[17] S. K. Godunov, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb. (N.S.), 47 (1959), pp. 271–306.

[18] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[19] D. Helbing, A. Hennecke, V. Shvetsov, and M. Treiber, *Micro- and macro-simulation of freeway traffic*, Math. Comput. Modelling, 35 (2002), pp. 517–547.

[20] D. Helbing and M. Treiber, *Critical discussion of "synchronized flow,"* Cooper@tive Tr@nsport@tion Dyn@mics, 1 (2002), pp. 2.1.–2.24.

[21] A. Hennecke, M. Treiber, and D. Helbing, *Macroscopic simulations of open systems and micro-macro link*, in Traffic and Granular Flow '99, D. Helbing, H. J. Herrmann, M. Schreckenberg, and D. E. Wolf, eds., Springer, Berlin, 2000, pp. 383–388.

[22] R. Herman and I. Prigogine, *Kinetic Theory of Vehicular Traffic*, American Elsevier, New York, 1971.

[23] M. Herty, S. Moutari, and M. Rascle, *Optimization criteria for modelling intersections of vehicular traffic flow*, Netw. Heterog. Media, 1 (2006), pp. 275–294.

[24] M. Herty and M. Rascle, *Coupling conditions for a class of second-order models for traffic flow*, SIAM J. Math. Anal., 38 (2006), pp. 595–616.

[25] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[26] S. N. Kružkov, *First order quasilinear equations with several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[27] J. Laval and C. F. Daganzo, *Lane-changing in traffic streams*, Trans. Res. B, 40 (2006), pp. 251–264.

[28] J. P. Lebacque, *Les modèles macroscopiques du traffic*, Ann. des Ponts, 67 (1993), pp. 24–45.

[29] J. P. Lebacque, *The Godunov scheme and what it means for first order traffic flow models*, in Transportation and Traffic Theory, J. B. Lessort, ed., Pergamon Press, Oxford, UK, 1996, pp. 647–678.

[30] M. Lighthill and J. Whitham, *On kinematic waves*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 281–345.

[31] L. Magne, S. Rabut, and J. F. Gabard, *Towards an hybrid macro-micro traffic flow simulation model*, in Proceedings of the INFORMS Spring Meeting, Salt Lake City, UT, 2000.

[32] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.

[33] H. Payne, *FREFLO: A macroscopic simulation model for freeway traffic*, Transportation Research Record, 722 (1979), pp. 68–75.

[34] L. Tong, *Well-posedness theory of an inhomogeneous traffic flow model*, Discrete Contin. Dyn. Syst. Ser. B, 2 (2002), pp. 401–414.

[35] D. H. Wagner, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.

[36] G. C. K. Wong and S. C. Wong, *A multi-class traffic flow model—an extension of LWR model with heterogeneous drivers*, Trans. Res. A, 36 (2002), pp. 827–841.

[37] H. M. Zhang, *A non-equilibrium traffic model devoid of gas-like behaviour*, Trans. Res. B, 36 (2002), pp. 275–298.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 437–460

RESTORATION OF COLOR IMAGES BY VECTOR VALUED BV FUNCTIONS AND VARIATIONAL CALCULUS[*]

MASSIMO FORNASIER[†] AND RICCARDO MARCH[‡]

Abstract. We analyze a variational problem for the recovery of vector valued functions and compute its numerical solution. The data of the problem are a small set of complete samples of the vector valued function and some significant incomplete information where the former are missing. The incomplete information is assumed as the result of a distortion, with values in a lower dimensional manifold. For the recovery of the function we minimize a functional which is formed by the discrepancy with respect to the data and total variation regularization constraints. We show the existence of minimizers in the space of vector valued bounded variation functions. For the computation of minimizers we provide a stable and efficient method. First, we approximate the functional by coercive functionals on $W^{1,2}$ in terms of Γ-convergence. Then we realize approximations of minimizers of the latter functionals by an iterative procedure to solve the PDE system of the corresponding Euler–Lagrange equations. The numerical implementation comes naturally by finite element discretization. We apply the algorithm to the restoration of color images from limited color information and gray levels where the colors are missing. The numerical experiments show that this scheme is very fast and robust. The reconstruction capabilities of the model are shown, also from very limited (randomly distributed) color data. Several examples are included from the real restoration problem of A. Mantegna's art frescoes in Italy.

Key words. color image processing, systems of partial differential equations, calculus of variations, finite element method

AMS subject classifications. 65M60, 94A08, 49M30, 49J45

DOI. 10.1137/060671875

1. Introduction and examples.

This paper concerns the analysis and the numerical implementation of a variational model for the restoration of vector valued functions. The restoration is obtained from few and sparse *complete* samples of the function and from significant *incomplete* information. The latter is assumed as the result of a nonlinear distortion and with values in a lower dimensional manifold. The applications we consider are in the field of digital signal and image restoration. Therefore, we deal with functional analysis in the space of bounded variation (BV) functions, which are actually considered a reasonable functional model for natural images and signals, usually characterized by discontinuities and piecewise smooth behavior. While in the literature on mathematical image processing mainly real valued BV functions and associated variational problems are discussed (see, for example, [7, 31]), in this contribution we consider vector valued functions.

Since the work of Mumford and Shah [29] and Rudin, Osher, and Fatemi [30], variational calculus techniques have been applied in several image processing problems. We refer the reader to the introductory book [6] for a presentation of this field, more details, and an extended literature.

[*] Received by the editors October 9, 2006; accepted for publication (in revised form) July 6, 2007; published electronically November 28, 2007.
http://www.siam.org/journals/siap/68-2/67187.html

[†] Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000 (mfornasi@math.princeton.edu). This author's work was supported by the European Union's Human Potential Programme under contract MOIF-CT-2006-039438.

[‡] Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo (IAC) "Mauro Picone," Viale del Policlinico 137, 00161 Roma, Italy (r.march@iac.cnr.it).

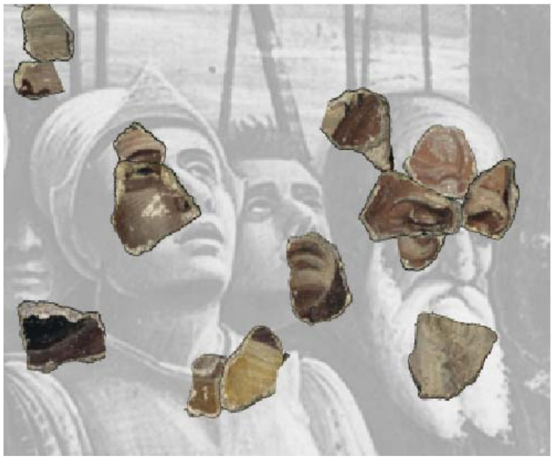


FIG. 1.1. *Fragments of A. Mantegna's frescoes* (1452)*, destroyed by a bombing in the Second World War. Computer based reconstruction by using efficient pattern matching techniques* [24] *is shown.*

Inspired by the fresco problem shown in Figure 1.1, a variational model has been proposed by one of the authors in [21]. The problem consists in recovering color in A. Mantegna's frescoes, which were partially destroyed by a bombing in the Second World War. Pieces of the frescoes with the original colors remain, while black and white photos, taken before the war, of the full frescoes are available. Unfortunately, the surface covered by the original fragments is only 77 m^2, while the original area was of several hundreds. This means that what we can currently reconstruct is just a *fraction* (estimated up to 8%) of what this inestimable artwork was. In particular, for most of the frescoes, the *original color* of the blanks is not known. So, natural questions arise: Is it possible to estimate *mathematically* the original colors of the frescoes by using the known fragments' information and the gray level of the pictures taken before the damage? And, how *faithful* is this estimation?

We now introduce some notations. Let Ω be an open, bounded, and connected subset of $\mathbb{R}^N$, and $D \subset \Omega$. The fresco problem is modeled as the reconstruction/restoration of a vector valued function $u : \Omega \to \mathbb{R}^M$ from a given observed couple of functions $(\bar{u}, \bar{v})$. The observed function $\bar{u}$ is assumed to represent correct information on $\Omega \backslash D$, and $\bar{v}$ is the result of a *nonlinear distortion* $\mathcal{L} : \mathbb{R}^M \to \mathbb{R}$ on D.

In particular, a digital image can be modeled as a function $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3_+$, so that, with each "point" $\mathbf{x}$ of the image, one associates the vector $u(\mathbf{x}) = (r(\mathbf{x}), g(\mathbf{x}), b(\mathbf{x})) \in \mathbb{R}^3_+$ of the color represented by the different channels: red, green, and blue. In particular, a digitalization of the image u corresponds to its sampling on a regular lattice $\tau\mathbb{Z}^2$, $\tau > 0$. Let us again write $u : \mathcal{N} \to \mathbb{R}^3_+$, $u(\mathbf{x}) = (r(\mathbf{x}), g(\mathbf{x}), b(\mathbf{x}))$ for $\mathbf{x} \in \mathcal{N} := \Omega \cap \tau\mathbb{Z}^2$.

Usually the gray level of an image can be described as a submanifold $\mathcal{M} \subset \mathbb{R}^3$ by

$$\mathcal{M} := \mathcal{M}_\sigma = \{\sigma(x) : x = \mathcal{L}(r, g, b) := L(\alpha r + \beta g + \gamma b),\ (r, g, b) \in \mathbb{R}^3_+\},$$

where $\alpha, \beta, \gamma > 0$, $\alpha + \beta + \gamma = 1$, $L : \mathbb{R} \to \mathbb{R}$ is a nonnegative increasing function, and $\sigma : \mathbb{R}_+ \to \mathbb{R}^3_+$ is a suitable section such that $\mathcal{L} \circ \sigma = \mathrm{id}_{\mathbb{R}_+}$. The function L is assumed smooth, nonlinear, and normally nonconvex and nonconcave. For example, Figure 1.2 describes the typical shape of an L function, which is estimated by fitting

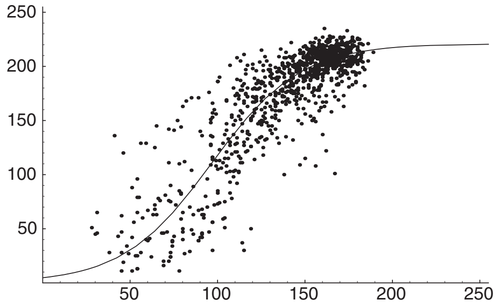


FIG. 1.2. *Estimate of the nonlinear curve L from a distribution of points with coordinates given by the linear combination $\alpha r + \beta g + \gamma b$ of the (red, green, blue) color fragments (abscissa) and by the corresponding underlying gray level of the original photographs dated to* 1920 *(ordinate). The sensitivity parameters α, β, γ to the different frequencies of red, green, and blue are chosen in order to minimize the total variance of the ordinates.*

a distribution of data from the real color fragments in Figure 1.1.

The variational problem proposed in [21] is the following:

$$\arg \inf_{u:\Omega\to\mathbb{R}^M} \Bigg\{ F(u) = \mu \int_{\Omega\setminus D} |u(x) - \bar{u}(x)|^p dx + \lambda \int_D |\mathcal{L}(u(x)) - \bar{v}(x)|^p dx \tag{1.1}$$
$$+ \int_\Omega \sum_{i=1}^M \phi(|\nabla u_i(x)|) dx \Bigg\},$$

where $p \geq 1$. For example, Figure 1.1 illustrates a typical situation where this model applies. In fact, in this case, there is an area $\Omega\setminus D$ of the domain $\Omega \subset \mathbb{R}^2$ of the image, where some fragments with colors are placed and complete information is available, and another area D (which we call the *inpainting region*), where only the gray-level information is known, modeled as the image of $\mathcal{L}$. The solution of the variational problem (1.1) produces in this case a new color image that extends the colors of the fragments in the gray region. Once the extended color image is transformed by means of $\mathcal{L}$, it is constrained to match the known gray level. We can consider this problem as a generalization of the well-known *image inpainting/disocclusion*; see, e.g., [3, 8, 9, 13, 14, 15, 16]. Several heuristic algorithms have been introduced for colorization of gray images; we refer the reader to the recently appeared paper [33] for related literature and to [26] for numerical examples. Nevertheless, our approach is theoretically founded, more general, and fits with many possible applications, for example, the recovery of a transmitted multichannel signal affected by a stationary (nonlinear) distortion.

For $N = p = 2$, we can compute the Euler–Lagrange equations associated with the functional F and obtain

$$0 = -\nabla \cdot \left(\frac{\phi'(|\nabla u_i|)}{|\nabla u_i|} \nabla u_i \right)$$
$$+ 2\mu(u_i - \bar{u}_i)1_{\Omega\setminus D} + 2\lambda(\mathcal{L}(u) - \bar{v}) \frac{\partial \mathcal{L}}{\partial u_i}(u) 1_D := \mathcal{E}_i(\mathcal{L}, u), \tag{1.2}$$

$i = 1, \ldots, M$, where $u = (u_1, \ldots, u_M)$ are the components of the function u. This is a system of coupled second order equations, and the analysis of the solutions itself constitutes a problem of independent interest. By using (1.2) and a finite difference approximation, a steepest-descent algorithm can be formulated as in [21].

Encouraged by the numerical evidence in [21], we discuss the existence of minimizers of the functional F in the context of vector valued BV functions. Our second goal is the formulation of efficient and stable algorithms for the computation of minimizers. Although the steepest-descent scheme recalled above gives appreciable results, it lacks a rigorous analysis and its convergence is usually very slow. For these reasons, we introduce new coercive functionals F_h on $W^{1,2}$ which approximate $\bar{F}$ (the relaxed functional of F with respect to the BV weak-$*$-topology) in terms of Γ-convergence. The computation of minimizers of F_h is performed by an iterative *double-minimization* algorithm; see also [12]. The reconstruction performances are very good, also from very limited (randomly distributed) color data. The virtues of our scheme can be summarized as follows.

1. It is derived as the minimization of a functional and its mathematical analysis and foundations are well described.
2. It implements a total variation (TV) minimization. It is well known [14, 15] that total variation inpainting is affected by two major drawbacks. The first one is that the TV model is only a linear interpolant; i.e., the broken isophotes are interpolated by straight lines. Thus it can generate corners along the inpainting boundary. The second one is that TV often fails to connect widely separated parts of a whole object, due to the high cost in TV measure of making long-distance connections. Due to the constraint on the gray level in the inpainting region, our scheme does not extend isophotes as straight lines and does not violate the *connectivity principle*.
3. As pointed out in [11, 23], while it is relatively easy to recover at higher resolution image portions with relatively uniform color, it might be difficult to recover jumps correctly. Not only should we preserve the morphology and enhance the detail of the discontinuities, but these properties must fit through the different color channels. An incorrect or uncoupled recovery in fact produces "rainbow effects" around jumps. In our functional, the constraint on the gray level in the inpainting region is formulated as a coupled combination of the color channels. In practice, this is sufficient to enforce the correct coupling of the channels at edges.
4. The numerical implementation of our double-minimization scheme is very simple. Its approximation by finite elements comes in a natural way. The scheme is fast and stable.

The paper is organized as follows. In section 2 we introduce the mathematical setting. We recall the main properties of BV functions and a definition of the space of BV functions with vector values. Section 3 is dedicated to results on convex functions and relaxed functionals of measures. In section 4 we collect the assumptions on the nonlinear function $\mathcal{L}$ we will need in our analysis. In section 5 the representation of the relaxed functional $\bar{F}$ of F with respect to the BV topology is given, and the existence and uniqueness of minimizers of $\bar{F}$ are discussed. In section 6 we introduce coercive functionals F_h on $W^{1,2}$ which are shown to Γ-converge to the relaxed functional described above. The double-minimization algorithm to compute minimizers of F_h is illustrated in section 7. Its numerical implementation is presented in section 8. We include several numerical experiments and discuss their results.

2. Vector valued BV functions.

In this section we want to introduce notations and preliminary results concerning vector valued BV functions.

We denote by $\mathcal{L}_N$ (and in the integrals dx) the Lebesgue N-dimensional measure in $\mathbb{R}^N$ and by $\mathcal{H}_\alpha$ the α-dimensional Hausdorff measure. Let Ω be an open, bounded, and connected subset of $\mathbb{R}^N$. With $\mathcal{B}(\Omega)$ we denote the family of Borel subsets of $\Omega \subset \mathbb{R}^N$. For a given vector valued measure $\mu : \mathcal{B}(\Omega) \to \mathbb{R}^M$, we denote by $|\mu|$ its total variation, i.e., the finite positive measure

$$|\mu|(A) := \sup \left\{ \sum_{j=1}^{M} \int_\Omega v_j d\mu_j : v = (v_1, \dots, v_M) \in C_0(A; \mathbb{R}^M), \ \|v\|_\infty \leq 1 \right\},$$

for every open set $A \subset \Omega$, where $C_0(A; \mathbb{R}^M) := \overline{C_c(A; \mathbb{R}^M)}^{\|\cdot\|_\infty}$, i.e., the sup-norm closure of the space of continuous function with compact support in A and vector values in $\mathbb{R}^M$. The set of the signed measures on Ω with bounded total variation is denoted by $\mathcal{M}(\Omega)$, coinciding in fact with the topological dual of $(C_0(A; \mathbb{R}^M), \| \cdot \|_\infty)$. Thus, the usual weak-$*$-topology on $\mathcal{M}(\Omega)$ is the weakest topology that makes the maps $\mu \to \int_\Omega f d\mu$ continuous for every continuous function $f \in C_0(A; \mathbb{R}^M)$. In the following we will make use of the notations $x \wedge y := \inf\{x, y\}$ and $x \vee y := \sup\{x, y\}$ for all $x, y \in \mathbb{R}$.

We say that $u \in L^1(\Omega)$ is a real function of bounded variation if its distributional derivative $Du = (D_{x_1} u, \dots, D_{x_N} u)$ is in $\mathcal{M}(\Omega)$. Then the space of bounded variation functions is denoted by

$$BV(\Omega) := \{u \in L^1(\Omega) : Du \in \mathcal{M}(\Omega)\}$$

and, endowed with the norm $\|u\|_{BV(\Omega)} := \|u\|_1 + |Du|(\Omega)$, is a Banach space [20]. More generally, we are interested in vector valued functions with bounded variation components, whose space is defined by

$$BV(\Omega; \mathbb{R}^M) := \{u = (u_1, \dots, u_M) \in L^1(\Omega; \mathbb{R}^M) : u_i \in BV(\Omega)\}.$$

To this space it will turn out to be convenient to attach the norm $\|u\|_{BV(\Omega;\mathbb{R}^M)} := \|u\|_{L^1(\Omega;\mathbb{R}^M)} + \sum_{i=1}^M |Du_i|(\Omega)$. With a slight abuse of notation, for $u \in BV(\Omega; \mathbb{R}^M)$ we denote

$$|Du| := \sum_{i=1}^{M} |Du_i|, \tag{2.1}$$

which again is a finite positive measure for Ω. The space $(BV(\Omega; \mathbb{R}^M), \| \cdot \|_{BV(\Omega;\mathbb{R}^M)})$ is a Banach space. Of course $BV(\Omega; \mathbb{R}^M) = BV(\Omega)$ for $M = 1$, and our notations are consistent with this case.

The product topology of the strong topology of $L^1(\Omega; \mathbb{R}^M)$ for u and of the weak-$*$-topology of measures for Du_i (for all $i = 1, \dots, M$) will be called the weak-$*$-topology of $BV(\Omega; \mathbb{R}^M)$ or the componentwise BV weak-$*$-topology. In the following, whenever the domain Ω and the dimension M will be clearly understood, we will write L^1 instead of $L^1(\Omega; \mathbb{R}^M)$ and BV instead of $BV(\Omega; \mathbb{R}^M)$.

We further recall the main structure properties of BV functions [1, 2, 20]. If $v \in BV(\Omega)$, then the Lebesgue decomposition of Dv with respect to the Lebesgue measure $\mathcal{L}_N$ is given by

$$Dv = \nabla v \cdot \mathcal{L}_N + D_s v,$$

where $\nabla u = \frac{d(Dv)}{dx} \in L^1(\Omega;\mathbb{R}^N)$ is the Radon–Nikodym derivative of Dv and $D_s v$ is singular with respect to $\mathcal{L}_N$.

For a function $v \in L^1(\Omega)$ one denotes by S_v the complement of the Lebesgue set of v, i.e.,

$$S_v := \{x \in \Omega : v^-(x) < v^+(x)\},$$

where

$$v^+(x) := \inf \left\{ t \in \bar{\mathbb{R}} : \lim_{\epsilon \to 0} \frac{\mathcal{L}_N(\{v > t\} \cap B(x,\epsilon))}{\epsilon^N} = 0 \right\}$$

and

$$v^-(x) := \sup \left\{ t \in \bar{\mathbb{R}} : \lim_{\epsilon \to 0} \frac{\mathcal{L}_N(\{v < t\} \cap B(x,\epsilon))}{\epsilon^N} = 0 \right\}.$$

Then S_v is countably rectifiable, and for $\mathcal{H}_{N-1}$-a.e. $x \in \Omega$ we can define the outer normal $\nu(x)$. We denote by $\tilde{v} : \Omega \setminus S_v \to \mathbb{R}$ the approximate limit of v defined as $\tilde{v}(x) = v^+(x) = v^-(x)$.

Following [1, 20] $D_s v$ can be expressed by $D_s v = C_v + J_v$, where $J_v = (v^+ - v^-)\nu \cdot \mathcal{H}_{N-1}|_{S_v}$ is the *jump part* and C_v is the *Cantor part* of Dv. Therefore, we can express the measure Dv by

$$Dv = \nabla v \cdot \mathcal{L}_N + C_v + (v^+ - v^-)\nu \cdot \mathcal{H}_{N-1}|_{S_v}, \tag{2.2}$$

and its total variation by

$$|Dv|(E) = \int_E |\nabla v| dx + \int_{E \setminus S_v} |C_v| + \int_{E \cap S_v} (v^+ - v^-) d\mathcal{H}_{N-1}, \tag{2.3}$$

for every Borel set E in the Borel σ-algebra $\mathcal{B}(\Omega)$ of Ω. For major details we refer the reader to [1]. By these properties of real BV functions, one obtains the following result for vector valued BV functions.

LEMMA 2.1 (Lebesgue decomposition for vector valued BV functions). *For $u \in BV(\Omega;\mathbb{R}^N)$, the positive measure $|Du|$ as defined in* (2.1) *has the following Lebesgue decomposition:*

$$|Du| = |D_a u| + |D_s u|, \tag{2.4}$$

where $|D_a u| = \sum_{i=1}^M |\nabla u_i| \mathcal{L}_N$ is the absolutely continuous part and $|D_s u| = \sum_{i=1}^M |C_{u_i}| + \sum_{i=1}^M (u_i^+ - u_i^-)\mathcal{H}_{N-1}|_{S_{u_i}}$ is the singular part of $|Du|$ with respect to the Lebesgue measure $\mathcal{L}_N$.

Proof. By definition it is $|Du| = \sum_{i=1}^M |Du_i|$ and by the Lebesgue decomposition (2.3) for each $|Du_i|$ it is $|Du| = \sum_{i=1}^M \left(|\nabla u_i| \mathcal{L}_N + |C_{u_i}| + (u_i^+ - u_i^-)\mathcal{H}_{N-1}|_{S_{u_i}} \right)$. Since $\sum_{i=1}^M |\nabla u_i| \mathcal{L}_N$ is absolutely continuous and $\sum_{i=1}^M |C_{u_i}| + \sum_{i=1}^M (u^+ - u^-)\mathcal{H}_{N-1}|_{S_{u_i}}$ is singular with respect to $\mathcal{L}_N$, one concludes the proof by the uniqueness of the Lebesgue decomposition. □

3. Convex functions and functionals of measures. In the following and throughout the paper we assume that

(A) $\phi : \mathbb{R} \to \mathbb{R}_+$ is an even and convex function, nondecreasing in $\mathbb{R}_+$, such that the following hold:

(i) $\phi(0) = 0$;

(ii) there exist $c > 0$ and $b \geq 0$ such that $cz - b \leq \phi(z) \leq cz + b$ for all $z \in \mathbb{R}$.

Under such conditions the asymptotic recession function ϕ^∞ defined by

$$\phi^\infty(z) := \lim_{y \to \infty} \frac{\phi(yz)}{y}$$

is well defined and bounded. It is $c = \lim_{y\to\infty} \frac{\phi(y)}{y} = \phi^\infty(1)$ and $\phi^\infty(z) = cz \cdot \operatorname{sign}(z)$.

Following [17, 25] we can define convex functions of measures. In particular, if $\mu \in \mathcal{M}(\Omega)$, then we can define

$$\phi(|\mu|) = \phi(|\mu_a|)\mathcal{L}_N + \phi^\infty(1)|\mu_s|,$$

where μ_a and μ_s are the absolutely continuous and singular parts of μ, respectively, with respect to $\mathcal{L}_N$. Therefore, according to Lemma 2.1, if $u \in BV(\Omega; \mathbb{R}^M)$, then

$$\begin{aligned} &\sum_{i=1}^M \phi(|Du_i|) \\ &= \sum_{i=1}^M \phi(|\nabla u_i|)\mathcal{L}_N + \phi^\infty(1) \left(\sum_{i=1}^M |C_{u_i}| + \sum_{i=1}^M (u^+ - u^-)\mathcal{H}_{N-1}|_{S_{u_i}} \right). \end{aligned} \tag{3.1}$$

Definition 3.1. *Let (X, τ) be a topological space satisfying the first axiom of countability and $F : X \to \bar{\mathbb{R}}$. The relaxed functional of F with respect to the topology τ is defined for every $x \in X$ as $\bar{F}(x) := \sup\{G(x) : G$ is τ-lower semicontinuous and $G \leq F\}$. In other words $\bar{F}$ is the maximal τ-lower semicontinuous functional that is smaller than F. We may also write*

$$\bar{F}(u) = \inf_{u^{(n)} \in X,\, u^{(n)} \xrightarrow{\tau} u} \left\{ \liminf_n F(u^{(n)}) \right\}.$$

We have the following result.

Lemma 3.2. *If $u \in BV(\Omega; \mathbb{R}^M)$ and ϕ is as in assumption* (A), *then*

$$\begin{aligned} E(u) &:= \int_\Omega \sum_{i=1}^M \phi(|Du_i|) := \sum_{i=1}^M \phi(|Du_i|)(\Omega) \\ &= \int_\Omega \sum_{i=1}^M \phi(|\nabla u_i|)dx + c \left(\sum_{i=1}^M \int_{\Omega \setminus S_{u_i}} |C_{u_i}| + \int_{S_{u_i}} (u_i^+ - u_i^-) d\mathcal{H}_{N-1} \right) \end{aligned}$$

is lower semicontinuous with respect to the componentwise BV weak-$*$-topology.

Proof. It is known that $u_i \to E_i(u_i) := \int_\Omega \phi(|\nabla u_i|)dx + c(\int_{\Omega\setminus S_{u_i}} |C_{u_i}| + \int_{S_{u_i}} (u_i^+ - u_i^-)d\mathcal{H}_{N-1})$ is lower semicontinuous for the BV weak-$*$-topology on $BV(\Omega)$ [25]. One concludes simply by observing that $E(u) = \sum_{i=1}^M E_i(u_i)$. □

4. Assumptions on the evaluation map $\mathcal{L}$. In the following we assume that

(L1) $\mathcal{L} : \mathbb{R}^M \to \mathbb{R}_+$ is a nondecreasing continuous function in the sense that $\mathcal{L}(x) \leq \mathcal{L}(y)$ for any $x, y \in \mathbb{R}^M$ such that $|x_i| \leq |y_i|$ for any $i \in \{1, \ldots, M\}$;

(L2) $\mathcal{L}(x) \leq a + b|x|^s$ for all $x \in \mathbb{R}^M$ and for fixed $s \geq p^{-1}$, $b > 0$, and $a \geq 0$.

Moreover, one of the two following conditions holds:

(L3-a) $\lim_{x\to\infty} \mathcal{L}(x) = +\infty$;

(L3-b) $\mathcal{L}(x) = \mathcal{L}(x_1, \dots, x_M) = \mathcal{L}((\ell_1 \wedge x_1 \vee -\ell_1), \dots, (\ell_M \wedge x_M \vee -\ell_M))$ for a suitable fixed vector $\ell = (\ell_1, \dots, \ell_M) \in \mathbb{R}^M_+$.

Observe that condition (L3-a) is equivalent to saying that for every $C > 0$ the set $\{\mathcal{L} \le C\}$ is bounded. Therefore, there exists $A \in \mathbb{R}^M$, with $A_i \ge 0$ for any $i \in \{1, \dots, M\}$, such that $\{\mathcal{L} \le C\} \subseteq \prod_{i=1}^M [-A_i, A_i]$.

In the following and throughout the paper D denotes a measurable subset of Ω, and we are given the couple $(\bar{u}, \bar{v})$ of bounded functions such that $\bar{u} : \Omega \setminus D \to \mathbb{R}^M$ and $\bar{v} : D \to \mathbb{R}$.

If condition (L3-a) holds, for any measurable function $u : \Omega \to \mathbb{R}^M$, we define the *truncation or clipping operator* as follows:

$$\text{tr}(u, \bar{u}, \Omega, D)(x) := ((\|\bar{u}_i|\Omega \setminus D\|_\infty \vee A_i) \wedge u_i(x) \vee (-\|\bar{u}_i|\Omega \setminus D\|_\infty \wedge -A_i))_{i=1}^M, \tag{4.1}$$

where $A \in \mathbb{R}^M$ is determined so that $\{\mathcal{L} \le \|\bar{v}|D\|_\infty\} \subseteq \prod_{i=1}^M [-A_i, A_i]$. Analogously we define the truncation operator in the case of condition (L3-b):

$$\text{tr}(u, \bar{u}, \bar{v}, \Omega, D)(x) := ((\|\bar{u}_i|\Omega \setminus D\|_\infty \vee \ell_i) \wedge u_i(x) \vee (-\|\bar{u}_i|\Omega \setminus D\|_\infty \wedge -\ell_i))_{i=1}^M. \tag{4.2}$$

In the case when it is clear which of the conditions (L3-a,b) holds and the set D and the functions $\bar{u}, \bar{v}$ are given, then it will be convenient to use the shorter notation $\hat{u} := \text{tr}(u, \bar{u}, \bar{v}, \Omega, D)$.

For any measurable function $u : \Omega \to \mathbb{R}^M$ we define

$$G_1(u) = \int_{\Omega \setminus D} |u(x) - \bar{u}(x)|^p dx, \tag{4.3}$$

$$G_2(u) = \int_D |\mathcal{L}(u(x)) - \bar{v}(x)|^p dx. \tag{4.4}$$

LEMMA 4.1. *For any $u \in BV(\Omega; \mathbb{R}^M)$ the truncation operator has the property that $\hat{u} \in BV(\Omega; \mathbb{R}^M)$, and*

$$G_i(\hat{u}) \le G_i(u), \quad i = 1, 2, \quad \textit{and } E(\hat{u}) \le E(u). \tag{4.5}$$

Proof. Let us assume that condition (L3-a) holds. If $x \in \Omega \setminus D$, the definition of the truncation operator implies that $|\hat{u}(x) - \bar{u}(x)| \le |u(x) - \bar{u}(x)|$, from which it follows that $G_1(\hat{u}) \le G_1(u)$. If $x \in D$ is such that $u(x) \in \prod_{i=1}^M [-\|\bar{u}_i|\Omega \setminus D\|_\infty \wedge -A_i, \|\bar{u}_i|\Omega \setminus D\|_\infty \vee A_i]$, then $\hat{u}(x) = u(x)$. Otherwise, $x \notin \prod_{i=1}^M [-A_i, A_i]$ and $|u_i(x)| \ge |\hat{u}_i(x)| \ge |\xi_i|$ for any ξ such that $\mathcal{L}(\xi) \le \|\bar{v}|D\|_\infty$ and any $i \in \{1, \dots, M\}$. Therefore, by the monotonicity assumption (L1) $\mathcal{L}(u(x)) \ge \mathcal{L}(\hat{u}(x)) \ge \|\bar{v}|D\|_\infty$, which implies that $|\mathcal{L}(\hat{u}(x)) - \bar{v}(x)| \le |\mathcal{L}(u(x)) - \bar{v}(x)|$ for any $x \in D$, and $G_2(\hat{u}) \le G_2(u)$. The proof is analogous if condition (L3-b) holds.

We now prove the corresponding statement for the functional E. Fix $i \in \{1, \dots, M\}$. By definition of the truncation operator, we have $\hat{u}_i = g_i \circ u_i$, where $g_i : \mathbb{R} \to \mathbb{R}$ is a Lipschitz function such that

$$g_i(t) = \begin{cases} t, & -c_i \le t \le d_i, \\ d_i, & t > d_i, \\ -c_i, & t < -c_i, \end{cases}$$

where $c_i, d_i > 0$ are determined by (4.1), (4.2). Using the chain rule for real valued BV functions (Theorem 3.99 of [2]), we have that $\hat{u} \in BV(\Omega; \mathbb{R}^M)$ and

$$D\hat{u}_i = g_i'(u_i) \nabla u_i \cdot \mathcal{L}_N + g_i'(\tilde{u}_i) C_{u_i} + \left(g_i(u_i^+) - g_i(u_i^-)\right) \nu_i \cdot \mathcal{H}_{N-1}|_{S_{u_i}},$$

where $\tilde{u}_i$ is the approximate limit of u_i. Then $\nabla\hat{u}_i(x) = \nabla u_i(x)$ if $-c_i < u_i(x) < d_i$, and $\nabla\hat{u}_i(x) = 0$ if either $u_i(x) > d_i$ or $u_i(x) < -c_i$. Moreover, by Proposition 3.73(c) of [2] it follows that $\nabla u_i(x) = 0$ for a.e. $x \in \{u_i(x) = d_i\}$ and a.e. $x \in \{u_i(x) = -c_i\}$. Hence $|\nabla\hat{u}_i(x)| \leq |\nabla u_i(x)|$ a.e., so that from assumption (A) of the function ϕ we get

$$\int_\Omega \phi(|\nabla\hat{u}_i|)dx \leq \int_\Omega \phi(|\nabla u_i|)dx. \tag{4.6}$$

Since $u_i^+(x) \geq u_i^-(x)$ for any $x \in S_{u_i}$, by the definition of the function g_i we have

$$S_{\hat{u}_i} \subseteq S_{u_i}, \qquad g_i(u_i^+(x)) - g_i(u_i^-(x)) \leq u_i^+(x) - u_i^-(x) \quad \text{for any } x \in S_{u_i}.$$

Then it follows that

$$\int_{S_{\hat{u}_i}} (\hat{u}_i^+ - \hat{u}_i^-)d\mathcal{H}_{N-1} \leq \int_{S_{u_i}} (u_i^+ - u_i^-)d\mathcal{H}_{N-1}. \tag{4.7}$$

By the definition of g_i we then have $0 \leq g_i'(\tilde{u}_i(x)) \leq 1$ for any $x \in \{x : \tilde{u}_i(x) \neq d_i\} \cap \{x : \tilde{u}_i(x) \neq -c_i\}$. Moreover, by Proposition 3.92(c) of [2], the Cantor part C_{u_i} vanishes on sets of the form $\tilde{u}_i^{-1}(Q)$ with $Q \subset \mathbb{R}$, $\mathcal{H}_1(Q) = 0$. It follows that C_{u_i} vanishes on the set $\{x : \tilde{u}_i(x) = d_i\} \cup \{x : \tilde{u}_i(x) = -c_i\}$, so that we get $|C_{\hat{u}_i}|(\Omega) \leq |C_{u_i}|(\Omega)$, i.e.,

$$\int_{\Omega\setminus S_{\hat{u}_i}} |C_{\hat{u}_i}| \leq \int_{\Omega\setminus S_{u_i}} |C_{u_i}|. \tag{4.8}$$

Collecting the inequalities (4.6)–(4.8) and summing over $i = 1, \dots, M$, we obtain $E(\hat{u}) \leq E(u)$, which concludes the proof. □

REMARK 4.2. *The truncation operator maps C_0^1 functions into $W^{1,q}$; i.e., for any $u \in C_0^1(\Omega;\mathbb{R}^M)$ we have $\operatorname{tr}(u, \bar{u}, \bar{v}, \Omega, D) \in W^{1,q}(\Omega;\mathbb{R}^M)$ for any $1 \leq q \leq \infty$.*

5. Relaxation and existence of minimizers.

The functional F is well defined in $L^\infty(\Omega;\mathbb{R}^M) \cap W^{1,1}(\Omega;\mathbb{R}^M)$. Since this space is not reflexive, and sequences that are bounded in $W^{1,1}$ are also bounded in BV, we extend F to the space $BV(\Omega;\mathbb{R}^M)$ in such a way that the extended functional is lower semicontinuous. By using the relaxation method of the calculus of variations, the natural candidate for the extended functional is the relaxed functional $\bar{F}$ of F with respect to the componentwise BV weak-$*$-topology [6].

In the following, without loss of generality, we set $\mu = \lambda = 1$.

5.1. Relaxation.

We set $X = \{u \in BV(\Omega;\mathbb{R}^M) : \|u_i\|_\infty \leq K_i,\ i = 1, \dots, M\}$, where, for any $i \in \{1, \dots, M\}$, the constant $K_i > 0$ is defined by $K_i = \max\{A_i, \|\bar{u}_i|\Omega\setminus D\|_\infty\}$ if condition (L3-a) holds, and by $K_i = \max\{\ell_i, \|\bar{u}_i|\Omega\setminus D\|_\infty\}$ if condition (L3-b) holds.

The following theorem extends to our case the relaxation result proved in [6, Theorem 3.2.1].

THEOREM 5.1. *The relaxed functional of F in X with respect to the componentwise BV weak-$*$-topology is given by*

$$\bar{F}(u) = \int_{\Omega\setminus D} |u(x) - \bar{u}(x)|^p dx + \int_D |\mathcal{L}(u(x)) - \bar{v}(x)|^p dx$$
$$+ \int_\Omega \sum_{i=1}^M \phi(|\nabla u_i|)dx + c\left(\sum_{i=1}^M \int_{\Omega\setminus S_{u_i}} |C_{u_i}| + \int_{S_{u_i}} (u_i^+ - u_i^-)d\mathcal{H}_{N-1}\right).$$

Proof. Let us define

$$f(u) := \begin{cases} F(u), & u \in X \cap W^{1,1}(\Omega;\mathbb{R}^M), \\ +\infty, & u \in X \setminus W^{1,1}(\Omega;\mathbb{R}^M). \end{cases}$$

Observe that $f(u) = \bar{F}(u)$ for $u \in W^{1,1}(\Omega;\mathbb{R}^M)$.

By property (L2) we have that $G_1(u), G_2(u) < +\infty$ for all $u \in X$. By using Fatou's lemma the functionals G_1 and G_2 are lower semicontinuous with respect to the strong L^1 topology and hence with respect to the componentwise BV weak-$*$-topology. Therefore, by Lemma 3.2, $\bar{F}$ is lower semicontinuous in X with respect to such topology.

Let $\bar{f}$ denote the relaxed functional of f in X with respect to the same topology. Since $\bar{F}(u) \leq f(u)$ for any $u \in X$, and $\bar{f}$ is the greatest lower semicontinuous functional less than or equal to f, we have $\bar{f}(u) \geq \bar{F}(u)$ for any $u \in X$. Then we have to show that $\bar{f}(u) \leq \bar{F}(u)$.

By [17, Theorems 2.2 and 2.3] for any $u \in X$ there exists a sequence $\{u^{(n)}\}_n \subset C_0^\infty(\Omega;\mathbb{R}^M) \cap W^{1,1}(\Omega;\mathbb{R}^M)$ such that $u^{(n)}$ converges to u in the componentwise BV weak-$*$-topology and $E(u) = \lim_n E(u^{(n)})$.

Let us now consider the sequence $\{\hat{u}^{(n)}\}_n$ of the truncated functions. By Lemma 4.1 we have

$$E(u) = \lim_n E(u^{(n)}) \geq \limsup_n E(\hat{u}^{(n)}). \tag{5.1}$$

With similar computations as those in the proof of Lemma 4.1

$$\int_\Omega |\hat{u}^{(n)}(x) - u(x)| dx \leq \int_\Omega |u^{(n)}(x) - u(x)| dx \to 0, \quad n \to \infty.$$

Moreover, since the truncated functions $\hat{u}^{(n)}$ are uniformly bounded in $L^\infty(\Omega;\mathbb{R}^M)$, then $\hat{u}^{(n)}$ converges to u in $L^q(\Omega;\mathbb{R}^M)$ for any $1 \leq q < \infty$.

Now the functional G_1 is continuous with respect to the strong $L^p(\Omega \setminus D;\mathbb{R}^M)$ topology. Moreover, since $\mathcal{L}$ is continuous, the functional G_2 is continuous with respect to the strong $L^q(D;\mathbb{R}^M)$ topology, with $q = sp \geq 1$ (see [19, Chapter 9, Lemma 3.2]).

Then, using (5.1), the continuity properties of G_1 and G_2, and Remark 4.2, we have $\hat{u}^{(n)} \in W^{1,1}(\Omega;\mathbb{R}^M)$, $\bar{F}(\hat{u}^{(n)}) = f(\hat{u}^{(n)})$, and

$$\begin{aligned} \bar{F}(u) = G_1(u) + G_2(u) + E(u) &\geq \lim_n (G_1(\hat{u}^{(n)}) + G_2(\hat{u}^{(n)})) + \limsup_n E(\hat{u}^{(n)}) \\ &\geq \limsup_n f(\hat{u}^{(n)}) \geq \liminf_n f(\hat{u}^{(n)}) \geq \inf_{u^{(n)} \in BV,\, u^{(n)} \overset{BV-w^*}{\to} u} \left\{ \liminf_n f(u^{(n)}) \right\} = \bar{f}(u). \end{aligned}$$

Then we have $\bar{F}(u) = \bar{f}(u)$ and the statement is proved. □

5.2. Existence and uniqueness of minimizers. In this section we shall prove the existence of minimizers of $\bar{F}$ in X and state the conditions for the uniqueness.

THEOREM 5.2. *There exists a solution of the following variational problem:*

$$\min_{u \in X} \left\{ \bar{F}(u) = \int_{\Omega \setminus D} |u(x) - \bar{u}(x)|^p dx + \int_D |\mathcal{L}(u(x)) - \bar{v}(x)|^p dx \right. \\ \left. + \int_\Omega \sum_{i=1}^M \phi(|\nabla u_i|) dx + c \left(\sum_{i=1}^M \int_{\Omega \setminus S_{u_i}} |C_{u_i}| + \int_{S_{u_i}} (u_i^+ - u_i^-) d\mathcal{H}_{N-1} \right) \right\}.$$

In particular, we have

$$\min_{u\in X} \bar{F}(u) = \inf_{u\in X} F(u).$$

Moreover, if $D \subsetneq \Omega$ and G_2 is a strictly convex functional, then the solution is unique.

Proof. Let $\{u^{(n)}\}_n$ be a minimizing sequence in BV. By assumption (A)(ii) in section 3, there exists a constant $C > 0$ such that $|Du^{(n)}|(\Omega) \leq C$ uniformly with respect to n. By Lemma 4.1 we can modify the minimizing sequence by truncation, obtaining a new minimizing sequence $\{\hat{u}^{(n)}\}_n \subset X$. By Lemma 4.1 this sequence is uniformly bounded in $BV(\Omega;\mathbb{R}^M)$, i.e.,

$$\|\hat{u}^{(n)}\|_\infty \leq \max_{i=1,\dots,M} K_i, \qquad |D\hat{u}^{(n)}|(\Omega) \leq C$$

for any n. Therefore, there exists a subsequence $\{\hat{u}^{(n_k)}\}_k$ converging with respect to the componentwise BV weak-$*$-topology to a function $u \in X$. Since the relaxed functional $\bar{F}$ is lower semicontinuous in X with respect to such a topology, we have

$$\bar{F}(u) \leq \liminf_k \bar{F}(u^{(n_k)}).$$

From the compactness and lower semicontinuity properties of $\bar{F}$ it follows that $u \in X$ is a minimizer of $\bar{F}$. Moreover, if $D \subsetneq \Omega$ and G_2 is a strictly convex functional, then $\bar{F}$ is strictly convex and the solution u is unique. Since F is coercive in X, one concludes by an application of [27, Theorem 3.8]. □

6. Approximation by Γ-convergence.

In this section we endow the space X with the L^1 strong topology, and we show that minimizers of $\bar{F}$ can be approximated in X by minimum points of functionals that are defined in $W^{1,2}(\Omega;\mathbb{R}^M)$.

For a positive decreasing sequence $\{\varepsilon_h\}_{h\in\mathbb{N}}$ such that $\lim_{h\to\infty}\varepsilon_h = 0$, and for $\phi \in C^1(\mathbb{R})$, we define

$$(6.1)\quad F_h(u) = \begin{cases} G_1(u) + G_2(u) + \displaystyle\int_\Omega \sum_{i=1}^M \phi_h(|\nabla u_i(x)|)dx, & u \in W^{1,2}(\Omega;\mathbb{R}^M), \\ +\infty, & u \in X \setminus W^{1,2}(\Omega;\mathbb{R}^M), \end{cases}$$

where

$$\phi_h(z) = \begin{cases} \dfrac{\phi'(\varepsilon_h)}{2\varepsilon_h} z^2 + \phi(\varepsilon_h) - \dfrac{\varepsilon_h \phi'(\varepsilon_h)}{2}, & 0 \leq z \leq \varepsilon_h, \\ \phi(z), & \varepsilon_h \leq z \leq \dfrac{1}{\varepsilon_h}, \\ \dfrac{\varepsilon_h \phi'(1/\varepsilon_h)}{2} z^2 + \phi\left(\dfrac{1}{\varepsilon_h}\right) - \dfrac{\phi'(1/\varepsilon_h)}{2\varepsilon_h}, & z \geq \dfrac{1}{\varepsilon_h}. \end{cases}$$

If $z \mapsto \frac{\phi'(z)}{z}$ is continuously decreasing, then $\phi_h(z) \geq \phi(z) \geq 0$ for any h and any z, and $\lim_h \phi_h(z) = \phi(z)$ for any z.

By means of standard arguments we have that for any h the functional F_h has a minimizer in $X \cap W^{1,2}(\Omega;\mathbb{R}^M)$; see, e.g., [31, Proposition 6.1]. Moreover, if $D \subsetneq \Omega$ and G_2 is a strictly convex functional, then the minimizer is unique. The following theorem extends to our case the Γ-convergence result proved in [31, Proposition 6.1]; see also Theorem 3.2.3 of [6]. We do not introduce the concept of Γ-convergence which is used here only as an auxiliary tool. We refer the reader to [27] and the relevant results therein for more details, in particular, [27, Proposition 5.7, Theorem 7.8, Corollary 7.20, Corollary 7.24].

THEOREM 6.1. *Let $\{u^{(h)}\}_h$ be a sequence of minimizers of F_h. Then $\{u^{(h)}\}_h$ is relatively compact in $L^1(\Omega;\mathbb{R}^M)$, each of its limit points minimizes the functional $\bar{F}$, and*

$$\min_{u\in X}\bar{F}(u)=\lim_{h\to\infty}\min_{u\in X\cap W^{1,2}}F_h(u).$$

Moreover, if $D\subsetneq\Omega$ and G_2 is a strictly convex functional, we have

$$\lim_{h\to\infty}u^{(h)}=u^{(\infty)}\ \textit{in } X,\qquad \lim_{h\to\infty}F_h(u^{(h)})=\bar{F}(u^{(\infty)}),\tag{6.2}$$

where $u^{(\infty)}$ is the unique minimizer of $\bar{F}$ in X.

Proof. We define

$$g(u)=\begin{cases} F(u), & u\in X\cap W^{1,2}(\Omega;\mathbb{R}^M),\\ +\infty, & u\in X\setminus W^{1,2}(\Omega;\mathbb{R}^M).\end{cases}$$

Observe that g is the restriction of F to functions $u\in W^{1,2}(\Omega;\mathbb{R}^M)$.

By construction we have that $\{F_h\}_h$ is a decreasing sequence of functionals that converges pointwise to g in $X\cap W^{1,2}(\Omega;\mathbb{R}^M)$. Therefore, by [27, Proposition 5.7], F_h Γ-converges to the relaxed functional $\bar{g}$ of g in X with respect to the $L^1(\Omega;\mathbb{R}^M)$ topology. Then we have to show that $\bar{F}=\bar{g}$.

Let $\{u^{(n)}\}_n\subset X$ be a sequence such that $u^{(n)}\to u$ in $L^1(\Omega;\mathbb{R}^M)$ and $\liminf_n \bar{F}(u^{(n)})<+\infty$. Up to the extraction of a subsequence we may assume that $\liminf_n \bar{F}(u^{(n)})=\lim_n\bar{F}(u^{(n)})$. Then $\bar{F}(u^{(n)})$ is uniformly bounded with respect to n, so that $\{u^{(n)}\}_n$ is uniformly bounded in BV. Then, up to a subsequence, $u^{(n)}$ converges to u in the componentwise BV weak-$*$-topology and, by Theorem 5.1, we have $\liminf_n\bar{F}(u^{(n)})\geq\bar{F}(u)$. Hence $\bar{F}$ is lower semicontinuous in X with respect to the $L^1(\Omega;\mathbb{R}^M)$ topology.

Then, arguing as in the proof of Theorem 5.1, for any function $u\in X$ there exists a sequence of truncated functions $\hat{u}^{(n)}\in W^{1,2}(\Omega;\mathbb{R}^M)\cap X$ such that

$$\hat{u}^{(n)}\to u \text{ in } L^1(\Omega;\mathbb{R}^M)\quad\text{and}\quad \bar{F}(u)\geq\liminf_{n\to\infty}g(\hat{u}^{(n)}).\tag{6.3}$$

Since $g\geq\bar{F}$, property (6.3) implies that $\bar{F}\geq\bar{g}$. Then, by the lower semicontinuity of $\bar{F}$ with respect to the $L^1(\Omega;\mathbb{R}^M)$ topology, we have $\bar{F}=\bar{g}$. Therefore, F_h Γ-converges to $\bar{F}$.

By construction $\phi_h(z)\geq\phi(z)$ for any $z\geq 0$, so that $F_h(u)\geq\bar{F}(u)$ for any h and any $u\in X$. Since $\bar{F}$ is coercive and lower semicontinuous in $L^1(\Omega;\mathbb{R}^M)$, it follows that the sequence $\{F_h\}_h$ is equicoercive in $L^1(\Omega;\mathbb{R}^M)$. In particular, any family $\{u^{(h)}\}_h$ of minimizers of F_h is relatively compact in $L^1(\Omega;\mathbb{R}^M)$. Then, using [27, Theorem 7.8], the limit points of sequences of minimizers of F_h minimize $\bar{F}$ and $\min_{u\in X}\bar{F}(u)=\lim_h\min_{u\in W^{1,2}}F_h(u)$.

Finally, if $D\subsetneq\Omega$ and G_2 is a strictly convex functional, by Theorem 5.2 there exists a unique minimizer of $\bar{F}$ in X. Therefore the limits (6.2) follow from Corollary 7.24 of [27]. □

REMARK 6.2. *So far we have considered evaluation maps $\mathcal{L}:\mathbb{R}^M\to\mathbb{R}$. However, the whole analysis can be generalized to the case $\mathcal{L}:\mathbb{R}^M\to\mathcal{M}$, $\mathcal{L}(x)=$*

$(\mathcal{L}_1(x),\dots,\mathcal{L}_D(x))$, where $\mathcal{M}\subset\mathbb{R}^M$ is a $(D\le M)$-dimensional submanifold. However, for $D=1$ and L usually being an invertible map, it is possible to "reequalize" the gray level so that $\mathcal{L}(x)=\frac{1}{M}(x_1+\cdots+x_M)$. Later in this paper, for simplicity purposes in numerical implementation, we will use such linearization for $\mathcal{L}$.

7. Euler–Lagrange equations and a relaxation algorithm. In this section we want to provide an algorithm to compute efficiently minimizers of the approximating functionals F_h. First, we want to derive the Euler–Lagrange equations associated with F_h. In the following we assume that both ϕ_h and $\mathcal{L}$ are continuously differentiable and that Ω is an open, bounded, and connected subset of $\mathbb{R}^N$ with Lipschitz boundary $\partial\Omega$. Moreover, $p=2$ if $N=1$ and $p=\frac{N}{N-1}$ for $N>1$, $1/p+1/p'=1$. By standard arguments we have the following result.

Proposition 7.1. *If u is a minimizer in $W^{1,2}(\Omega;\mathbb{R}^M)$ of F_h, then u solves the following system of Euler–Lagrange equations:*

$$\begin{cases} 0=-\operatorname{div}\left(\frac{\phi_h'(|\nabla u_i|)}{|\nabla u_i|}\nabla u_i\right)+p|u-\bar{u}|^{p-2}(u_i-\bar{u}_i)1_{\Omega\setminus D}+p|\mathcal{L}(u)\\ \qquad -\bar{v}|^{p-2}(\mathcal{L}(u)-\bar{v})\frac{\partial\mathcal{L}}{\partial u_i}(u)1_D,\\ \frac{\phi_h'(|\nabla u_i|)}{|\nabla u_i|}\frac{\partial u_i}{\partial\nu}=0 \text{ on } \partial\Omega,\quad i=1,\dots,M.\end{cases}$$

The former equalities hold in the sense of distributions and in $L^{p'}(\Omega;\mathbb{R}^M)$.

The previous equations yield a necessary condition for the computation of minimizers of F_h. Again we are not ensured of the uniqueness in general, unless G_2 is strictly convex. The system is composed of M second order *nonlinear* equations which are coupled on terms of order 0. Both the nonlinear term $\operatorname{div}\left(\frac{\phi_h'(|\nabla u_i|)}{|\nabla u_i|}\nabla u_i\right)$ and the coupled terms of order 0 constitute a complication for the numerical solution of these equations.

Based on the work [12, 18, 32], we propose in the following a method to compute efficiently solutions of the Euler–Lagrange equations, which simplifies the problem of the nonlinearity. Since we want to illustrate concrete applications for color image recovery, for simplicity, we limit our analysis to the case $N=p=2$ and $\phi(t)=|t|$ for all $t\in\mathbb{R}$. Let us introduce a new functional given by

$$\mathcal{E}_h(u,w):=2\left(G_1(u)+G_2(u)\right)+\int_\Omega\sum_{i=1}^M\left(w_i|\nabla u_i(x)|^2+\frac{1}{w_i}\right)dx, \tag{7.1}$$

where $u\in W^{1,2}(\Omega;\mathbb{R}^M)$, and $w\in L^2(\Omega;\mathbb{R}^M)$ is such that $\varepsilon_h\le w_i\le\frac{1}{\varepsilon_h}$, $i=1,\dots,M$. While the variable u again is the function to be reconstructed, we call the variable w the *gradient weight*. In the following, since we assume h fixed, we drop the index h from the functional $\mathcal{E}_h$.

For any given $u^{(0)}\in X\cap W^{1,2}(\Omega;\mathbb{R}^M)$ and $w^{(0)}\in L^2(\Omega;\mathbb{R}^M)$ (for example, $w^{(0)}:=1$), we define the following iterative double-minimization algorithm:

$$\begin{cases} u^{(n+1)}=\arg\min\limits_{u\in W^{1,2}(\Omega;\mathbb{R}^M)}\mathcal{E}(u,w^{(n)}),\\ w^{(n+1)}=\arg\min\limits_{\varepsilon_h\le w\le\frac{1}{\varepsilon_h}}\mathcal{E}(u^{(n+1)},w).\end{cases} \tag{7.2}$$

We have the following convergence result.

Theorem 7.2. *The sequence $\{u^{(n)}\}_{n\in\mathbb{N}}$ has subsequences that converge strongly in $L^2(\Omega;\mathbb{R}^M)$ and weakly in $W^{1,2}(\Omega;\mathbb{R}^M)$ to a stationary point $u^{(\infty)}$ of F_h; i.e., $u^{(\infty)}$*

solves the Euler–Lagrange equations in Proposition 7.1. *Moreover, if F_h has a unique minimizer u^*, then $u^{(\infty)} = u^*$ and the full sequence $\{u^{(n)}\}_{n\in\mathbb{N}}$ converges to u^*.*

Proof. Observe that

$$\begin{aligned}\mathcal{E}(u^{(n)}, w^{(n)}) - \mathcal{E}(u^{(n+1)}, w^{(n+1)}) &= \underbrace{\Big(\mathcal{E}(u^{(n)}, w^{(n)}) - \mathcal{E}(u^{(n+1)}, w^{(n)})\Big)}_{A_n} \\ &\quad + \underbrace{\Big(\mathcal{E}(u^{(n+1)}, w^{(n)}) - \mathcal{E}(u^{(n+1)}, w^{(n+1)})\Big)}_{B_n} \geq 0.\end{aligned}$$

Therefore, $\mathcal{E}(u^{(n)}, w^{(n)})$ is a nonincreasing sequence and, moreover, it is bounded from below, since

$$\inf_{\varepsilon_h \leq w \leq 1/\varepsilon_h} \int_\Omega \sum_{i=1}^M \left(w_i |\nabla u_i(x)|^2 + \frac{1}{w_i} \right) dx \geq 0.$$

This implies that $\mathcal{E}(u^{(n)}, w^{(n)})$ converges. Moreover, we can write

$$B_n = \int_\Omega \sum_{i=1}^M c(w_i^{(n)}(x), |\nabla u_i^{(n+1)}(x)|) - c(w_i^{(n+1)}(x), |\nabla u_i^{(n+1)}(x)|) dx,$$

where $c(t, z) := tz^2 + \frac{1}{t}$. By Taylor's formula, we have

$$c(w_i^{(n)}, z) = c(w_i^{(n+1)}, z) + \frac{\partial c}{\partial t}(w_i^{(n+1)}, z)(w_i^{(n)} - w_i^{(n+1)}) + \frac{1}{2}\frac{\partial^2 c}{\partial t^2}(\xi, z)|w_i^{(n)} - w_i^{(n+1)}|^2$$

for $\xi \in \operatorname{conv}(w_i^{(n)}, w_i^{(n+1)})$. By definition of $w_i^{(n+1)}$ and taking into account that $\varepsilon_h \leq w_i^{(n+1)} \leq \frac{1}{\varepsilon_h}$, we have

$$\frac{\partial c}{\partial t}(w_i^{(n+1)}, |\nabla u_i^{(n+1)}(x)|)(w_i^{(n)} - w_i^{(n+1)}) \geq 0,$$

and $\frac{\partial^2 c}{\partial t^2}(t, z) = \frac{2}{t^3} \geq 2\varepsilon_h^3$, for any $t \leq 1/\varepsilon_h$. This implies that

$$\mathcal{E}(u^{(n)}, w^{(n)}) - \mathcal{E}(u^{(n+1)}, w^{(n+1)}) \geq B_n \geq \varepsilon_h^3 \int_\Omega \sum_{i=1}^M |w_i^{(n)}(x) - w_i^{(n+1)}(x)|^2 dx,$$

and since $\mathcal{E}(u^{(n)}, w^{(n)})$ is convergent, we have $\sum_{i=1}^M \int_\Omega |w_i^{(n)}(x) - w_i^{(n+1)}(x)|^2 dx \to 0$ for $n \to \infty$. In fact it holds that

$$\|w_i^{(n)} - w_i^{(n+1)}\|_{L^q} \to 0, \quad i = 1, \ldots, M, \tag{7.3}$$

for $n \to \infty$ and for any $1 \leq q < \infty$. Since $u^{(n+1)}$ is a minimizer of $\mathcal{E}(u, w^{(n)})$, it solves the following system of variational equations:

$$\begin{aligned}\int_\Omega \Big(& w_i^{(n)} \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2(u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ & + 2(\mathcal{L}(u^{(n+1)}(x)) - \bar{v}(x)) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}(x)) 1_D(x) \Big) \varphi_i(x) dx = 0 \end{aligned} \tag{7.4}$$

for $i = 1, \ldots, M$ and for all $\varphi \in W^{1,2}(\Omega; \mathbb{R}^M)$. Therefore, we can write

$$\begin{aligned}\int_\Omega \Big(& w_i^{(n+1)} \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2(u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ & + 2(\mathcal{L}(u^{(n+1)}(x)) - \bar{v}(x)) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}(x)) 1_D(x) \Big) \varphi_i(x) dx \\ & = \int_\Omega (w_i^{(n+1)} - w_i^{(n)}) \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) dx.\end{aligned}$$

For $\frac{1}{q} + \frac{1}{q'} + \frac{1}{2} = 1$, we have

$$\begin{aligned}\Bigg| \int_\Omega \Big(& w_i^{(n+1)} \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2(u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ & + 2(\mathcal{L}(u^{(n+1)}(x)) - \bar{v}(x)) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}(x)) 1_D(x) \Big) \varphi_i(x) dx \Bigg| \\ & \leq \| w_i^{(n+1)} - w_i^{(n)} \|_{L^q} \| \nabla u_i^{(n+1)} \|_{L^{q'}} \| \nabla \varphi_i \|_{L^2}.\end{aligned}$$

Since $u^{(n+1)}$ is a minimizer of $\mathcal{E}(u, w^{(n)})$, we may assume without loss of generality that $\hat{u}_i^{(n+1)} = u_i^{(n+1)}$ for all $i = 1, \ldots, M$, where $\hat{}$ is the truncation operator. Consequently $\|u_i^{(n+1)}\|_\infty \leq C < +\infty$ uniformly with respect to n. We can use the results in [28] to show that there exists $q' > 2$ such that

$$\| \nabla u_i^{(n+1)} \|_{L^{q'}} \leq C < +\infty$$

uniformly with respect to n (see also [4, 5, 12] for similar arguments). Therefore, using (7.3), we can conclude that

$$- \operatorname{div}(w_i^{(n+1)} \nabla u_i^{(n+1)}) + 2 \left((u_i^{(n+1)} - \bar{u}_i) 1_{\Omega \setminus D} + (\mathcal{L}(u^{(n+1)}) - \bar{v}) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}) 1_D \right) \to 0,$$

for $n \to \infty$, in $(W^{1,2}(\Omega; \mathbb{R}^M))'$. This also shows that $\{u^{(n)}\}_n$ is uniformly bounded in $W^{1,2}(\Omega; \mathbb{R}^M)$. Therefore, there exists a subsequence $\{u^{(n_k)}\}_k$ that converges strongly in L^2 and weakly in $W^{1,2}(\Omega; \mathbb{R}^M)$ to a function $u^{(\infty)} \in W^{1,2}(\Omega; \mathbb{R}^M)$. Since $w_i^{(n+1)} = \frac{\phi_h'(|\nabla u_i^{(n+1)}|)}{|\nabla u_i^{(n+1)}|}$, with standard arguments for monotone operators (see the proof of [12, Proposition 3.1] and [10]), we show that in fact

(7.5)
$$- \operatorname{div} \left(\frac{\phi_h'(|\nabla u_i^{(\infty)}|)}{|\nabla u_i^{(\infty)}|} \nabla u_i^{(\infty)} \right) + 2 \left((u_i^{(\infty)} - \bar{u}_i) 1_{\Omega \setminus D} + (\mathcal{L}(u^{(\infty)}) - \bar{v}) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(\infty)}) 1_D \right) = 0,$$

for $i = 1, \ldots, M$, in $(W^{1,2}(\Omega; \mathbb{R}^M))'$. The latter are the Euler–Lagrange equations associated with the functional F_h, and therefore $u^{(\infty)}$ is a stationary point for F_h.

Assume now that F_h has a unique minimizer u^*. Then necessarily $u^{(\infty)} = u^*$. Since every subsequence of $\{u^{(n)}\}_n$ has a subsequence converging to u^*, the full sequence $\{u^{(n)}\}_n$ converges to u^*. □

Since both F_h and $\mathcal{E}_h(\cdot, w)$ admit minimizers, their uniqueness is equivalent to the uniqueness of the solutions of the corresponding Euler–Lagrange equations. If uniqueness of the solution is satisfied, then the algorithm (7.2) can be reformulated equivalently as the following two-step iterative procedure:

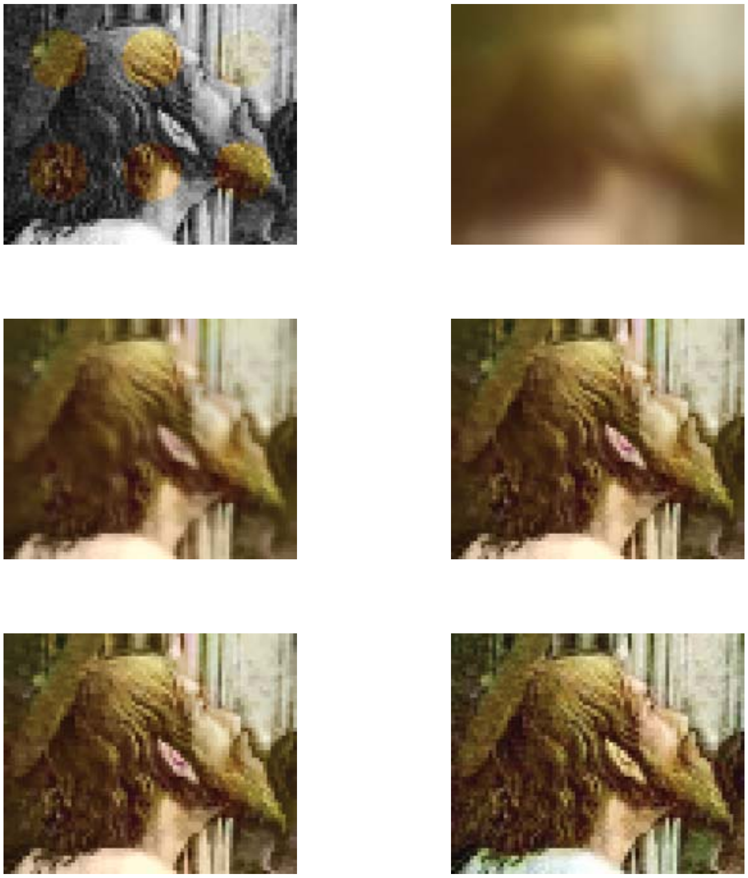


Fig. 7.1. *The datum* $(\bar{u}, \bar{v})$ *is illustrated in the top-left position. The image has dimensions* 64×78 *pixels. The first four iterations of the algorithms are listed from left to right, starting from the first row. The original color image (A. Mantegna's frescoes, photo by Alinari dated to* 1940*) to be reconstructed is illustrated in the bottom-right position. This image serves as a ground truth for the numerical experiments. The parameters we have used are* $\varepsilon_h = 10^{-4}$, $\lambda = \mu = 150$.

- Find $u^{(n+1)}$, which solves

$$\int_\Omega \Big(w_i^{(n)}(x) \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2(u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ + 2(\mathcal{L}(u^{(n+1)}(x)) - \bar{v}(x)) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}(x)) 1_D(x) \Big) \varphi_i(x) dx = 0$$

  for $i = 1, \ldots, M$ and for all $\varphi \in W^{1,2}(\Omega; \mathbb{R}^M)$.
- Compute directly $w^{(n+1)}$ by

$$w_i^{(n+1)} = \varepsilon_h \vee \frac{1}{|\nabla u_i^{(n+1)}|} \wedge \frac{1}{\varepsilon_h}, \quad i = 1, \ldots, M.$$

There are cases for which one can ensure uniqueness of solutions:

1. If G_2 is strictly convex, then the minimizers are unique as well as the solutions of the equations.

2. Modify the equations by again inserting the parameters $\lambda, \mu > 0$:

$$\int_\Omega \Bigg(w_i^{(n)} \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2\mu (u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ + 2\lambda (\mathcal{L}(u^{(n+1)}(x)) - \bar{v}(x)) \frac{\partial \mathcal{L}}{\partial u_i}(u^{(n+1)}(x)) 1_D(x) \Bigg) \varphi_i(x) dx = 0$$

for $i = 1, \ldots, M$ and for all $\varphi \in W^{1,2}(\Omega; \mathbb{R}^M)$. By a standard fixed point argument, it is not difficult to show that for $\mu \sim \lambda \sim \varepsilon_h$ the solution of the previous equations is unique. Unfortunately the condition $\mu \sim \lambda \sim \varepsilon_h$ is acceptable only for those applications where the constraints on the data are weak, for example, when the data are affected by a strong noise.

3. In the following section we illustrate the finite element approximation of the Euler–Lagrange equations. Since we are interested in color image applications, we restrict the numerical experiments to the case $\mathcal{L}(u_1, u_2, u_3) = \frac{1}{3}(u_1 + u_2 + u_3)$. By this choice, the numerical results confirm that the linear systems arising from the finite element discretization are uniquely solvable for a rather large set of possible parameters λ, μ.

8. Numerical implementation and results. In this section we want to present the numerical implementation of the iterative double-minimization algorithm (7.2) for color image restoration. As the second step of the scheme (which amounts to the update of the gradient weight) can be explicitly done once $u^{(n+1)}$ is computed, we are left essentially to provide a numerical implementation of the first step, i.e., the solution of the Euler–Lagrange equations.

8.1. Finite element approximation of the Euler–Lagrange equations. For the solution of the Euler–Lagrange equations we use a finite element approximation. We illustrate the implementation with the concrete aim of the reconstruction of a digital color image supported in $\Omega = [0,1]^2$ from few color fragments supported in $\Omega \setminus D$ and the gray-level information where colors are missing. By the nature of this problem, we can choose a regular triangulation $\mathcal{T}$ of the domain Ω with nodes distributed on a regular grid $\mathcal{N} := \tau \mathbb{Z}^2 \cap \Omega$, corresponding to the pixels of the image. Associated with $\mathcal{T}$ we fix the following finite element spaces:

$$\mathcal{U} = \{u \in C^0(\Omega) : u|T \in \mathbb{P}^1, \, T \in \mathcal{T}\}, \\ \mathcal{V} = \{w \in L^2(\Omega) : w|T \in \mathbb{P}^0, \, T \in \mathcal{T}\}.$$

The space $\mathcal{U}$ induces the finite element space of color images given by

$$U := \{u \in W^{1,2}(\Omega, \mathbb{R}^3) : u_i \in \mathcal{U}, \, i = 1,2,3\}.$$

The space $\mathcal{V}$ induces the finite element space of *gradient weights* given by

$$V := \{w \in L^2(\Omega, \mathbb{R}^3) : w_i \in \mathcal{V}, \, i = 1,2,3\}.$$

In order to avoid the nonlinearity in the coupled terms of order 0, we restrict our functional to the case $\mathcal{L}(u_1, u_2, u_3) = \frac{1}{3}(u_1 + u_2 + u_3)$. For further simplicity we have not considered truncations which in fact are not necessary in practice.

For a given $w^{(n)} \in V$, the first step of our approximation of the double-minimization scheme amounts to the computation of $u^{(n+1)} \in U$, which solves, using (7.4),

$$\int_\Omega \Bigg(w_i^{(n)}(x) \nabla u_i^{(n+1)}(x) \cdot \nabla \varphi_i(x) + 2\mu (u_i^{(n+1)}(x) - \bar{u}_i(x)) 1_{\Omega \setminus D}(x) \\ + \frac{2}{3}\lambda \left(\frac{1}{3}(u_1^{(n+1)}(x) + u_2^{(n+1)}(x) + u_3^{(n+1)}(x)) - \bar{v}(x) \right) 1_D(x) \Bigg) \varphi_i(x) dx = 0 \tag{8.1}$$

for $i = 1, 2, 3$ and for all $\varphi \in U$. To the spaces $\mathcal{U}$ and $\mathcal{V}$ are attached the corresponding nodal bases $\{\varphi_k\}_{k \in \mathcal{N}}$ and $\{\chi_k\}_{k \in \mathcal{N}}$, respectively. Therefore, we also have that

$$U = \left\{ u : u = \left(\sum_{k \in \mathcal{N}} u_{i,k} \varphi_k \right)_{i=1,2,3} \right\}, \quad V = \left\{ w : w = \left(\sum_{k \in \mathcal{N}} w_{i,k} \chi_k \right)_{i=1,2,3} \right\}.$$

With these bases we can construct the following matrices:

$$\mathbf{K}_i^{(n+1)} := \left(\int_\Omega w_i^{(n)}(x) \nabla \varphi_k(x) \cdot \nabla \varphi_h(x) dx \right)_{k,h \in \mathcal{N}}, \tag{8.2}$$

$$\mathbf{M}_{\Omega \setminus D} := \left(2\mu \int_\Omega 1_{\Omega \setminus D}(x) \varphi_k(x) \varphi_h(x) dx \right)_{k,h \in \mathcal{N}}, \tag{8.3}$$

$$\mathbf{M}_D := \left(\frac{2\lambda}{9} \int_\Omega 1_D(x) \varphi_k(x) \varphi_h(x) dx \right)_{k,h \in \mathcal{N}}. \tag{8.4}$$

By these building blocks, we can assemble

$$\mathbf{K}^{(n+1)} := \begin{pmatrix} \mathbf{K}_1^{(n+1)} + \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D & \mathbf{M}_D & \mathbf{M}_D \\ \mathbf{M}_D & \mathbf{K}_2^{(n+1)} + \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D & \mathbf{M}_D \\ \mathbf{M}_D & \mathbf{M}_D & \mathbf{K}_3^{(n+1)} + \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D \end{pmatrix}$$

and

$$\mathbf{M} := \begin{pmatrix} \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D & \mathbf{M}_D & \mathbf{M}_D \\ \mathbf{M}_D & \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D & \mathbf{M}_D \\ \mathbf{M}_D & \mathbf{M}_D & \mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D \end{pmatrix}. \tag{8.5}$$

Furthermore, let us denote the vector of the nodal values of the solution by

$$\mathbf{u}^{(n+1)} = (u_{1,k_1}^{(n+1)}, \dots, u_{1,k_{\#\mathcal{N}}}^{(n+1)}, u_{2,k_1}^{(n+1)}, \dots, u_{2,k_{\#\mathcal{N}}}^{(n+1)}, u_{3,k_1}^{(n+1)}, \dots, u_{3,k_{\#\mathcal{N}}}^{(n+1)})^T \tag{8.6}$$

assembled as a column vector containing the nodal values of each channel in order, where $k_i \in \mathcal{N}$ are nodes which are suitably ordered. In a similar way the nodal values of the data $\bar{u}, \bar{v}$ are assembled in the vector

$$\bar{\mathbf{u}} = (\bar{u}_{1,k_1}, \dots, \bar{u}_{1,k_j}, \bar{v}_{1,k_{j+1}}, \dots, \bar{v}_{1,k_{\#\mathcal{N}}}, \bar{u}_{2,k_1}, \dots, \bar{u}_{2,k_j}, \bar{v}_{2,k_{j+1}}, \dots, \bar{v}_{2,k_{\#\mathcal{N}}}, \\ \bar{u}_{3,k_1}, \dots, \bar{u}_{3,k_j}, \bar{v}_{3,k_{j+1}}, \dots, \bar{v}_{3,k_{\#\mathcal{N}}})^T. \tag{8.7}$$

For the right-hand side we have the additional requirement that $\bar{v}_{i,k} = \bar{v}_{\ell,k}$ for $i \neq \ell$, representing the gray-level values. Moreover, the order of the nodes $\{k_l : l = 1, \dots, \#\mathcal{N}\}$ is such that

$$(\mathbf{M}_{\Omega \setminus D} + \mathbf{M}_D)(\bar{u}_{i,k_1}, \dots, \bar{u}_{i,k_j}, \bar{v}_{i,k_{j+1}}, \dots, \bar{v}_{i,k_{\#\mathcal{N}}})^T = \mathbf{M}_{\Omega \setminus D} \begin{pmatrix} \bar{\mathbf{u}}_\mathbf{i} \\ 0 \end{pmatrix} + \mathbf{M}_D \begin{pmatrix} 0 \\ \bar{\mathbf{v}}_\mathbf{i} \end{pmatrix}.$$

With these notations and conventions, the solution of the system of equations (8.1) is equivalent to the solution of the following algebraic linear system:

$$\mathbf{K}^{(n+1)}\mathbf{u}^{(n+1)} = \mathbf{M}\bar{\mathbf{u}}. \tag{8.8}$$

8.2. Numerical implementation of the double-minimization algorithm. We have now all the ingredients to assemble our numerical scheme into the following algorithm.

Algorithm 1. **DOUBLE_MINIMIZATION**

```
Input:        Data vector $\bar{\mathbf{u}}$, $\varepsilon_h > 0$, initial gradient weight $w^{(0)}$ with $\varepsilon_h \leq w_{i,k}^{(0)} \leq$
              $1/\varepsilon_h$, number $n_{\max}$ of outer iterations.
Parameters:   Positive weights $\lambda, \mu \geq 0$.
Output:       Approximation $u^*$ of the minimizer of $F_h$
$\mathbf{u}^{(0)} := 0$;
$\mathbf{f} := \mathbf{M}\bar{\mathbf{u}}$;
for $n := 0$ to $n_{\max}$ do
    Assemble the matrix $\mathbf{K}^{(n+1)}$ as in (8.2);
    Compute $\mathbf{u}^{(n+1)}$ such that $\mathbf{K}^{(n+1)}\mathbf{u}^{(n+1)} := \mathbf{f}$;
    Assemble the solution $u^{(n+1)} = (\sum_{k\in\mathcal{N}} u_{i,k}^{(n+1)}\varphi_k)_{i=1,2,3}$;
    Compute the gradient $\nabla u^{(n+1)} = (\sum_{k\in\mathcal{N}} u_{i,k}^{(n+1)}\nabla\varphi_k)_{i=1,2,3}$;
    $w_i^{(n+1)} := \varepsilon_h \vee \frac{1}{|\nabla u_i^{(n+1)}|} \wedge \frac{1}{\varepsilon_h}, \quad i = 1, \ldots, M$;
endfor
$u^* := u^{(n+1)}$.
```

8.3. Numerical experiments in color image restoration and results. In this section we show numerical results dealing with applications of the algorithm to color image restoration. We assume as in Figure 1.1 to have available few color fragments of the image and the gray levels of the missing parts. In all the experiments we report, we also furnish a corresponding ground truth image for comparison. The support of the image is $\Omega = [0,1]^2$, where we construct a grid of dimensions $h \times w$, and w and h are the width and height of the image in pixels, respectively. On this grid a regular triangulation is defined. The values of the images are in $[0, 255]$ channelwise.

The algorithm converges to a stationary situation in a limited number of iterations. In our numerical tests 3–4 iterations are sufficient; see Figures 7.1 and 8.4. The quality of the reconstruction increases for increasing the amount of correct color information of the datum. Nevertheless we observe that the geometrical distribution of the color datum is more crucial for a better reconstruction. A remarkable result is illustrated in Figures 8.1 and 8.2. In the bottom-left positions we illustrate data with only 3% of the original color information, randomly distributed. From this very limited complete information the algorithm still produces a rather good reconstruction of the original color images. Let us emphasize this once more:

It is sufficient to have a very limited guess of possible colors which are nicely distributed in the image to recolor all of the image.

This result has a significant impact for several possible applications. Besides the problem of the restoration of the fresco colors (where we have available 8% of the total color surface), we can use this algorithm in old black and white video and image restoration and for extreme compression of color images. For the sake of further reference, in Figure 8.3 we illustrate a comparison between the algorithm proposed in [26] and our reconstruction. In our experiments the results appear visually equivalent, although our method tends to reproduce more accurately the luminosity of the image,

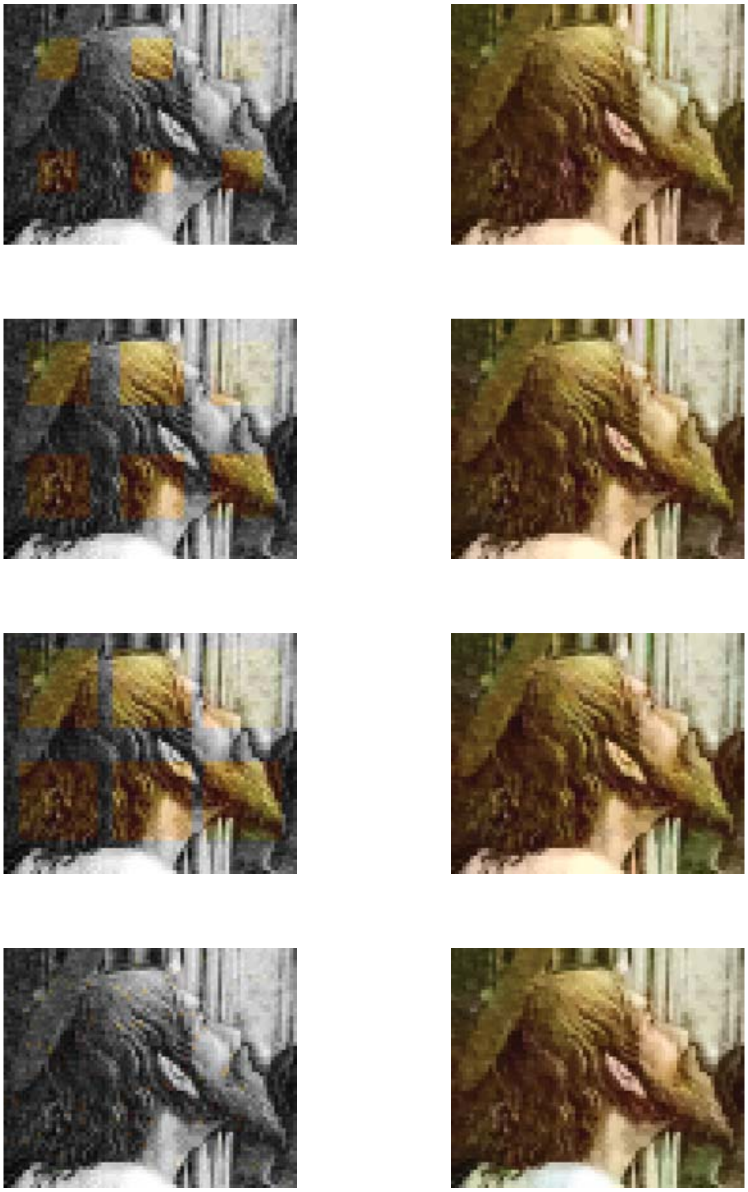


Fig. 8.1. *The first column illustrates a sequence of different data. The second column illustrates the corresponding* 10*th iteration of the algorithm. The original color image to be reconstructed is illustrated in Figure* 7.1 *in the bottom-right position. The parameters we have used are* $\varepsilon_h = 10^{-4}$, $\lambda = \mu = 150$. *In the bottom-left position we illustrate a datum with only* 3% *of the original color information, randomly distributed.*

because of the enforced constraint of the gray-level reproduction in the inpainting region. The method proposed in [26] can perform slightly better pick signal-to-noise ratios (PSNRs), though in most of the cases they do not differ significantly. For a more specific discussion on fidelity in recolorization, we refer the reader to [22]. We

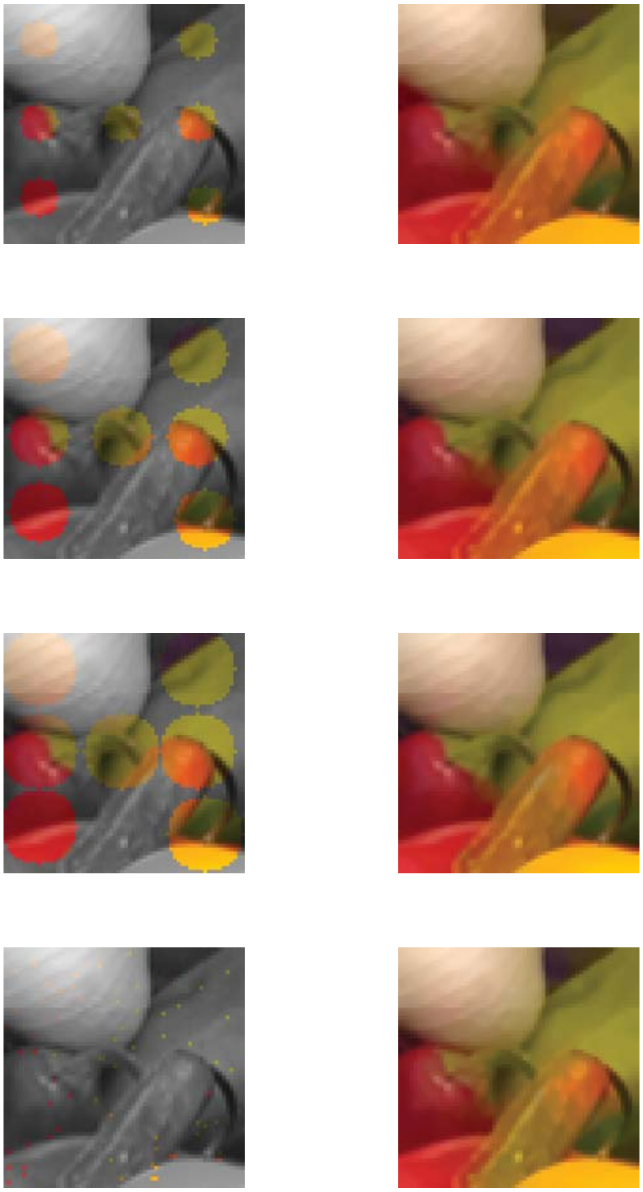


FIG. 8.2. *The first column illustrates a sequence of different data. The second column illustrates the corresponding* 10*th iteration of the algorithm. The parameters we have used are* $\varepsilon_h = 10^{-4}$, $\lambda = \mu = 150$. *In the bottom-left position we illustrate a datum with only* 3% *of the original color information, randomly distributed. The original color image to be reconstructed is illustrated in the top-right position of Figure* 8.3. *This image serves as a ground truth for the numerical experiments.*

conclude with a brief discussion on the parameters $\lambda, \mu, \varepsilon_h$. In Figure 8.4 we show the history of the residual error with respect to the original color image for increasing choices of the parameters λ, μ. These numerical results confirm the regularization effect due to the total variation constraint. The choice of ε_h has a twofold function. It

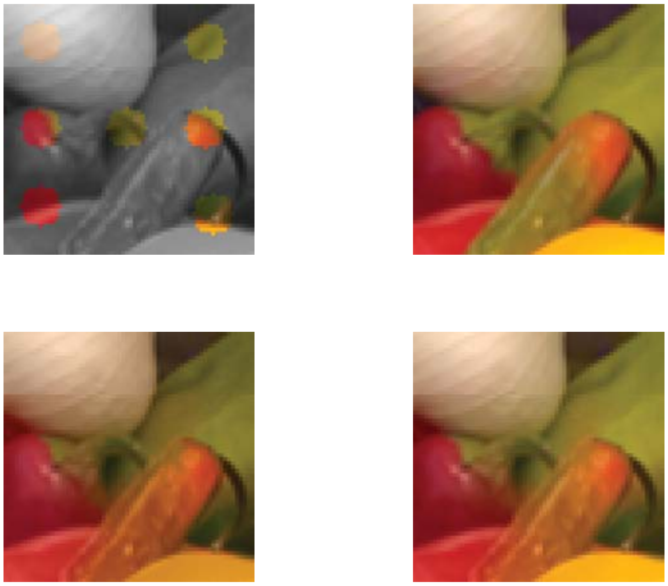


FIG. 8.3. *The figure on the top-left reports the initial data, and on the top-right we have the original image. The figure on the bottom-left is the recolorization by means of the algorithm presented in* [26] *generated by the MATLAB code provided at http://www.cs.huji.ac.il/∼yweiss/Colorization/colorization.zip. On the bottom-right we again report the reconstruction due to our algorithm. The PSNR of a color image u with respect to a ground truth image $\bar{u}$ is defined by PSNR* $= 10\log_{10}\big(\frac{255^2}{\frac{1}{3hw}\sum_{i=1}^{3}\|u_i-\bar{u}_i\|_2^2}\big)$, *where h, w are the height and width of the image, respectively. Although the PSNRs are* 31.61 *dB and* 29.48 *dB, respectively, particularly visible is a more accurate luminosity restoration, e.g., in the yellow region, due to our algorithm.*

serves as a regularization parameter; i.e., the visual smoothness of the reconstruction depends on ε_h. The larger values of ε_h give smoother reconstructed images. This effect is due to the fact that if ε_h gets large, then the corresponding differential operator $\nabla \cdot \big(\frac{\phi_h'(|\nabla u_i|)}{|\nabla u_i|}\nabla u_i\big)$ becomes more and more isotropic. Moreover, since in discrete images the gradients are always bounded, if ε_h is smaller than a threshold $T > 0$ depending on the mesh size τ—in our experiments $T = (255 \max\{h, w\})^{-1}$—then the lower bound on the gradient weight becomes irrelevant in the algorithm. However, the second purpose of this parameter is also for the sake of numerical stability. Depending on the size of the image, this parameter cannot be too small (i.e., minimal); otherwise the corresponding stiffness matrices $\mathbf{K}^{(n)}$ might be significantly ill-conditioned, and suitable preconditioners (multigrid and subspace correction/domain decomposition methods) should be invoked in this case.

Acknowledgments. The first author acknowledges the hospitality of NuHAG (the Numerical Harmonic Analysis Group), Faculty of Mathematics, University of Vienna, during the early preparation of this work. The second author wishes to thank Prof. N. Fusco for a helpful discussion. Both authors would like to thank the anonymous referees for their useful comments and suggestions.

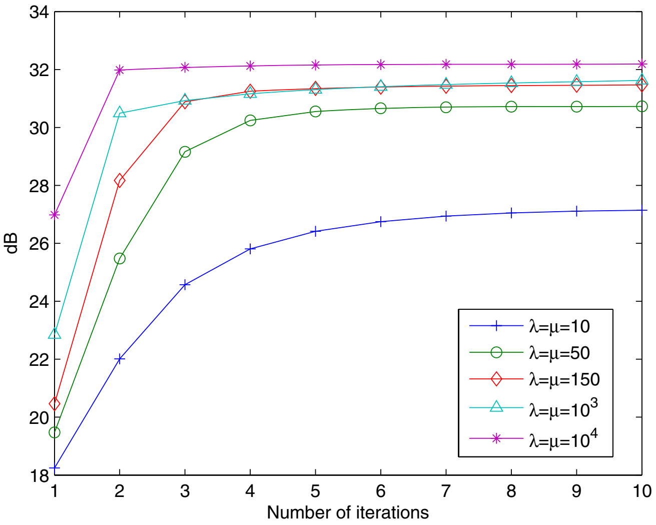


FIG. 8.4. *The plots illustrate the PSNR for different iterations of the algorithm applied to the image in Figure* 7.1 *and for different values of the parameters* λ, μ.

REFERENCES

[1] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variation*, Boll. Un. Mat. Ital. B (7), 3 (1989), pp. 857–881.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Clarendon Press, Oxford University Press, New York, 2000.

[3] L. Ambrosio and S. Masnou, *A direct variational approach to a problem arising in image reconstruction*, Interfaces Free Bound., 5 (2003), pp. 63–81.

[4] G. Aubert, R. Deriche, and P. Kornprobst, *Computing optical flow via variational techniques*, SIAM J. Appl. Math., 60 (1999), pp. 156–182.

[5] G. Aubert and P. Kornprobst, *A mathematical study of the relaxed optical flow problem in the space* $BV(\Omega)$, SIAM J. Math. Anal., 30 (1999), pp. 1282–1308.

[6] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing. Partial Differential Equations and the Calculus of Variations*, Springer-Verlag, New York, 2002.

[7] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[8] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, *Filling-in by joint interpolation of vector fields and gray levels*, IEEE Trans. Image Process., 10 (2001), pp. 1200–1211.

[9] M. Beltramio, G. Sapiro, V. Caselles, and B. Ballester, *Image inpainting*, in Proceedings of the 27th Annual ACM Conference on Computer Graphics, 2000, pp. 417–424.

[10] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland Math. Stud. 5, North–Holland, Amsterdam, London, American Elsevier, New York, 1973.

[11] A. Brook, R. Kimmel, and N. A. Sochen, *Variational restoration and edge detection for color images*, J. Math. Imaging Vision, 18 (2003), pp. 247–268.

[12] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[13] T. F. Chan, S. H. Kang, and J. Shen, *Euler's elastica and curvature-based inpainting*, SIAM J. Appl. Math., 63 (2002), pp. 564–592.

[14] T. F. Chan and J. Shen, *Inpainting based on nonlinear transport and diffusion*, in Inverse Problems, Image Analysis, and Medical Imaging, Contemp. Math. 313, AMS, Providence, RI, 2002, pp. 53–65.

[15] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.

[16] T. F. Chan and J. Shen, *Variational image inpainting*, Comm. Pure Appl. Math., 58 (2005), pp. 579–619.

[17] D. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana Univ. Math. J., 33 (1984), pp. 673–709.

[18] D. C. Dobson and C. R. Vogel, *Convergence of an iterative method for total variation denoising*, SIAM J. Numer. Anal., 34 (1997), pp. 1779–1791.

[19] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North–Holland, Amsterdam, Oxford, American Elsevier, New York, 1976; reprinted, SIAM, Philadelphia, 1999.

[20] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[21] M. Fornasier, *Nonlinear projection recovery in digital inpainting for color image restoration*, J. Math. Imaging Vision, 24 (2006), pp. 359–373.

[22] M. Fornasier, *Faithful recovery of vector valued functions from incomplete data. Recolorization and art restoration*, in Proceedings of the First International Conference on Scale Space and Variational Methods in Computer Vision, Lecture Notes in Comput. Sci. 4485, F. Sgallari, A. Murli, and N. Paragios, eds., Springer-Verlag, Berlin, 2007, pp. 116–127.

[23] M. Fornasier and H. Rauhut, *Recovery algorithms for vector valued data with joint sparsity constraints*, SIAM J. Numer. Anal., to appear.

[24] M. Fornasier and D. Toniolo, *Fast, robust and efficient 2D pattern recognition for re-assembling fragmented images*, Pattern Recognition, 38 (2005), pp. 2074–2087.

[25] C. Goffman and P. A. Raviart, *Sublinear functions of measures and variational integrals*, Duke Math. J., 31 (1964), pp. 159–178.

[26] A. Levin, D. Lischinski, and Y. Weiss, *Colorization using optimization*, ACM Trans. Graph., 23 (2004), pp. 689–694.

[27] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.

[28] N. Meyers, *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.

[29] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–684.

[30] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

[31] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Appl. Math. Optim., 44 (2001), pp. 131–161.

[32] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput., 17 (1996), pp. 227–238.

[33] L. Yatziv and G. Sapiro, *Fast image and video colorization using chrominance blending*, IEEE Trans. Image Process., 15 (2006), pp. 1120–1129.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 461–485

ASYMMETRIC CREEPING MOTION OF A RIGID SPINDLE-SHAPED BODY IN A VISCOUS FLUID[*]

M. ZABARANKIN[†]

Abstract. Exact solutions to the three-dimensional problems of asymmetric creeping translation and rotation of a rigid spindle-shaped body in a viscous incompressible fluid have been obtained. In both problems, the velocity field has been represented in the form of Dean and O'Neill, and under certain conditions, the equation of continuity has been reduced to a three-contour equation for an analytic function related to the density in a Fourier integral, representing the pressure in bispherical coordinates. Then, the three-contour equation has been reduced to a Fredholm integral equation of the second kind with a quasi-difference kernel by the complex Fourier transform. As an illustration for the obtained solutions, the pressure at the surface of the body has been calculated and analyzed. The resisting force and torque have been obtained for an arbitrary body of revolution via limits of certain harmonic functions at infinity and, as an example, have been computed for various values of a geometrical parameter of the spindle-shaped body for asymmetric translation and rotation, respectively.

Key words. spindle-shaped body, asymmetric Stokes flow, exact solution, three-contour equation, complex Fourier transform, Fredholm integral equation, Riemann boundary-value problem, drag force, resisting torque

AMS subject classifications. 30E20, 35Q15, 35Q30, 76D07

DOI. 10.1137/070679958

1. Introduction.

This paper presents exact solutions to the three-dimensional (3D) problems of *asymmetric creeping translation* and *rotation* of a rigid spindle-shaped body in a quiescent viscous incompressible fluid. The problems are formulated in the framework of the linearized Navier–Stokes equations that govern slow flows of viscous incompressible fluids and neglect inertial effects:

$$\operatorname{grad}\wp = \mu\,\boldsymbol{\Delta}\mathbf{u}, \qquad \operatorname{div}\mathbf{u} = 0, \tag{1}$$

where $\mathbf{u}$ is the fluid velocity field, $\wp$ is the pressure in the fluid, μ is the shear viscosity, and $\boldsymbol{\Delta}\mathbf{u} \equiv \operatorname{grad}(\operatorname{div}\mathbf{u}) - \operatorname{curl}(\operatorname{curl}\mathbf{u})$. The model (1) describes so-called *Stokes (creeping) flows* and is known as the Stokes equations [7, 10]. We also consider that the velocity field $\mathbf{u}$ and pressure $\wp$ vanish at infinity:

$$\mathbf{u}|_\infty = 0, \qquad \wp|_\infty = 0. \tag{2}$$

Constructing *exact* solutions to the model (1) has been and continues to be one of the central themes in analytical hydrodynamics. Much of the work in this research area has been dedicated to studying Stokes flows due to axially symmetric and asymmetric motions of a rigid body of *revolution*. By asymmetric motion of the body, we will understand translation along and rotation around axes orthogonal to the body's axis of revolution.

The problem of 3D Stokes flows arising from *axially symmetric rotation* of bodies of revolution reduces to finding merely one harmonic function [8] and is the simplest

[*] Received by the editors January 11, 2007; accepted for publication (in revised form) July 9, 2007; published electronically November 28, 2007.
http://www.siam.org/journals/siap/68-2/67995.html

[†] Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030 (mzabaran@stevens.edu).

from this class. On the other hand, the 3D Stokes flow problem corresponding to *axially symmetric translation* of bodies of revolution can be reduced to determining a biharmonic *stream function* [7, 19] and has been solved for *sphere* [17], *prolate* and *oblate spheroids* [12, 7], *circular disk* [12, 7], *spherical cap* [13, 2, 19], *two spheres* [16], *torus* [14, 21], *spindle-shaped body* [15, 23], and *lens-shaped body* [13, 22]. However, it is well known that the stream function approach cannot be extended to solving *asymmetric* 3D Stokes flow problems.

For arbitrary boundary conditions, Dean and O'Neill [3] suggested the solution form

$$\mathbf{u} = \tfrac{1}{2\mu}\,\mathfrak{r}\,\wp + \mathfrak{F}, \tag{3}$$

where $\mathfrak{r}$ is the radius vector and $\mathfrak{F}$ is an arbitrary harmonic vector ($\boldsymbol{\Delta}\mathfrak{F} = 0$). In this case, the continuity equation $\operatorname{div}\mathbf{u} = 0$ reduces to

$$3\wp + (\mathfrak{r}\cdot \operatorname{grad}\wp) + 2\mu \operatorname{div}\mathfrak{F} = 0. \tag{4}$$

The representation (3) was used to construct exact solutions to the Stokes flow problems for the asymmetric translation and rotation of a rigid torus [6, 21] and bispheres (two spheres of equal size) [20].

Let (x, y, z) and (r, φ, z) be systems of Cartesian and cylindrical coordinates with bases $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ and $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$, respectively, which have the same z-axis and are related in an ordinary way, and let S be the surface of the rigid body of revolution, whose axis of revolution (symmetry) coincides with the z-axis. We also introduce the so-called k-harmonic operator, Δ_k, by

$$\Delta_k = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2} - \frac{k^2}{r^2}.$$

Without loss of generality, we consider the following asymmetric Stokes flow problems (which, to the best of our knowledge, have not been addressed for the spindle-shaped body).

Problem I (asymmetric translation). *Let the body of revolution translate along the x-axis with the constant velocity v_x. Then the vector $\mathbf{u}$ solves* (1) *and* (2) *and satisfies the* no-slip *boundary condition on the body's surface S:*

$$\mathbf{u}|_S = v_x \mathbf{i},$$

which in component form in the cylindrical coordinates is reformulated as

$$u_r|_S = v_x \cos\varphi, \qquad u_\varphi|_S = -v_x \sin\varphi, \qquad u_z|_S = 0. \tag{5}$$

Problem II (asymmetric rotation). *Let the body of revolution rotate around the y-axis with the constant angular velocity ϖ_y. Then the vector $\mathbf{u}$ solves* (1) *and* (2) *and satisfies the* no-slip *boundary condition on the body's surface S:*

$$\mathbf{u}|_S = \left[\varpi_y\,\mathbf{j} \times (x\,\mathbf{i} + z\,\mathbf{k})\right],$$

which in component form in the cylindrical coordinates reduces to

$$u_r|_S = \varpi_y\, z\, \cos\varphi, \qquad u_\varphi|_S = -\varpi_y\, z\, \sin\varphi, \qquad u_z|_S = -\varpi_y\, r\, \cos\varphi. \tag{6}$$

It is seen from the boundary conditions (5) and (6) that in both cases, the pressure $\wp$ and solution form (3) can be represented by

$$\wp(r,\varphi,z) = \mu\,\mathit{\Theta}(r,z)\,\cos\varphi, \tag{7}$$

$$\begin{aligned} u_r(r,\varphi,z) &= u_r^{(1)}(r,z)\cos\varphi = \tfrac{1}{2}\,\left(r\,\mathit{\Theta}(r,z) + \mathit{\Upsilon}(r,z) + \mathit{\Phi}(r,z)\right)\cos\varphi,\\ u_\varphi(r,\varphi,z) &= u_\varphi^{(1)}(r,z)\sin\varphi = \tfrac{1}{2}\,\left(\mathit{\Phi}(r,z) - \mathit{\Upsilon}(r,z)\right)\sin\varphi,\\ u_z(r,\varphi,z) &= u_z^{(1)}(r,z)\cos\varphi = \left(\tfrac{1}{2}\,z\,\mathit{\Theta}(r,z) + \mathit{\Psi}(r,z)\right)\cos\varphi, \end{aligned} \tag{8}$$

where the functions $\mathit{\Theta}$ and $\mathit{\Upsilon}$, $\mathit{\Phi}$, and $\mathit{\Psi}$ are associated with the first harmonics of the pressure $\wp$ and vector-function $\mathfrak{F}$, respectively, and satisfy

$$\Delta_1\,\mathit{\Theta} = 0, \qquad \Delta_0\,\mathit{\Upsilon} = 0, \qquad \Delta_2\,\mathit{\Phi} = 0, \qquad \Delta_1\,\mathit{\Psi} = 0. \tag{9}$$

For (8), (4) reduces to

$$\left(r\frac{\partial}{\partial r} + z\frac{\partial}{\partial z} + 3\right)\mathit{\Theta} + \frac{\partial}{\partial r}\mathit{\Upsilon} + \left(\frac{\partial}{\partial r} + \frac{2}{r}\right)\mathit{\Phi} + 2\,\frac{\partial}{\partial z}\,\mathit{\Psi} = 0. \tag{10}$$

A known approach for solving the problem (2), (8), (9), and (10) for both boundary conditions (5) and (6) is to represent the harmonic functions $\mathit{\Theta}$, $\mathit{\Upsilon}$, $\mathit{\Phi}$, and $\mathit{\Psi}$ in the form of integrals or series in curvilinear coordinates, associated with the geometry of the body of revolution, then to express $\mathit{\Upsilon}$, $\mathit{\Phi}$, and $\mathit{\Psi}$ via $\mathit{\Theta}$ from the corresponding boundary conditions, and finally to substitute the series or integrals representing these functions into (10).

For bispheres and a torus, the function $\mathit{\Theta}$ can be represented in bispherical and toroidal coordinates, respectively, in the form of a series, and by this approach, (10) reduces to a *second-order difference equation* with respect to series coefficients X_k [20, 21, 6]:

$$a_k\,X_{k+1} + b_k\,X_k + c_k\,X_{k-1} = F_k, \qquad k \geq 1, \tag{11}$$

where a_k, b_k, c_k, and F_k are known functions such that $a_k \to 0$, $b_k \to 0$, $c_k \to 0$, and $F_k \to 0$ as $k \to \infty$. In [20, 21, 6], (11), also known as an infinite tridiagonal algebraic system (arising in 3D problems of elastic media [9]), was reduced to finite systems for some large k and solved numerically.

In this paper, we show that for the spindle-shaped body, the described approach reduces (10) to a *three-contour equation* for an analytic function $X(s)$ related to the density in a Fourier integral representing the function $\mathit{\Theta}$ in bispherical coordinates:

$$a(s)\,X(s+1) + b(s)\,X(s) + c(s)\,X(s-1) = F(s), \qquad \operatorname{Re} s = 0, \tag{12}$$

where s is a complex variable and $a(s)$, $b(s)$, $c(s)$, and $F(s)$ are known functions. We obtain (12) under the condition that $X(s)$ is an analytic function in the strip $|\operatorname{Re} s| \leq 1$ and show that this condition holds true for the body with $\eta_0 < 2.103$ (the surface of the spindle-shaped body is determined by fixing the coordinate η in the bispherical coordinates (ξ,η,φ); i.e., $\eta = \eta_0$). Then, we reduce (12) to a three-contour equation with $a(s) = c(s) = 1$, and show that in this special case, the latter reduces to a Fredholm integral equation of the second kind by the complex Fourier transform. As an illustration for the suggested method, we calculate the pressure at the surface of the body in the xz–half-plane ($\varphi = 0$) for various values of η_0 for the asymmetric translation and rotation of the body.

We also derive formulas for the resisting force and torque experienced by an arbitrary rigid body of revolution in asymmetric creeping motion in a viscous incompressible fluid. For the asymmetric translation with the boundary conditions (5), the resisting force is expressed via the limit of the function Θ at $z = 0$ and $r \to \infty$, and for the asymmetric rotation with the boundary conditions (6), the resisting torque is found via the limit of the function Ψ at $z = 0$ and $r \to \infty$. These formulas resemble Payne and Pell's formula for the resisting force for a body of revolution in axially symmetric translation, which is given by the limit of a stream function at infinity; see [7]. Using the derived formulas, we compute the resisting force and torque for the spindle-shaped body for various values of η_0 ($\eta_0 < 2.103$) for the asymmetric translation and rotation, respectively.

The paper is organized into six sections. Section 2 derives formulas for the resisting force and torque for an arbitrary body of revolution slowly moving in the viscous fluid. Section 3 reduces the 3D Stokes flow problems for asymmetric creeping motions of the rigid spindle-shaped body to the Fredholm integral equation of the second kind. Sections 4 and 5 obtain exact solutions to the problems in the cases of asymmetric translation and rotation, respectively, and compute the resisting force and torque for various values of the geometrical parameter η_0. Section 6 concludes the paper. Appendices A and B present proofs for the propositions dealing with computation of the resisting force and torque, respectively.

2. Resisting force and torque.

This section derives the formulas for the resisting force and torque experienced by an arbitrary rigid body of revolution in asymmetric creeping motion in a quiescent viscous incompressible fluid. The formulas resemble the one of Payne and Pell for the resisting force for the body in axially symmetric translation, which is given by the limit of a stream function at infinity; see [13].

2.1. Resisting force.

Proposition 1 (resisting force). *The resisting force for a rigid body of revolution, slowly moving in the fluid, is given by*

$$\mathbf{F} = -\iint_{\widetilde{S}} \left(\mu \left[\mathfrak{n} \times \boldsymbol{\omega}\right] + \wp\, \mathfrak{n}\right) dS, \tag{13}$$

where $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$ is the vorticity, $\widetilde{S}$ is an arbitrary smooth surface encompassing the body, and $\mathfrak{n}$ is the outer normal of $\widetilde{S}$.

Proof. The proof is presented in Appendix A.

Let the body whose axis of revolution is determined by the z-axis[1] translate in the fluid along the x-axis. In this case, the velocity field of the fluid satisfies the boundary conditions (5) and can be represented in the form (8).

Theorem 2 (resisting force for asymmetric translation). *In terms of the solution form* (8), *the resisting force for the body having the z-axis of revolution and translating along the x-axis is given by*

$$F_x = -4\pi\mu \lim_{r\to\infty} \left(r^2\, \Theta(r, z)\big|_{z=0} \right). \tag{14}$$

Proof. We calculate the resisting force by formula (13) assuming that the surface $\widetilde{S}$ is a sphere with large radius R_0. Let (R, ϑ, φ) be a system of spherical coordinates

[1] The Cartesian and cylindrical coordinates have the same z-axis.

with the basis $(\mathbf{e}_R, \mathbf{e}_\vartheta, \mathbf{e}_\varphi)$ and relating to the cylindrical coordinates (r, φ, z) in the ordinary way. For the sphere $\widetilde{S}$, we have $dS = R_0^2 \sin\vartheta \, d\vartheta \, d\varphi$, $\mathfrak{r} = R_0 \mathbf{e}_R$, and $\mathfrak{n} = \mathbf{e}_R$, and thus, (13) reduces to

$$\mathbf{F} = -\int_0^{2\pi} \int_0^{\pi} \left(\mu \left[\mathbf{e}_R \times \boldsymbol{\omega}\right] + \wp \, \mathbf{e}_R\right) R_0^2 \sin\vartheta \, d\vartheta \, d\varphi. \tag{15}$$

In the case of the asymmetric translation of the body along the x-axis, the velocity field is given by (8), and consequently, in the spherical coordinates, the vector $\mathbf{u}$ takes the form

$$\begin{aligned} \mathbf{u} = \mathbf{e}_R & \left\{ \tfrac{1}{2} R \, \Theta(R,\vartheta) + \tfrac{1}{2} \left[\varUpsilon(R,\vartheta) + \varPhi(R,\vartheta) \right] \sin\vartheta + \varPsi(R,\vartheta) \cos\vartheta \right\} \cos\varphi \\ & + \mathbf{e}_\vartheta \left\{ \tfrac{1}{2} \left[\varUpsilon(R,\vartheta) + \varPhi(R,\vartheta) \right] \cos\vartheta - \varPsi(R,\vartheta) \cos\vartheta \right\} \cos\varphi \\ & + \mathbf{e}_\varphi \; \tfrac{1}{2} \left(\varPhi(R,\vartheta) - \varUpsilon(R,\vartheta) \right) \sin\varphi, \end{aligned} \tag{16}$$

where the functions Θ, $\varUpsilon$, $\varPhi$, and $\varPsi$ satisfy (9) and can be represented by

(17)

$$\begin{aligned} \Theta(R,\vartheta) &= \sum_{n=1}^{\infty} A_n R^{-n-1} \, \mathrm{P}_n^{(1)}(\cos\vartheta) = -A_1 R^{-2} \sin\vartheta - \tfrac{3}{2} A_2 R^{-3} \sin[2\vartheta] + \mathcal{O}\left(R^{-4}\right), \\ \varUpsilon(R,\vartheta) &= \sum_{n=0}^{\infty} B_n R^{-n-1} \, \mathrm{P}_n(\cos\vartheta) = B_0 R^{-1} + B_1 R^{-2} \cos\vartheta + \mathcal{O}\left(R^{-3}\right), \\ \varPhi(R,\vartheta) &= \sum_{n=2}^{\infty} C_n R^{-n-1} \, \mathrm{P}_n^{(2)}(\cos\vartheta) = \mathcal{O}\left(R^{-3}\right), \\ \varPsi(R,\vartheta) &= \sum_{n=1}^{\infty} D_n R^{-n-1} \, \mathrm{P}_n^{(1)}(\cos\vartheta) = -D_1 R^{-2} \sin\vartheta + \mathcal{O}\left(R^{-3}\right). \end{aligned}$$

Here, $\mathrm{P}_n^{(k)}(\cos\vartheta)$ is the associated Legendre polynomial of the first kind of order n and rank k. For $k = 0$, the superscript is omitted.

Then, substituting (17) into (10), we obtain

$$(A_1 + B_0) \, R^{-2} \, \mathrm{P}_1^{(1)}(\cos\vartheta) + (B_1 - 2D_1) \, R^{-3} \, \mathrm{P}_2^{(1)}(\cos\vartheta) + \mathcal{O}\left(R^{-4}\right) = 0, \tag{18}$$

whence

$$A_1 + B_0 = 0, \qquad B_1 - 2D_1 = 0, \qquad \ldots. \tag{19}$$

For (16) with (17) and (19), the vorticity $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$ reduces to

$$\begin{aligned} \boldsymbol{\omega} = A_1 R^{-2} & \left\{ \mathbf{e}_\vartheta \sin\varphi + \mathbf{e}_\varphi \cos\vartheta \cos\varphi \right\} + \tfrac{1}{2} R^{-3} \left\{ \mathbf{e}_R \left(4D_1 \sin\vartheta \sin\varphi\right) \right. \\ & \left. + \mathbf{e}_\vartheta \left(3A_2 - 2D_1\right) \cos\vartheta \sin\varphi - \mathbf{e}_\varphi \left(2D_1 - 3A_2 \cos[2\vartheta]\right) \cos\varphi \right\} + \mathcal{O}\left(R^{-4}\right). \end{aligned} \tag{20}$$

In the case of the asymmetric translation along the x-axis, the resisting force has the component in the direction $\mathbf{i}$ only. We substitute (20) and the function Θ in the form (17) into (15) and obtain

$$F_x = (\mathbf{F} \cdot \mathbf{i}) = 4\pi\mu \, A_1 + \mathcal{O}\left(R_0^{-1}\right), \tag{21}$$

where A_1 can be determined from (17) by

$$A_1 = -\lim_{R \to \infty} \left\{ R^2 \; \Theta(R,\vartheta)\big|_{\vartheta = \frac{\pi}{2}} \right\}.$$

Consequently, passing R_0 to infinity in (21), we obtain (14). □

Obviously, for the translation along any axis orthogonal to the z-axis with the same velocity v_x, the formula for the resisting force will be the same, in particular, $F_y = F_x$.

2.2. Resisting torque.

Proposition 3 (resisting torque). *The resisting torque for a rigid body of revolution, slowly rotating in the fluid, is given by*

$$\mathbf{T} = -\mu \iint_{\widetilde{S}} \left([\mathfrak{n} \times [\mathfrak{r} \times \boldsymbol{\omega}]] + [\mathfrak{n} \times \mathbf{u}] - \tfrac{1}{\mu} \, [\mathfrak{n} \times \mathfrak{r}] \, \wp + (\mathfrak{r} \cdot \boldsymbol{\omega}) \, \mathfrak{n} \right) dS, \tag{22}$$

where $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$ *is the vorticity,* $\mathfrak{r}$ *is the radius vector,* $\widetilde{S}$ *is an arbitrary smooth surface encompassing the body, and* $\mathfrak{n}$ *is the outer normal of* $\widetilde{S}$.

Proof. The proof is presented in Appendix B.

Let the body with the z-axis of revolution rotate in the fluid around the y-axis. In this case, the velocity field of the fluid satisfies the boundary conditions (6) and, as in the case of the asymmetric translation along the x-axis, can be represented in the form (8).

Theorem 4 (resisting torque for asymmetric rotation). *In terms of the solution form* (8), *the resisting torque for the body having the z-axis of revolution and rotating around the y-axis is given by*

$$T_y = 8\pi\mu \lim_{r \to \infty} \left(r^2 \, \Psi(r, z) \big|_{z=0} \right). \tag{23}$$

Proof. As in the proof of Theorem 2, let $\widetilde{S}$ be the sphere of large radius R_0, and let (R, ϑ, φ) be the system of the spherical coordinates. For $\widetilde{S}$, we have $dS = R_0^2 \sin\vartheta \, d\vartheta \, d\varphi$, $\mathfrak{r} = R_0 \mathbf{e}_R$, and $\mathfrak{n} = \mathbf{e}_R$, and thus, (22) reduces to

$$\mathbf{T} = -\mu \int_0^{2\pi} \int_0^{\pi} \left(2R_0 \, (\mathbf{e}_R \cdot \boldsymbol{\omega}) \, \mathbf{e}_R - R_0 \, \boldsymbol{\omega} + [\mathbf{e}_R \times \mathbf{u}] \right) R_0^2 \sin\vartheta \, d\vartheta \, d\varphi. \tag{24}$$

In the case of the asymmetric rotation around the y-axis, the resisting torque has the component in the direction $\mathbf{j}$ only. Then, substituting (16), (17), and (20) into (24) and using (19), we obtain

$$T_y = (\mathbf{T} \cdot \mathbf{j}) = -8\pi\mu \, D_1 + \mathcal{O}\left(R_0^{-1}\right), \tag{25}$$

where D_1 can be determined from (17) by

$$D_1 = - \lim_{R \to \infty} \left\{ R^2 \, \Psi(R, \vartheta) \big|_{\vartheta = \frac{\pi}{2}} \right\}.$$

Consequently, passing R_0 to infinity in (25), we obtain (23). □

Obviously, for the rotation around any axis orthogonal to the z-axis with the same angular velocity ϖ_y, the formula for the resisting torque will be the same, in particular, $T_x = T_y$.

3. Fredholm integral equation for asymmetric motion of a rigid spindle-shaped body.

Let (ξ, η, φ) be a system of bispherical coordinates, in which the angular coordinate φ coincides with the one in the cylindrical coordinates (r, φ, z) and coordinates ξ and η are related to r and z by

$$r = c \, \frac{\sin\eta}{\cosh\xi - \cos\eta}, \qquad z = c \, \frac{\sinh\xi}{\cosh\xi - \cos\eta}, \qquad \begin{array}{c} -\infty < \xi < \infty, \\ 0 \le \eta \le \pi, \end{array} \tag{26}$$

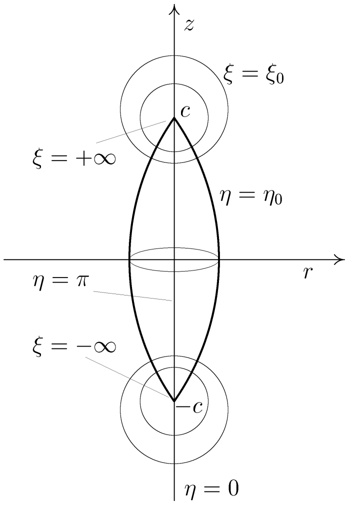


FIG. 1. *The bispherical coordinates and spindle-shaped body.*

where c is a metric parameter of the bispherical coordinates.

The spindle-shaped body is the body of revolution whose surface S is formed by rotating a circle arc around the z-axis[2] and can be described in the bispherical coordinates by fixing the coordinate η, i.e., $\eta = \eta_0$ (see Figure 1). For $\eta_0 = \frac{\pi}{2}$, the surface of the body forms a sphere, and for $\eta_0 < \frac{\pi}{2}$ and $\eta_0 > \frac{\pi}{2}$, the body's shape resembles an "apple" and "lemon," respectively.

For the boundary conditions (5) and (6), the velocity field $\mathbf{u}$ can be represented by (8), where the functions $\varTheta$, $\varUpsilon$, $\varPhi$, and $\varPsi$ satisfy (9) and (10). For convenience, we denote

$$(27)\qquad f_1 = \left(u_r^{(1)} - u_\varphi^{(1)}\right)\Big|_{\eta=\eta_0}, \qquad f_2 = \left(u_r^{(1)} + u_\varphi^{(1)}\right)\Big|_{\eta=\eta_0}, \qquad f_3 = u_z^{(1)}\Big|_{\eta=\eta_0},$$

and reformulate (5) and (6) for $\varTheta$, $\varUpsilon$, $\varPhi$, and $\varPsi$ in a general form:

$$(28)\quad \left(\tfrac{1}{2}\, r\, \varTheta + \varUpsilon\right)\big|_{\eta=\eta_0} = f_1, \quad \left(\tfrac{1}{2}\, r\, \varTheta + \varPhi\right)\big|_{\eta=\eta_0} = f_2, \quad \left(\tfrac{1}{2}\, z\, \varTheta + \varPsi\right)\big|_{\eta=\eta_0} = f_3.$$

For the domain exterior to the spindle-shaped body, i.e., for $\eta \in [0, \eta_0]$, the functions $\varTheta$, $\varUpsilon$, $\varPhi$, and $\varPsi$ can be represented in the bispherical coordinates by Fourier integrals with respect to ξ:

$$(29)\qquad \varTheta(\xi,\eta) = \frac{1}{2\pi i}\sqrt{\cosh\xi - \cos\eta}\int_{-i\infty}^{+i\infty} A(s)\, \mathrm{P}^{(1)}_{-\frac{1}{2}+s}(\cos\eta)\, \mathrm{e}^{-\xi s} ds, \quad \eta \le \eta_0,$$

$$(30)\qquad \varUpsilon(\xi,\eta) = \frac{1}{2\pi i}\sqrt{\cosh\xi - \cos\eta}\int_{-i\infty}^{+i\infty} B(s)\, \mathrm{P}_{-\frac{1}{2}+s}(\cos\eta)\, \mathrm{e}^{-\xi s} ds, \quad \eta \le \eta_0,$$

$$(31)\qquad \varPhi(\xi,\eta) = \frac{1}{2\pi i}\sqrt{\cosh\xi - \cos\eta}\int_{-i\infty}^{+i\infty} C(s)\, \mathrm{P}^{(2)}_{-\frac{1}{2}+s}(\cos\eta)\, \mathrm{e}^{-\xi s} ds, \quad \eta \le \eta_0,$$

[2] The z-axis is the axis of revolution of the spindle-shaped body.

$$\mathit{\Psi}(\xi,\eta)=\frac{1}{2\pi i}\sqrt{\cosh\xi-\cos\eta}\int_{-i\infty}^{+i\infty}D(s)\,\mathrm{P}^{(1)}_{-\frac{1}{2}+s}(\cos\eta)\,\mathrm{e}^{-\xi s}ds,\quad \eta\le\eta_0, \tag{32}$$

where $\mathrm{P}^{(k)}_{-\frac{1}{2}+s}(\cos\eta)$ is the associated Legendre function of the first kind of complex index s and rank k, behaving as $\frac{|\operatorname{Im}s|^k}{\sqrt{2\pi\sin\eta}}\,\mathrm{e}^{\eta|\operatorname{Im}s|}$ at $\operatorname{Re}s=0$ and $|\operatorname{Im}s|\to\infty$; see [1]. For $k=0$, the superscript is omitted.

We assume $A(s)$, $B(s)$, $C(s)$, and $D(s)$ in the corresponding Fourier integrals (29)–(32) to be meromorphic functions in the strip $|\operatorname{Re}s|\le 1$, vanishing as $\mathcal{O}(\mathrm{e}^{-\gamma|\operatorname{Im}s|})$ at $\operatorname{Re}s=0$ and $|\operatorname{Im}s|\to\infty$, where $\gamma>\eta_0$.

Substituting (29)–(32) into (10), where the derivatives of the functions $\mathit{\Theta}$, $\mathit{\Upsilon}$, $\mathit{\Phi}$, and $\mathit{\Psi}$ in (10) are derived similarly to formulas (2.1) and (2.2) in [22], we obtain

$$\begin{aligned} &c\left(\left(s+\tfrac{3}{2}\right)A(s+1)+5A(s)-\left(s-\tfrac{3}{2}\right)A(s-1)\right)+B(s+1)-2B(s)+B(s-1)\\ &\quad-\left(\left(s+\tfrac{5}{2}\right)\left(s+\tfrac{3}{2}\right)C(s+1)-2\left(s^2-\tfrac{9}{4}\right)C(s)+\left(s-\tfrac{5}{2}\right)\left(s-\tfrac{3}{2}\right)C(s-1)\right)\\ &\quad+2\left(\left(s+\tfrac{3}{2}\right)D(s+1)-2s\,D(s)+\left(s-\tfrac{3}{2}\right)D(s-1)\right)=0, \end{aligned} \tag{33}$$

under the condition that

$$\begin{aligned} &\operatorname*{Res}_{0\le\operatorname{Re}s\le1}\left\{\left(\left(s+\tfrac{1}{2}\right)\left(c\,A(s)-\left(s+\tfrac{3}{2}\right)C(s)+2D(s)\right)+B(s)\right)\mathrm{P}^{(1)}_{-\frac{3}{2}+s}(\cos\eta)\right\}=0,\\ &\operatorname*{Res}_{-1\le\operatorname{Re}s\le0}\left\{\left(\left(s-\tfrac{1}{2}\right)\left(c\,A(s)+\left(s-\tfrac{3}{2}\right)C(s)-2D(s)\right)-B(s)\right)\mathrm{P}^{(1)}_{\frac{1}{2}+s}(\cos\eta)\right\}=0, \end{aligned} \tag{34}$$

which follow from the assumption that $A(s)$, $B(s)$, $C(s)$, and $D(s)$ may admit poles in $|\operatorname{Re}s|\le1$.

Let $\mathcal{A}_{[-1,1]}$ define the class of functions *analytic* in the strip $|\operatorname{Re}s|\le1$, i.e., with no poles in $|\operatorname{Re}s|\le1$, and vanishing within the strip at $|s|\to\infty$. Also, to simplify further notation, let

$$\alpha_k(s)=\mathrm{P}^{(k)}_{-\frac{1}{2}+s}(\cos\eta_0),\qquad \beta_k(s)=\mathrm{P}^{(k)}_{-\frac{3}{2}+s}(\cos\eta_0),\qquad \gamma_k(s)=\mathrm{P}^{(k)}_{\frac{1}{2}+s}(\cos\eta_0). \tag{35}$$

We introduce a new function $X(s)$ related to $\mathit{\Theta}(\xi,\eta_0)$ by the complex Fourier transform:

$$X(s)=\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathit{\Theta}(\xi,\eta_0)\,\mathrm{e}^{\xi s}}{(\cosh\xi-\cos\eta_0)^{\frac{3}{2}}}\,d\xi,\qquad s\in\mathbb{C}, \tag{36}$$

$$\mathit{\Theta}(\xi,\eta_0)=\frac{1}{\pi i}\,(\cosh\xi-\cos\eta_0)^{\frac{3}{2}}\int_{-i\infty}^{+i\infty}X(s)\,\mathrm{e}^{-\xi s}ds. \tag{37}$$

The critical condition for the rest of the analysis is to consider that $X(s)\in\mathcal{A}_{[-1,1]}$. As will be shown in the end of this section, this condition holds true for the spindle-shaped body with $\eta_0<2.103$.

Since $\mathit{\Theta}(\xi,\eta_0)$ is the boundary value for the function $\mathit{\Theta}(\xi,\eta)$, we obtain from (29) that

$$A(s)=\frac{1}{\alpha_1(s)}\int_{-\infty}^{\infty}\frac{\mathit{\Theta}(\xi,\eta_0)\,\mathrm{e}^{\xi s}\,d\xi}{\sqrt{\cosh\xi-\cos\eta_0}}=\frac{1}{\alpha_1(s)}\left(X(s+1)-2X(s)\cos\eta_0+X(s-1)\right). \tag{38}$$

Similarly, from the boundary conditions (28) and representations (30), (31), and (32), we can express the functions $B(s)$, $C(s)$, and $D(s)$ via $X(s)$:

$$\begin{aligned} B(s) &= \frac{1}{\alpha_0(s)}\left(F_1(s) - c\,X(s)\sin\eta_0\right), \\ C(s) &= \frac{1}{\alpha_2(s)}\left(F_2(s) - c\,X(s)\sin\eta_0\right), \\ D(s) &= \frac{1}{\alpha_1(s)}\left(F_3(s) - \tfrac{c}{2}\left(X(s+1) - X(s-1)\right)\right), \end{aligned} \tag{39}$$

where

$$F_j(s) = \int_{-\infty}^{\infty} \frac{f_j(\xi)\,\mathrm{e}^{\xi s}}{\sqrt{\cosh\xi - \cos\eta_0}}\,d\xi, \qquad j = 1, 2, 3. \tag{40}$$

With (38) and (39), equation (33) reduces to a so-called three-contour equation for the function $X(s)$:

$$\begin{aligned} &\left(\sin\eta_0\left(\frac{\left(s+\frac52\right)\left(s+\frac32\right)}{\gamma_2(s)} - \frac{1}{\gamma_0(s)}\right) - \frac{2\left(s+\frac32\right)\cos\eta_0}{\gamma_1(s)} + \frac{2\left(s+\frac52\right)}{\alpha_1(s)}\right)X(s+1) \\ &+ 2\left(\sin\eta_0\left(\frac{1}{\alpha_0(s)} - \frac{\left(s^2-\frac94\right)}{\alpha_2(s)}\right) - \frac{5\cos\eta_0}{\alpha_1(s)} + \frac{\left(s+\frac32\right)}{\gamma_1(s)} - \frac{\left(s-\frac32\right)}{\beta_1(s)}\right)X(s) \\ &+ \left(\sin\eta_0\left(\frac{\left(s-\frac52\right)\left(s-\frac32\right)}{\beta_2(s)} - \frac{1}{\beta_0(s)}\right) + \frac{2\left(s-\frac32\right)\cos\eta_0}{\beta_1(s)} - \frac{2\left(s-\frac52\right)}{\alpha_1(s)}\right)X(s-1) \\ &+ R_1(s+1) + R_2(s) + R(s-1) = 0, \end{aligned} \tag{41}$$

where

$$\begin{aligned} R_1(s) &= \frac{1}{c}\left(\frac{F_1(s)}{\alpha_0(s)} - \frac{\left(s+\frac12\right)\left(s+\frac32\right)F_2(s)}{\alpha_2(s)} + \frac{2\left(s+\frac12\right)F_3(s)}{\alpha_1(s)}\right), \\ R_2(s) &= -\frac{2}{c}\left(\frac{F_1(s)}{\alpha_0(s)} - \frac{\left(s^2-\frac94\right)F_2(s)}{\alpha_2(s)} + \frac{2sF_3(s)}{\alpha_1(s)}\right), \\ R_3(s) &= \frac{1}{c}\left(\frac{F_1(s)}{\alpha_0(s)} - \frac{\left(s-\frac12\right)\left(s-\frac32\right)F_2(s)}{\alpha_2(s)} + \frac{2\left(s-\frac12\right)F_3(s)}{\alpha_1(s)}\right). \end{aligned} \tag{42}$$

The crucial point in the analysis of (41) is to notice that the coefficients at $X(s+1)$ and $X(s-1)$ can be represented as $\widetilde{\mathfrak{D}}(s+1)/\alpha_1(s)$ and $\widetilde{\mathfrak{D}}(s-1)/\alpha_1(s)$, respectively,[3] where the function $\widetilde{\mathfrak{D}}(s)$ is defined by

$$\widetilde{\mathfrak{D}}(s) = \frac{\mathfrak{D}(s)}{\alpha_0(s)\alpha_1(s)\alpha_2(s)} \tag{43}$$

with

$$\begin{aligned} \mathfrak{D}(s) = {}& 2\alpha_1^3(s)\cos^2\eta_0 + \left(s^2-\tfrac14\right)^2\alpha_0^3(s)\sin[2\eta_0] + 2\left(s^2-\tfrac14\right)\alpha_0^2(s)\alpha_1(s)\cos[2\eta_0] \\ &+ 2\left(\left(s^2-\tfrac54\right)\sin^2\eta_0 - 2\right)\alpha_0(s)\alpha_1^2(s)\cot\eta_0. \end{aligned} \tag{44}$$

[3] Noticing this, in fact, requires a great deal of manipulation.

In terms of $\widetilde{\mathfrak{D}}(s)$, the condition (34) is reformulated as

$$\begin{aligned}
&\operatorname*{Res}_{0\le \operatorname{Re} s\le 1}\left\{\left(\frac{2\left(s+\frac{1}{2}\right)X(s-1)}{\alpha_1(s)}+\frac{\left(\widetilde{\mathfrak{D}}(s)-2\left(s+\frac{3}{2}\right)\right)X(s)}{\beta_1(s)}+R_1(s)\right)\right.\\
&\qquad\qquad\qquad\qquad\qquad\left.\times\,\mathrm{P}^{(1)}_{-\frac{3}{2}+s}(\cos\eta)\right\}=0,\\
&\operatorname*{Res}_{-1\le \operatorname{Re} s\le 0}\left\{\left(-\frac{2\left(s-\frac{1}{2}\right)X(s+1)}{\alpha_1(s)}+\frac{\left(\widetilde{\mathfrak{D}}(s)+2\left(s-\frac{3}{2}\right)\right)X(s)}{\gamma_1(s)}+R_3(s)\right)\right.\\
&\qquad\qquad\qquad\qquad\qquad\left.\times\,\mathrm{P}^{(1)}_{\frac{1}{2}+s}(\cos\eta)\right\}=0,
\end{aligned}\tag{45}$$

and the three-contour equation (41) takes the form

$$\widetilde{\mathfrak{D}}(s+1)X(s+1)+\frac{2K(s)\,X(s)}{\alpha_0(s)\alpha_2(s)\beta_1(s)\gamma_1(s)}+\widetilde{\mathfrak{D}}(s-1)X(s-1)=\mathcal{F}(s),\quad \operatorname{Re} s=0,\tag{46}$$

where

$$\begin{aligned}
K(s)=\ &(\beta_1(s)+\gamma_1(s))\left(3\alpha_0(s)\alpha_1(s)\alpha_2(s)-\tfrac{1}{2}\,\mathfrak{D}(s)\right)\\
&+3\alpha_0(s)\beta_1(s)\gamma_1(s)\left(\alpha_1(s)\sin\eta_0-2\alpha_2(s)\cos\eta_0\right)
\end{aligned}$$

and

$$\mathcal{F}(s)=-\alpha_1(s)\left(R_1(s+1)+R_2(s)+R_3(s-1)\right).\tag{47}$$

Now we analyze the condition (45) provided that $X(s)$ is analytic in the strip $|\operatorname{Re} s|\le 1$. Note that (45) holds true if each term in the left-hand sides of equations (45) is analytic in the corresponding strip ($0\le \operatorname{Re} s\le 1$ or $-1\le \operatorname{Re} s\le 0$). However, the converse statement is not correct.

1. We begin with the first terms in the left-hand sides of (45).
   (i) The analyticity of $X(s)$ in $|\operatorname{Re} s|\le 1$ implies that the functions $X(s-1)$ and $X(s+1)$ are analytic in $0\le \operatorname{Re} s\le 1$ and $-1\le \operatorname{Re} s\le 0$, respectively.
   (ii) The function $\alpha_1(s)=\mathrm{P}^{(1)}_{-\frac{1}{2}+s}(\cos\eta_0)$ (as well as $\alpha_2(s)=\mathrm{P}^{(2)}_{-\frac{1}{2}+s}(\cos\eta_0)$) has only "generic"[4] zeros $s=\pm\frac{1}{2}$ in $|\operatorname{Re} s|\le 1$ for all $\eta_0\in(0,\pi)$, which are, however, compensated by zeros $s=\pm\frac{1}{2}$ of the multipliers $\mathrm{P}^{(1)}_{-\frac{3}{2}+s}(\cos\eta)$ and $\mathrm{P}^{(1)}_{\frac{1}{2}+s}(\cos\eta)$ in the strips $0\le \operatorname{Re} s\le 1$ and $-1\le \operatorname{Re} s\le 0$, respectively.
2. We proceed to the second terms in the left-hand sides of (45).
   (i) By the same reasoning as in item 1(ii), we conclude that

$$\mathrm{P}^{(1)}_{-\frac{3}{2}+s}(\cos\eta)\Big/\beta_1(s)\quad\text{and}\quad \mathrm{P}^{(1)}_{\frac{1}{2}+s}(\cos\eta)\Big/\gamma_1(s)$$

[4] Zeros $s=\pm\frac{1}{2}$, which $\alpha_1(s)$, $\alpha_2(s)$, and $\mathfrak{D}(s)$ have for all $\eta_0\in(0,\pi)$, are referred to as "generic," in contrast to "individual" zeros, which depend on η_0.

have no poles for all $\eta_0 \in (0, \pi)$ in $0 \le \operatorname{Re} s \le 1$ and $-1 \le \operatorname{Re} s \le 0$, respectively.
(ii) According to (43) and (44), poles of $\widetilde{\mathfrak{D}}(s)$ are determined by zeros of the product $\alpha_0(s)\,\alpha_1(s)\,\alpha_2(s)$. However, the only zeros $s = \pm\frac{1}{2}$ of $\alpha_1(s)$ and $\alpha_2(s)$ in $|\operatorname{Re} s| \le 1$ (see item 1(ii)) are compensated by zeros $s = \pm\frac{1}{2}$ of multiplicity 2 of $\mathfrak{D}(s)$. Consequently, since the function $\alpha_0(s) = \mathrm{P}_{-\frac{1}{2}+s}(\cos\eta_0)$ has zeros in $|\operatorname{Re} s| \le 1$ for $\eta_0 \ge 2.281$ only, $\widetilde{\mathfrak{D}}(s)$ has no poles in $|\operatorname{Re} s| \le 1$ for $\eta_0 < 2.281$.

3. At last, the functions $R_1(s)$ and $R_3(s)$ in (45) may admit only simple poles $s = \pm\frac{1}{2}$ in $0 \le \operatorname{Re} s \le 1$ and $-1 \le \operatorname{Re} s \le 0$, respectively (because of the corresponding multipliers $\mathrm{P}^{(1)}_{-\frac{3}{2}+s}(\cos\eta)$ and $\mathrm{P}^{(1)}_{\frac{1}{2}+s}(\cos\eta)$). This condition will be verified in sections 4 and 5 for the problems of asymmetric translation and rotation, respectively.

Concluding the analysis of (45), we assume that $\eta_0 < 2.281$ and require the functions $R_1(s)$ and $R_3(s)$ to admit simple poles $s = \pm\frac{1}{2}$ in $|\operatorname{Re} s| \le 1$ only. Consequently, these conditions along with the analyticity of $X(s)$ in $|\operatorname{Re} s| \le 1$ are sufficient for (45) to hold true (however, they may not be necessary).

Finally, introducing a new function

$$\widetilde{X}(s) = \widetilde{\mathfrak{D}}(s) X(s) \tag{48}$$

and denoting

$$\widetilde{K}(s) = \frac{\alpha_1(s)}{\beta_1(s)\gamma_1(s)}\,\frac{K(s)}{\mathfrak{D}(s)}, \tag{49}$$

we obtain

$$\widetilde{X}(s+1) + 2\widetilde{K}(s)\widetilde{X}(s) + \widetilde{X}(s-1) = \mathcal{F}(s), \qquad \operatorname{Re} s = 0. \tag{50}$$

To solve (50), we first need to determine a class of functions for $\widetilde{X}(s)$. The condition $\eta_0 < 2.281$ guarantees that $\widetilde{\mathfrak{D}}(s)$ is analytic in $|\operatorname{Re} s| \le 1$ (see item 2(ii) in the analysis of (45)), and consequently, the analyticity of $\widetilde{\mathfrak{D}}(s)$ and $X(s)$ in $|\operatorname{Re} s| \le 1$ implies that $\widetilde{X}(s)$, defined by (48), is analytic in $|\operatorname{Re} s| \le 1$. Moreover, from (43) and the asymptotic behavior of $\alpha_k(s)$ for $k = 0$, 1 and 2, it follows that $\widetilde{\mathfrak{D}}(s) \to 2\sin^2\eta_0$ at $\operatorname{Re} s = 0$ and $|\operatorname{Im} s| \to \infty$, which, along with $X(s) \to 0$ at $\operatorname{Re} s = 0$ and $|\operatorname{Im} s| \to \infty$, results in having $\widetilde{X}(s) \to 0$ at $\operatorname{Re} s = 0$ and $|\operatorname{Im} s| \to \infty$. Thus, we conclude that

$$X(s) \in \mathcal{A}_{[-1,1]} \quad \text{and} \quad \eta_0 < 2.281 \quad \Longrightarrow \quad \widetilde{X}(s) \in \mathcal{A}_{[-1,1]}.$$

We will seek to find a solution to the three-contour equation (50) in the class $\mathcal{A}_{[-1,1]}$. Although obtaining a closed-form solution to (50) is still an open issue, (50) can be reduced to a Fredholm integral equation of the second kind for the function

$$H(s) = \widetilde{X}(s+1) + \widetilde{X}(s-1). \tag{51}$$

Provided that $\widetilde{X}(s) \in \mathcal{A}_{[-1,1]}$, we apply the complex Fourier transform to (51) and obtain

$$\int_{-i\infty}^{+i\infty} H(\tau)\,\mathrm{e}^{\tau\, t}\,d\tau = 2\cosh t \int_{-i\infty}^{+i\infty} \widetilde{X}(\tau)\,\mathrm{e}^{\tau\, t}\,d\tau,$$

whence $\widetilde{X}(s)$ is expressed via $H(s)$ by the inverse Fourier transform:

$$\widetilde{X}(s) = \frac{1}{4\pi i}\int_{-i\infty}^{+i\infty}\left(\int_{-\infty}^{\infty}\frac{\mathrm{e}^{(\tau-s)\,t}}{\cosh t}\,dt\right)H(\tau)\,d\tau = \frac{1}{4i}\int_{-i\infty}^{+i\infty}\frac{H(\tau)\,d\tau}{\cos\left[\frac{\pi}{2}(\tau-s)\right]}. \tag{52}$$

To show that $\widetilde{X}$, determined by (52), is unique, we need to prove that (51), as an equation with respect to $\widetilde{X}$, has only zero homogeneous solution in the class of functions analytic in the strip $|\operatorname{Re} s| \le 1$ and vanishing within the strip at $|s| \to \infty$.

PROPOSITION 5. *The equation $\widetilde{X}(s+1) + \widetilde{X}(s-1) = 0$ has only zero solution in the class $\mathcal{A}_{[-1,1]}$.*

Proof. This equation reduces to a Riemann boundary-value problem for an analytic function by the conformal mapping $w = -i\cot[\pi s/2]$ of the strip $|\operatorname{Re} s| \le 1$ into the complex plane w with the branch cut along the segment $[-1,1]$; see Figure 2. The lines $s = -1 + i\tau$ and $s = 1 + i\tau$, $\tau \in \mathbb{R}$, correspond to the upper and lower banks of the branch cut with the counterclockwise orientation as shown in Figure 2, and infinite points of the strip $|\operatorname{Re} s| \le 1$, i.e., $|s| \to \infty$, correspond to the points $w = \pm 1$.

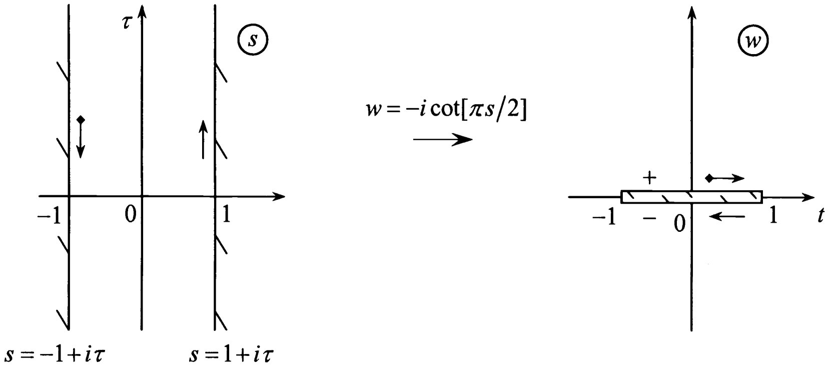


FIG. 2. *The function $w = -i\cot[\pi s/2]$ maps the strip $-1 \le \operatorname{Re} s \le 1$ of the complex plane s into the complex plane w with the branch cut along the segment* $[-1,1]$.

Let $Y(w) = \widetilde{X}(s)$; then $Y^{+}(t) = \widetilde{X}(-1+i\tau)$ and $Y^{-}(t) = \widetilde{X}(1+i\tau)$, $\tau \in \mathbb{R}$, are the boundary values of $Y(w)$ at the upper and lower banks of the branch cut, respectively. Thus, the original equation reduces to $Y^{+}(t) = -Y^{-}(t)$, $t \in [-1,1]$, which is the Riemann boundary-value problem for determining the analytic function $Y(w)$ in the complex plane w with the branch cut along the segment $[-1,1]$. An analytic function that solves this problem and vanishes at $w = \pm 1$ is given by $Y(w) = \sqrt{1-w^2}\,\mathcal{P}_n(w)$, where $\mathcal{P}_n(w)$ is a polynomial of degree n to be determined from the behavior of $Y(w)$ at $|w| \to \infty$; see [5]. But since $\widetilde{X}(s)$ is analytic at $s = 0$, the function $Y(w)$ is bounded at $|w| \to \infty$, and, consequently, $\mathcal{P}_n(w) \equiv 0$, and $Y(w)$ is the zero function. □

Thus, with (51) and (52), the three-contour equation (50) reduces to a Fredholm integral equation of the second kind for $H(s)$:

$$H(s) + \frac{\widetilde{K}(s)}{2i}\int_{-i\infty}^{+i\infty}\frac{H(\tau)\,d\tau}{\cos\left[\frac{\pi}{2}(\tau-s)\right]} = \mathcal{F}(s), \qquad \operatorname{Re} s = 0. \tag{53}$$

The analytic function $\widetilde{X}(s)$ is determined in $|\operatorname{Re} s| \le 1$ by (52), and its value at

TABLE 1
First "individual" zero of $\mathfrak{D}(s)$ for various η_0.

η_0	s_0	η_0	s_0
$\pi/12$	20.33	$7\pi/12$	1.197
$2\pi/12$	10.19	$8\pi/12$	1.005
$3\pi/12$	6.817	2.103[†]	1.000
$4\pi/12$	5.142	$9\pi/12$	0.873
$5\pi/12$	2.153	$10\pi/12$	0.777
$6\pi/12$	1.5	$11\pi/12$	0.696

[†] For $\eta_0 \geq 2.103$, the first "individual" zero lies within the strip $|\operatorname{Re} s| \leq 1$.

the contour $\operatorname{Re} s = 0$ is expressed from (53) by

$$\widetilde{X}(s) = \frac{\mathcal{F}(s) - H(s)}{2\widetilde{K}(s)}, \qquad \operatorname{Re} s = 0. \tag{54}$$

Finally, it follows from (48), along with (54) and (49), that

$$X(s) = \frac{\beta_1(s)\gamma_1(s)}{\alpha_1(s)} \frac{(\mathcal{F}(s) - H(s))}{2K(s)}, \qquad \operatorname{Re} s = 0. \tag{55}$$

Substituting (55) into the inverse Fourier transform (37), we obtain $\Theta(\xi, \eta_0)$.

Now, based on the obtained solution $\widetilde{X}(s)$, we can establish when $X(s)$ is analytic in $|\operatorname{Re} s| \leq 1$. The analyticity of the ratio $X(s) = \widetilde{X}(s)/\widetilde{\mathfrak{D}}(s)$ is guaranteed by the analyticity of $\widetilde{X}(s)$ and $1/\widetilde{\mathfrak{D}}(s)$. Since both $\mathfrak{D}(s)$ and the multiplier $\alpha_1(s)\,\alpha_2(s)$ in the denominator of $\widetilde{\mathfrak{D}}(s)$ have "generic" zeros $s = \pm\frac{1}{2}$ of multiplicity 2 that compensate each other, poles of $1/\widetilde{\mathfrak{D}}(s)$ are determined only by "individual" zeros of $\mathfrak{D}(s)$. Table 1 presents the first "individual" zero of $\mathfrak{D}(s)$ for various η_0 ($\mathfrak{D}(s)$ is an even function) and shows that $\mathfrak{D}(s)$ has no "individual" zero in the strip $|\operatorname{Re} s| \leq 1$ for $\eta_0 < 2.103$. Consequently, we conclude that

$$\widetilde{X}(s) \in \mathcal{A}_{[-1,1]} \quad \text{and} \quad \eta_0 < 2.103 \quad \Longrightarrow \quad X(s) \in \mathcal{A}_{[-1,1]}.$$

This means that the formula (55) is valid only for $\eta_0 < 2.103$.

In the next sections, we will solve (53) for the asymmetric translation of the spindle-shaped body along the x-axis and for the asymmetric rotation of the body around the y-axis.

4. Asymmetric translation. In the case of the asymmetric translation of the rigid spindle-shaped body in the fluid along the x-axis, the velocity field satisfies the boundary conditions (2) and (5) and is represented in the form (8).

From (5) and (27), we have

$$f_1 = 2v_x, \qquad f_2 = 0, \qquad f_3 = 0.$$

To calculate Fourier integrals (40), we use the representation for the associated Legendre function

$$\mathrm{P}^{(k)}_{-\frac{1}{2}+s}(\cos\eta) = \frac{(2k-1)!!}{2^k\sqrt{2}\,\pi}\cos[\pi s]\int_{-\infty}^{\infty} \frac{\left(\sin^k \eta\right) \mathrm{e}^{\tau s}\, d\tau}{(\cosh\tau + \cos\eta)^{k+\frac{1}{2}}}, \tag{56}$$

where $k \geq 0$ and $(2k-1)!! = \prod_{j=1}^{k}(2j-1)$; see [1]. For $k=0$, we define $(-1)!! = 1$. To simplify further notation, we denote

$$\alpha_0^-(s) = \mathrm{P}_{-\frac{1}{2}+s}(-\cos\eta_0), \qquad \alpha_1^-(s) = \mathrm{P}^{(1)}_{-\frac{1}{2}+s}(-\cos\eta_0).$$

Using (56), we have

$$F_1(s) = v_x\,\frac{2\sqrt{2}\,\pi}{\cos[\pi s]}\,\alpha_1^-(s), \qquad F_2(s) = 0, \qquad F_3(s) = 0,$$

and, consequently, it follows from (42) that

$$R_1(s) = R_3(s) = \frac{v_x}{c}\,\frac{2\sqrt{2}\,\pi}{\cos[\pi s]}\,\frac{\alpha_1^-(s)}{\alpha_0(s)}.$$

For $\eta_0 < 2.103$, the functions $R_1(s)$ and $R_3(s)$ are analytic in $|\operatorname{Re} s| \leq 1$ (zeros $s = \pm\frac{1}{2}$ of $\cos[\pi s]$ are compensated by those of $\alpha_1^-(s)$), and, thus, condition (45) holds true.

Then, with the convolution relationship (see [1])

$$\alpha_0^-(s)\,\alpha_1(s) + \alpha_0(s)\,\alpha_1^-(s) = \frac{2}{\pi}\,\frac{\cos[\pi s]}{\sin\eta_0}, \tag{57}$$

we obtain from (42) and (47) that

$$\mathcal{F}(s) = -\frac{v_x}{c}\,4\sqrt{2}\,\frac{\alpha_1(s)\,(2\alpha_1(s)\sin\eta_0 + \alpha_0(s)\cos\eta_0)}{\left(s^2 - \frac{1}{4}\right)\alpha_0(s)\beta_0(s)\gamma_0(s)}. \tag{58}$$

Thus, the Fredholm integral equation (53) is solved for (58), and the function $\Theta(\xi, \eta_0)$ is determined by (37) with (55).

In the case of a sphere, i.e., $\eta_0 = \frac{\pi}{2}$, equation (53) has the closed-form solution

$$H_{\text{sphere}}(s) = \frac{v_x}{c}\,\frac{6\sqrt{2}}{\mathrm{P}_{-\frac{1}{2}+s}(0)},$$

and, consequently, we obtain

$$X_{\text{sphere}}(s) = -\frac{v_x}{c}\,\frac{\sqrt{2}\left(s^2 - \frac{1}{4}\right)}{\mathrm{P}^{(1)}_{-\frac{1}{2}+s}(0)}, \qquad \Theta_{\text{sphere}}(\xi, \eta_0)\big|_{\eta_0 = \frac{\pi}{2}} = \frac{v_x}{c}\,\frac{3}{2\cosh\xi},$$

and

$$\begin{aligned}
\Theta_{\text{sphere}}(r, z) &= \frac{3}{2} v_x\, c\, \frac{r}{(r^2 + z^2)^{\frac{3}{2}}},\\
\Upsilon_{\text{sphere}}(r, z) &= \frac{3}{2} v_x\, c\, \frac{1}{\sqrt{r^2 + z^2}} + \frac{1}{4} v_x\, c^3\, \frac{2z^2 - r^2}{(r^2 + z^2)^{\frac{5}{2}}},\\
\Phi_{\text{sphere}}(r, z) &= -\frac{3}{4} v_x\, c^3\, \frac{r^2}{(r^2 + z^2)^{\frac{5}{2}}},\\
\Psi_{\text{sphere}}(r, z) &= -\frac{3}{4} v_x\, c^3\, \frac{r\,z}{(r^2 + z^2)^{\frac{5}{2}}}.
\end{aligned}$$

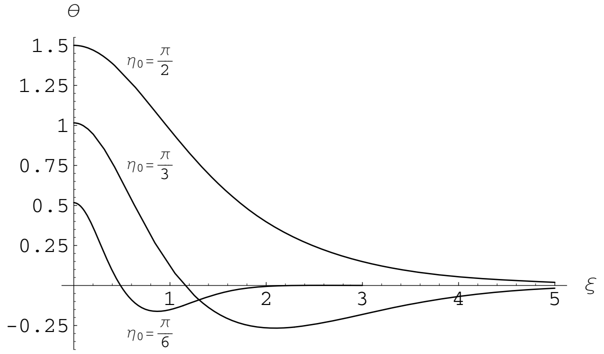


FIG. 3. *Asymmetric translation: the function* $\frac{c}{v_x}\,\Theta(\xi,\eta_0)$ *for* $\eta_0 = \frac{\pi}{6}$, $\frac{\pi}{3}$, *and* $\frac{\pi}{2}$ *("apples").*

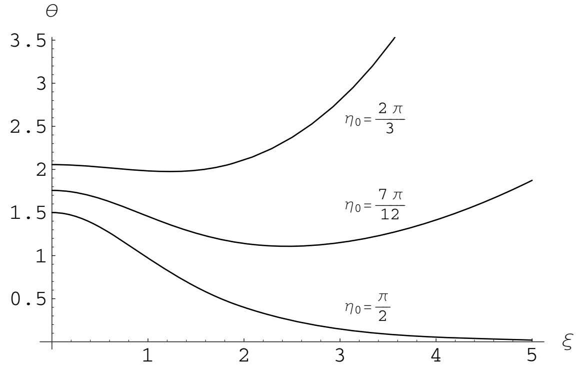


FIG. 4. *Asymmetric translation: the function* $\frac{c}{v_x}\,\Theta(\xi,\eta_0)$ *for* $\eta_0 = \frac{\pi}{2}$, $\frac{7\pi}{12}$, *and* $\frac{2\pi}{3}$ *("lemons").*

The pressure $\wp$ and function Θ are related by (7). Figures 3 and 4 illustrate $\frac{c}{v_x}\,\Theta(\xi,\eta_0)$ as a function of ξ for $\eta_0 = \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{7\pi}{12}$, and $\frac{2\pi}{3}$. Figures 5 and 6 show epures of the normalized pressures, $\frac{c}{v_x\mu}\,\wp$ and $\frac{c}{2v_x\mu}\,\wp$, at the surface of the rigid spindle-shaped body in the xz–half-plane ($\varphi = 0$) for $\eta_0 = \frac{\pi}{3}$ ("apple") and $\eta_0 = \frac{7\pi}{12}$ ("lemon"), respectively.

In the case of the asymmetric translation along the x-axis, the resisting force is given by (14), which for the function $\Theta(\xi,\eta)$ in the form (29) with (38) reduces to

$$\begin{aligned} F_x &= -2\sqrt{2}\,i\mu c^2 \int_{-i\infty}^{+i\infty} \left(s^2 - \tfrac{1}{4}\right) A(s)\,ds \\ &= -2\sqrt{2}\,i\mu c^2 \int_{-i\infty}^{+i\infty} X(s)\left(\frac{\left(s-\frac{3}{2}\right)\left(s-\frac{1}{2}\right)}{\beta_1(s)} - 2\frac{\left(s^2-\frac{1}{4}\right)\cos\eta_0}{\alpha_1(s)} + \frac{\left(s+\frac{1}{2}\right)\left(s+\frac{3}{2}\right)}{\gamma_1(s)}\right)ds, \end{aligned}$$

where in obtaining the last integral, we used the condition that $X(s)$ has no poles within the strip $|\operatorname{Re} s| \le 1$. This means that the above formula for F_x is valid only for $\eta_0 < 2.103$.

In the case of the *axially symmetric* translation of the spindle-shaped body along the z-axis with the constant velocity v_z (see [15, 23]), the resisting force has the

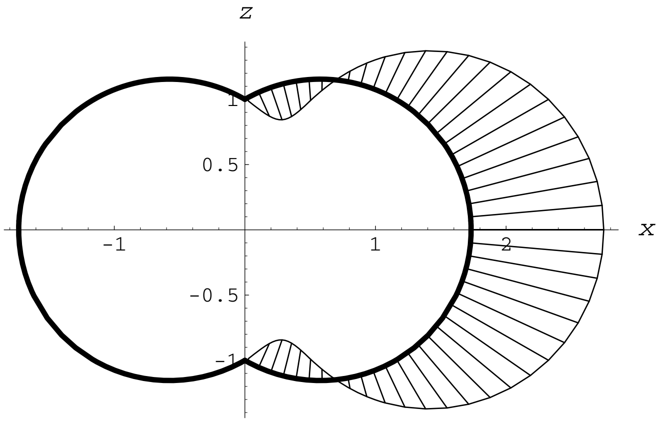


FIG. 5. *Asymmetric translation: epure of the normalized pressure* $\frac{c}{v_x \mu}\,\wp$ *at the surface of the rigid spindle-shaped body in the xz–half-plane* ($\varphi = 0$) *for* $\eta_0 = \frac{\pi}{3}$ *("apple"). At a particular point on the contour, the value of the function is depicted by the length of the outward normal line if the value is positive and by the length of the inward normal line if the value is negative.*

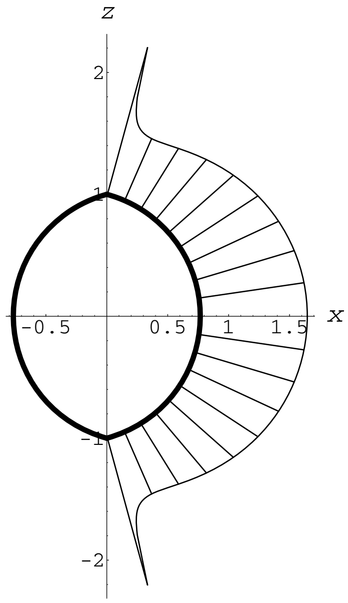


FIG. 6. *Asymmetric translation: epure of the normalized pressure* $\frac{c}{2v_x \mu}\,\wp$ *at the surface of the rigid spindle-shaped body in the xz–half-plane* ($\varphi = 0$) *for* $\eta_0 = \frac{7\pi}{12}$ *("lemon").*

component in the direction $\mathbf{k}$ only, i.e., F_z, which is determined by (see [23])

$$F_z = -4v_z\,\mu\,c\,i\int_{-i\infty}^{+i\infty}\left(\frac{\alpha_2(s)\sin\eta_0}{\mathfrak{D}_0(s)} + \frac{\pi\,\alpha_0^-(s)}{\cos[\pi s]}\right)\frac{ds}{\alpha_0(s)},$$

where $\mathfrak{D}_0(s)$ is the determinant of the axially symmetric translation problem:

$$\mathfrak{D}_0(s) = \left(1+\cos^2\eta_0\right)\alpha_0(s)\alpha_1(s) + \tfrac{1}{2}\,\sin[2\eta_0]\left(\alpha_1^2(s) + \left(s^2 - \tfrac{1}{4}\right)\alpha_0^2(s)\right).$$

Table 2
Normalized resisting forces, d_z and d_x, as functions of η_0.

η_0	d_z	$d_x = d_y$	η_0	d_z	$d_x = d_y$
$0^\dagger$	$0.9346^\ddagger$	$0.8434^\S$	$6\pi/12$	1	1
$\pi/12$	0.9365	0.8466	$7\pi/12$	1.038	1.084
$2\pi/12$	0.9405	0.8568	$8\pi/12$	1.100	1.213
$3\pi/12$	0.9475	0.8747	$9\pi/12$	1.209	–
$4\pi/12$	0.9585	0.9022	$10\pi/12$	1.430	–
$5\pi/12$	0.9751	0.9423	$11\pi/12$	2.067	–

†The case $\eta_0 = 0$ corresponds to a closed torus (torus with no opening).
‡The value is reported in [19], while 0.9353 and 0.953 are reported in [4] and [18], respectively.
§The value is reported in [11].

Table 2 compares the normalized resisting forces $d_z = \frac{\tan[\eta_0/2]}{6\pi\mu v_z c} F_z$ and $d_x = \frac{\tan[\eta_0/2]}{6\pi\mu v_x c} F_x$ for the spindle-shaped body for the axially symmetric and asymmetric translations, respectively, where $c\,\cot[\eta_0/2]$ determines the radius of the sphere inscribed into “lemon” (the body with $\eta_0 > \frac{\pi}{2}$) or circumscribed about “apple” (the body with $\eta_0 < \frac{\pi}{2}$).

5. Asymmetric rotation. In the case of the asymmetric rotation of the rigid spindle-shaped body in the fluid around the y-axis, the velocity field satisfies the boundary conditions (2) and (6) and is represented in the form (8).

From (6) and (27), we have

$$f_1 = \varpi_y c\,\frac{2\sinh\xi}{\cosh\xi - \cos\eta_0}, \qquad f_2 = 0, \qquad f_3 = -\varpi_y c\,\frac{\sin\eta_0}{\cosh\xi - \cos\eta_0},$$

and using (56), we calculate Fourier integrals (40):

$$F_1(s) = \varpi_y\, c\,\frac{4\sqrt{2}\,\pi\, s}{\cos[\pi s]}\,\alpha_0^-(s), \qquad F_2(s) = 0, \qquad F_3(s) = -\varpi_y\, c\,\frac{2\sqrt{2}\,\pi}{\cos[\pi s]}\,\alpha_1^-(s).$$

It follows from (42) that

$$R_1(s) = \varpi_y\,\frac{4\sqrt{2}\,\pi}{\cos[\pi s]}\left(\frac{s\,\alpha_0^-(s)}{\alpha_0(s)} - \frac{\left(s+\frac{1}{2}\right)\alpha_1^-(s)}{\alpha_1(s)}\right),$$
$$R_3(s) = \varpi_y\,\frac{4\sqrt{2}\,\pi}{\cos[\pi s]}\left(\frac{s\,\alpha_0^-(s)}{\alpha_0(s)} - \frac{\left(s-\frac{1}{2}\right)\alpha_1^-(s)}{\alpha_1(s)}\right).$$

For $\eta_0 < 2.103$, the above functions $R_1(s)$ and $R_3(s)$ admit only simple poles $s = \pm\frac{1}{2}$ in the strip $|\operatorname{Re} s| \le 1$, and consequently, the condition (45) holds true.

Then, with the convolution relationship (57), we obtain from (42) and (47) that

$$\mathcal{F}(s) = \varpi_y\, 8\sqrt{2}\, s\left(\frac{\alpha_1(s)\,(\alpha_0(s)\cos\eta_0 - 2\alpha_1(s)\sin\eta_0)}{\left(s^2-\frac{1}{4}\right)\alpha_0(s)\beta_0(s)\gamma_0(s)} - \frac{2\alpha_2(s)\sin\eta_0}{\beta_1(s)\gamma_1(s)}\right). \tag{59}$$

Thus, the Fredholm integral equation (53) is solved for (59), and the function $\mathit{\Theta}(\xi,\eta_0)$ is determined by (37) with (55).

In the case of a sphere, i.e., $\eta_0 = \frac{\pi}{2}$, expression (59) reduces to zero. Thus, $H_{\text{sphere}} \equiv 0$, $X_{\text{sphere}} \equiv 0$, and $\mathit{\Theta}_{\text{sphere}} \equiv 0$, and the functions $\mathit{\Upsilon}_{\text{sphere}}$, $\mathit{\Phi}_{\text{sphere}}$, and

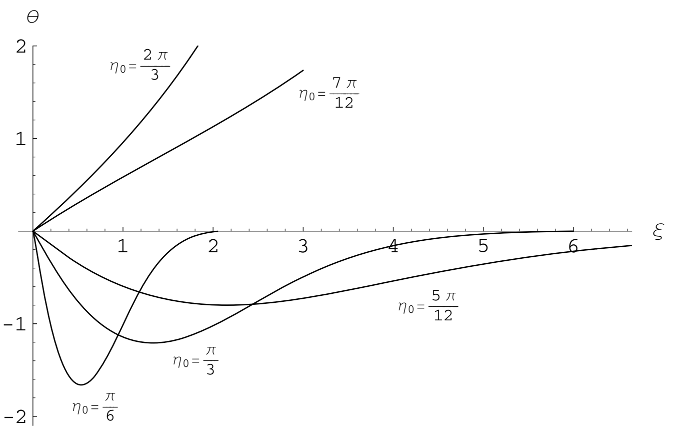
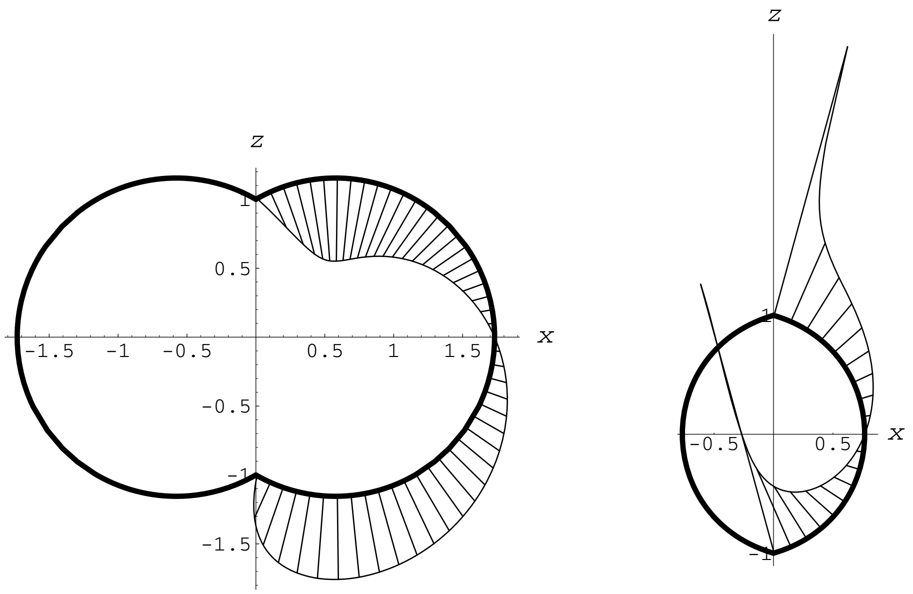


FIG. 7. *Asymmetric rotation: the function* $\frac{1}{\varpi_y}\,\Theta(\xi,\eta_0)$ *for* $\eta_0 = \frac{\pi}{6},\ \frac{\pi}{3},\ \frac{5\pi}{12},\ \frac{7\pi}{12}$*, and* $\frac{2\pi}{3}$.



$\eta_0 = \pi/3$

$\eta_0 = 7\pi/12$

FIG. 8. *Asymmetric rotation: epures of the normalized pressure* $\frac{1}{2\varpi_y\mu}\,\wp$ *at the surface of the rigid spindle-shaped body in the xz–half-plane (*$\varphi = 0$*) for* $\eta_0 = \frac{\pi}{3}$ *("apple") and* $\eta_0 = \frac{7\pi}{12}$ *("lemon").*

Ψ_{sphere} take the form

$$\Upsilon_{\text{sphere}} = \varpi_y c^3 \frac{2z}{(r^2+z^2)^{\frac{3}{2}}}, \qquad \Phi_{\text{sphere}} \equiv 0, \qquad \Psi_{\text{sphere}} = -\varpi_y c^3 \frac{r}{(r^2+z^2)^{\frac{3}{2}}}.$$

As in the case of the asymmetric translation, the pressure $\wp$ and function Θ are related by (7). Figure 7 illustrates $\Theta(\xi,\eta_0)$ as a function of ξ for $\eta_0 = \frac{\pi}{6},\ \frac{\pi}{3},\ \frac{5\pi}{12},\ \frac{7\pi}{12}$, and $\frac{2\pi}{3}$. Figure 8 shows epures of the normalized pressure $\frac{1}{2\varpi_y\mu}\,\wp$ at the surface of the rigid spindle-shaped body in the xz–half-plane ($\varphi = 0$) for $\eta_0 = \frac{\pi}{3}$ (“apple”) and $\eta_0 = \frac{7\pi}{12}$ (“lemon”).

For the asymmetric rotation around the y-axis, the torque is represented by (23), which for the function Ψ in the form (32) with (39) reduces to

$$\begin{aligned} T_y &= 4\sqrt{2}\, i\, \mu\, c^2 \int_{-i\infty}^{+i\infty} \left(s^2 - \tfrac{1}{4}\right) D(s)\, ds \\ &= -16\pi\varpi_y \mu\, c^3\, i \int_{-i\infty}^{+i\infty} \frac{\left(s^2 - \frac{1}{4}\right)}{\cos[\pi s]} \frac{\alpha_1^-(s)}{\alpha_1(s)}\, ds \\ &\quad - 4\sqrt{2}\, i\, \mu\, c^3 \sin\eta_0 \int_{-i\infty}^{+i\infty} \frac{s\, \alpha_2(s)}{\beta_1(s)\gamma_1(s)}\, X(s)\, ds, \end{aligned}$$

where in obtaining the last integral, as in the case of deriving the expression for F_x, we used the condition that $X(s)$ has no poles within the strip $|\operatorname{Re} s| \le 1$. This means that the above formula for T_y is valid only for $\eta_0 < 2.103$.

In the case of the *axially symmetric* rotation of a body of revolution around the z-axis with the constant angular velocity ϖ_z, the no-slip boundary condition for the velocity field $\mathbf{u}$ is determined by $\mathbf{u}|_S = [\varpi_z\, \mathbf{k} \times r\, \mathbf{e}_r]$, which in component form reduces to

$$u_r|_S = 0, \qquad u_\varphi|_S = \varpi_z\, r, \qquad u_z|_S = 0.$$

The vector $\mathbf{u}$ that solves (1), (2), and the above boundary condition can be represented by

$$\mathbf{u} = \Omega(r, z)\, \mathbf{e}_\varphi,$$

where Ω satisfies $\Delta_1 \Omega = 0$ (see [8]). In this case, the resisting torque has the component in the direction $\mathbf{k}$ only, i.e., T_z, and similarly to the derivation of the formula (23) for T_y, it can be shown that

$$T_z = -8\pi\mu \lim_{r\to\infty} \left(r^2 \Omega(r, z)\big|_{z=0} \right),$$

which for the spindle-shaped body reduces to

$$T_z = -16\pi\varpi_z \mu\, c^3\, i \int_{-i\infty}^{+i\infty} \frac{\left(s^2 - \frac{1}{4}\right)}{\cos[\pi s]} \frac{\alpha_1^-(s)}{\alpha_1(s)}\, ds.$$

Table 3 shows the normalized torques $t_z = -\frac{\tan^3[\eta_0/2]}{8\pi\mu\varpi_z c^3}\, T_z$ and $t_y = -\frac{\tan^3[\eta_0/2]}{8\pi\mu\varpi_y c^3}\, T_y$ for the rigid spindle-shaped body for the axially symmetric and asymmetric rotations, respectively, where $c \cot[\eta_0/2]$ determines the radius of the sphere inscribed into “lemon” (the body with $\eta_0 > \frac{\pi}{2}$) or circumscribed about “apple” (the body with $\eta_0 < \frac{\pi}{2}$).

6. Conclusions. We have obtained exact solutions to the 3D Stokes flow problems for asymmetric translation and rotation of a rigid spindle-shaped body. Representing the velocity field in the form (3), we have reduced both problems to the three-contour equation (46) for the analytic function $X(s)$, related to the density in the Fourier integral representing the pressure. The equation has been obtained under the condition $\eta_0 < 2.103$ that guarantees the analyticity of $X(s)$ in the strip

TABLE 3
Normalized resisting torques, t_z and t_y, as functions of η_0.

η_0	t_z	$t_y = t_x$	η_0	t_z	$t_y = t_x$
0[†]	0.7969[‡]	0.6498[§]	$6\pi/12$	1	1
$\pi/12$	0.8010	0.6551	$7\pi/12$	1.113	1.293
$2\pi/12$	0.8140	0.6716	$8\pi/12$	1.292	1.912
$3\pi/12$	0.8370	0.7026	$9\pi/12$	1.601	–
$4\pi/12$	0.8724	0.7553	$10\pi/12$	2.241	–
$5\pi/12$	0.9242	0.8443	$11\pi/12$	4.219	–

[†] The case $\eta_0 = 0$ corresponds to a closed torus (torus with no opening).
[‡] The value is reported in [4].
[§] The value is reported in [11].

$|\operatorname{Re} s| \leq 1$. This condition follows from the fact that the function $\widetilde{\mathfrak{D}}(s)$ has zeros in $|\operatorname{Re} s| \leq 1$ for $\eta_0 \geq 2.103$. Then, we have reduced (46) to the three-contour equation (50) with the unit coefficients at $\widetilde{X}(s+1)$ and $\widetilde{X}(s-1)$ and finally have reduced the latter to the Fredholm integral equation (53) by the complex Fourier transform. In the case of a sphere, i.e., $\eta_0 = \frac{\pi}{2}$, the integral equation has a closed-form solution for the asymmetric translation and has a zero solution for the asymmetric rotation.

We have derived formulas for the resisting force, $\mathbf{F}$, and torque, $\mathbf{T}$, experienced by a rigid body of revolution in arbitrary slow motion in a viscous incompressible fluid, and have shown that for the asymmetric translation along the x-axis and asymmetric rotation around the y-axis, F_x and T_y can be expressed via the limits of the functions $\Theta(r,z)$ and $\Psi(r,z)$ at infinity. We have computed F_x and T_y for the corresponding asymmetric motions of the spindle-shaped body for various values of η_0 ($\eta_0 < 2.103$) and compared them with F_z and T_z for the axially symmetric translation and rotation of the body, respectively. In the case of "apple" ($\eta_0 < \frac{\pi}{2}$), $F_x < F_z$ and $T_y < T_z$, and in the case of "lemon" ($\eta_0 > \frac{\pi}{2}$), $F_x > F_z$ and $T_y > T_z$; see Tables 2 and 3. Also, F_x and T_y for the spindle-shaped body are in accordance with F_x and T_y for a closed torus.

For both problems of asymmetric motion, we have calculated the pressure at the surface of the spindle-shaped body in the xz–half-plane ($\varphi = 0$) for various values of η_0 ($\eta_0 < 2.103$). It follows from (37) that the function $\Theta(\xi, \eta_0)$, associated with the pressure, is infinite at $\xi \to \pm\infty$ when $X(s)$ has a pole (except $\frac{1}{2}$) with the real part less than $\frac{3}{2}$. Since the poles of $X(s)$ are determined by zeros of $\mathfrak{D}(s)$, we conclude based on Table 1 that in both problems, $\Theta(\xi, \eta_0) \to 0$ for "apples" ($\eta_0 < \frac{\pi}{2}$), and $|\Theta(\xi, \eta_0)| \to \infty$ for "lemons" ($\eta_0 > \frac{\pi}{2}$) when $\xi \to \pm\infty$. This conclusion is supported by numerical calculations; see Figures 3–8. Notably, in the case of *axially symmetric* translation, the pressure at the surface at $\xi \to \pm\infty$ is finite for $\eta_0 \leq \frac{2\pi}{3}$ and is infinite for $\eta_0 > \frac{2\pi}{3}$; see [23].

Obtaining exact solutions for the asymmetric Stokes flow problems for the rigid spindle-shaped body in the case $\eta_0 \geq 2.103$ as well as calculating the corresponding resisting force and torque is still an open issue.

Appendix A. Proof of Proposition 1.

This section proves Proposition 1.

Let (r, φ, z) be a system of cylindrical coordinates with the basis $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$, in which the z-axis determines the body's axis of revolution, and let $\mathfrak{n} = n_r\, \mathbf{e}_r + n_z\, \mathbf{k}$ define the outer normal to the body's surface S, where $n_r = \frac{\partial r}{\partial n}$ and $n_z = \frac{\partial z}{\partial n}$.

By definition, the force, exerted by the fluid on the elementary surface dS with the normal $\mathfrak{n}$, is given by

$$\mathbf{P}_n = 2\mu\,(\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u} + \mu\,[\mathfrak{n}\times\operatorname{curl}\mathbf{u}] - \wp\,\mathfrak{n} \tag{60}$$

(see [19]), and the total force is defined as the integral over the body's surface S:

$$\mathbf{F} = \iint_S \left(2\mu\,(\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u} + \mu\,[\mathfrak{n}\times\operatorname{curl}\mathbf{u}] - \wp\,\mathfrak{n}\right) dS. \tag{61}$$

Let (s,φ,n) be a characteristic coordinate system with the right-handed orthogonal basis $(\mathfrak{s},\mathbf{e}_\varphi,\mathfrak{n})$. Then, in (s,φ,n), the operators $\operatorname{grad}\wp$, $\operatorname{div}\mathbf{u}$, and $\operatorname{curl}\mathbf{u}$ take the form

$$\begin{aligned}
\operatorname{grad}\wp &= \mathfrak{s}\,\frac{\partial\wp}{\partial s} + \mathbf{e}_\varphi\,\frac{1}{r}\,\frac{\partial\wp}{\partial\varphi} + \mathfrak{n}\,\frac{\partial\wp}{\partial n},\\
\operatorname{div}\mathbf{u} &= \frac{1}{r}\,\frac{\partial}{\partial s}\,(r\,u_s) + \frac{1}{r}\,\frac{\partial}{\partial\varphi}\,u_\varphi + \frac{1}{r}\,\frac{\partial}{\partial n}\,(r\,u_n),\\
\operatorname{curl}\mathbf{u} &= \mathfrak{s}\left(\frac{1}{r}\,\frac{\partial}{\partial\varphi}\,u_n - \frac{1}{r}\,\frac{\partial}{\partial n}\,(r\,u_\varphi)\right) + \mathbf{e}_\varphi\left(\frac{\partial}{\partial n}\,u_s - \frac{\partial}{\partial s}\,u_n\right)\\
&\qquad + \mathfrak{n}\left(\frac{1}{r}\,\frac{\partial}{\partial s}\,(r\,u_\varphi) - \frac{1}{r}\,\frac{\partial}{\partial\varphi}\,u_s\right),
\end{aligned} \tag{62}$$

where (u_s,u_φ,u_n) are the components of $\mathbf{u}$ in $(\mathfrak{s},\mathbf{e}_\varphi,\mathfrak{n})$; see [7]. The derivatives $\frac{\partial}{\partial s}$, $\frac{\partial}{\partial\varphi}$, and $\frac{\partial}{\partial n}$ of the unit vectors $\mathfrak{s}$, $\mathbf{e}_\varphi$, and $\mathfrak{n}$ are given by

$$\begin{aligned}
&\tfrac{\partial}{\partial s}\,\mathfrak{s} = 0, && \tfrac{\partial}{\partial\varphi}\,\mathfrak{s} = \tfrac{\partial r}{\partial s}\,\mathbf{e}_\varphi, && \tfrac{\partial}{\partial n}\,\mathfrak{s} = 0,\\
&\tfrac{\partial}{\partial s}\,\mathbf{e}_\varphi = 0, && \tfrac{\partial}{\partial\varphi}\,\mathbf{e}_\varphi = -\tfrac{\partial r}{\partial s}\,\mathfrak{s} - \tfrac{\partial r}{\partial n}\,\mathfrak{n}, && \tfrac{\partial}{\partial n}\,\mathbf{e}_\varphi = 0,\\
&\tfrac{\partial}{\partial s}\,\mathfrak{n} = 0, && \tfrac{\partial}{\partial\varphi}\,\mathfrak{n} = \tfrac{\partial r}{\partial n}\,\mathbf{e}_\varphi, && \tfrac{\partial}{\partial n}\,\mathfrak{n} = 0.
\end{aligned} \tag{63}$$

It can be shown that

$$(\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u} = -\,[\mathfrak{n}\times\operatorname{curl}\mathbf{u}] + \frac{1}{r}\left(\frac{\partial}{\partial\varphi}\,[\mathbf{u}\times\mathfrak{s}] - \frac{\partial}{\partial s}\,(r\,[\mathbf{u}\times\mathbf{e}_\varphi])\right). \tag{64}$$

Indeed, using the formula

$$\operatorname{grad}(\mathfrak{a}\cdot\mathfrak{b}) = (\mathfrak{a}\cdot\operatorname{grad})\,\mathfrak{b} + (\mathfrak{b}\cdot\operatorname{grad})\,\mathfrak{a} + [\mathfrak{a}\times\operatorname{curl}\mathfrak{b}] + [\mathfrak{b}\times\operatorname{curl}\mathfrak{a}]$$

for $\mathfrak{a} = \mathfrak{n}$ and $\mathfrak{b} = \mathbf{u}$, along with the identity $\operatorname{curl}\mathfrak{n} = 0$, we obtain

$$(\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u} = -\,[\mathfrak{n}\times\operatorname{curl}\mathbf{u}] + \operatorname{grad}(\mathfrak{n}\cdot\mathbf{u}) - (\mathbf{u}\cdot\operatorname{grad})\,\mathfrak{n}. \tag{65}$$

Then, using (62) and (63) and the fact that $\operatorname{div}\mathbf{u} = 0$, we have

$$\begin{aligned}
\frac{\partial}{\partial\varphi}\,[\mathbf{u}\times\mathfrak{s}] - \frac{\partial}{\partial s}\,(r\,[\mathbf{u}\times\mathbf{e}_\varphi]) &= \frac{\partial}{\partial\varphi}\,(u_n\,\mathbf{e}_\varphi - u_\varphi\,\mathfrak{n}) + \frac{\partial}{\partial s}\,(r\,(u_n\,\mathfrak{s} - u_s\,\mathfrak{n}))\\
&= \frac{\partial u_n}{\partial\varphi}\,\mathbf{e}_\varphi - \frac{\partial u_\varphi}{\partial\varphi}\,\mathfrak{n} - u_n\left(\frac{\partial r}{\partial s}\,\mathfrak{s} + \frac{\partial r}{\partial n}\,\mathfrak{n}\right) - u_\varphi\,\frac{\partial r}{\partial n}\,\mathbf{e}_\varphi\\
&\quad + \frac{\partial}{\partial s}\,(ru_n)\,\mathfrak{s} - \frac{\partial}{\partial s}\,(ru_s)\,\mathfrak{n}\\
&= r\,\frac{\partial u_n}{\partial s}\,\mathfrak{s} + \frac{\partial u_n}{\partial\varphi}\,\mathbf{e}_\varphi + r\,\frac{\partial u_n}{\partial n}\,\mathfrak{n} - u_\varphi\,\frac{\partial r}{\partial n}\,\mathbf{e}_\varphi\\
&\equiv r\,(\operatorname{grad}(\mathfrak{n}\cdot\mathbf{u}) - (\mathbf{u}\cdot\operatorname{grad})\,\mathfrak{n}).
\end{aligned} \tag{66}$$

The relationship (64) follows from (66) and (65).

Let ℓ denote the contour of the surface S in the rz–half-plane. If ℓ is a closed curve with no intersections, e.g., the body is a torus, then obviously $\iint_S \frac{1}{r}\frac{\partial}{\partial s}\left(r\left[\mathbf{u}\times\mathbf{e}_\varphi\right]\right)dS = 0$, where $dS = r\,ds\,d\varphi$. Otherwise, since the body is bounded, ℓ intersects the z-axis at least twice. Assuming that s varies from the first intersection of the contour ℓ with the z-axis to the last one, and having that at all the intersections $r = 0$, we again obtain $\iint_S \frac{1}{r}\frac{\partial}{\partial s}\left(r\left[\mathbf{u}\times\mathbf{e}_\varphi\right]\right)dS = 0$. Consequently, integrating (64) over the surface S with $dS = r\,ds\,d\varphi$, we have

$$\iint_S (\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u}\,dS = -\iint_S \left[\mathfrak{n}\times\operatorname{curl}\mathbf{u}\right]dS. \tag{67}$$

With (67) and the notation $\boldsymbol{\omega} = \operatorname{curl}\mathbf{u}$, the resisting force (61) reduces to

$$\mathbf{F} = -\iint_S \left(\mu\left[\mathfrak{n}\times\boldsymbol{\omega}\right] + \wp\,\mathfrak{n}\right)dS. \tag{68}$$

Finally, let $\widetilde{S}$ be an arbitrary smooth surface encompassing the body, and let V be the volume between the surfaces S and $\widetilde{S}$ ($\mathfrak{n}$ will denote the outer normal for the corresponding surface). The Stokes equations (1) can be rewritten as

$$\operatorname{grad}\wp = -\mu\operatorname{curl}\left(\operatorname{curl}\mathbf{u}\right), \qquad \operatorname{div}\mathbf{u} = 0. \tag{69}$$

Integrating the first equation in (69) over the volume V and using Gauss's theorem, we obtain

$$\iiint_V \left(\mu\operatorname{curl}\boldsymbol{\omega} + \operatorname{grad}\wp\right)dV = \iint_{\widetilde{S}} \left(\mu\left[\mathfrak{n}\times\boldsymbol{\omega}\right] + \wp\,\mathfrak{n}\right)dS - \iint_S \left(\mu\left[\mathfrak{n}\times\boldsymbol{\omega}\right] + \wp\,\mathfrak{n}\right)dS = 0,$$

which implies the equivalence of (68) and (13), and, thus, the proposition is proved.

Appendix B. Proof of Proposition 3.

This section proves Proposition 3.

As in Appendix A, let (r,φ,z) and (s,φ,n) be the systems of the cylindrical and characteristic coordinates, respectively, and let the z-axis coincide with the body's axis of revolution.

By definition, the resultant torque, exerted on the body, is given by the integral over the body's surface S:

$$\mathbf{T} = \iint_S \left[\mathfrak{r}\times\mathbf{P}_n\right]dS = \iint_S \left[\mathfrak{r}\times\left(2\mu\left(\mathfrak{n}\cdot\operatorname{grad}\right)\mathbf{u} + \mu\left[\mathfrak{n}\times\operatorname{curl}\mathbf{u}\right] - \wp\,\mathfrak{n}\right)\right]dS, \tag{70}$$

where $\mathfrak{r} = r\,\mathbf{e}_r + z\,\mathbf{k}$ is the radius vector and $\mathbf{P}_n$ is determined by (60).

Producing the vectorial product of (64) with $\mathfrak{r}$, we obtain

$$\left[\mathfrak{r}\times(\mathfrak{n}\cdot\operatorname{grad})\,\mathbf{u}\right] = -\left[\mathfrak{r}\times\left[\mathfrak{n}\times\operatorname{curl}\mathbf{u}\right]\right] + \frac{1}{r}\left[\mathfrak{r}\times\left(\frac{\partial}{\partial\varphi}\left[\mathbf{u}\times\mathfrak{s}\right] - \frac{\partial}{\partial s}\left(r\left[\mathbf{u}\times\mathbf{e}_\varphi\right]\right)\right)\right]. \tag{71}$$

With the relationships $\frac{\partial\mathfrak{r}}{\partial\varphi} = r\,\mathbf{e}_\vartheta$ and $\frac{\partial\mathfrak{r}}{\partial s} = \frac{\partial r}{\partial s}\,\mathbf{e}_r + \frac{\partial z}{\partial s}\,\mathbf{k} \equiv \mathfrak{s}$, the second term in the right-hand side in (71) can be represented by

$$\frac{1}{r}\left(\frac{\partial}{\partial\varphi}\left(\left[\mathfrak{r}\times\left[\mathbf{u}\times\mathfrak{s}\right]\right]\right) - \frac{\partial}{\partial s}\left(r\;\left[\mathfrak{r}\times\left[\mathbf{u}\times\mathbf{e}_\varphi\right]\right]\right)\right) - \left(\left[\mathbf{e}_\varphi\times\left[\mathbf{u}\times\mathfrak{s}\right]\right] - \left[\mathfrak{s}\times\left[\mathbf{u}\times\mathbf{e}_\varphi\right]\right]\right),$$

where

$$\begin{aligned}[\mathbf{e}_\varphi \times [\mathbf{u} \times \mathfrak{s}]] - [\mathfrak{s} \times [\mathbf{u} \times \mathbf{e}_\varphi]] &= -\mathfrak{s}\,(\mathbf{u} \cdot \mathbf{e}_\varphi) + \mathbf{e}_\varphi\,(\mathbf{u} \cdot \mathfrak{s}) \\ &= [\mathfrak{n} \times \mathbf{e}_\varphi]\,(\mathbf{u} \cdot \mathbf{e}_\varphi) + [\mathfrak{n} \times \mathfrak{s}]\,(\mathbf{u} \cdot \mathfrak{s}) = [\mathfrak{n} \times \mathbf{u}]\,.\end{aligned}$$

With the last two formulas, the relationship (71) takes the from

$$\begin{aligned}[\mathfrak{r} \times (\mathfrak{n} \cdot \operatorname{grad})\,\mathbf{u}] = &-[\mathfrak{r} \times [\mathfrak{n} \times \operatorname{curl}\mathbf{u}]] - [\mathfrak{n} \times \mathbf{u}] \\ &+ \frac{1}{r}\left(\frac{\partial}{\partial\varphi}\left([\mathfrak{r} \times [\mathbf{u} \times \mathfrak{s}]]\right) - \frac{\partial}{\partial s}\left(r\ [\mathfrak{r} \times [\mathbf{u} \times \mathbf{e}_\varphi]]\right)\right).\end{aligned} \tag{72}$$

By the same reasoning as for integrating (64) over the body's surface S (see Appendix A), we obtain

$$\iint_S [\mathfrak{r} \times (\mathfrak{n} \cdot \operatorname{grad})\,\mathbf{u}]\,dS = -\iint_S \left([\mathfrak{r} \times [\mathfrak{n} \times \operatorname{curl}\mathbf{u}]] + [\mathfrak{n} \times \mathbf{u}]\right) dS. \tag{73}$$

With (73), the torque (70) reduces to

$$\mathbf{T} = -\mu \iint_S \left([\mathfrak{r} \times [\mathfrak{n} \times \operatorname{curl}\mathbf{u}]] + 2\,[\mathfrak{n} \times \mathbf{u}] + \tfrac{1}{\mu}\ [\mathfrak{r} \times \mathfrak{n}]\,\wp\right) dS. \tag{74}$$

Then, for the first equation in (69) and for an arbitrary volume V, we can write

$$\iiint_V \left[\mathfrak{r} \times \left(\operatorname{curl}(\operatorname{curl}\mathbf{u}) + \tfrac{1}{\mu}\ \operatorname{grad}\wp\right)\right] dV = 0. \tag{75}$$

Summing up two formulas

$$\begin{aligned}\operatorname{curl}[\mathfrak{a} \times \mathfrak{b}] &= (\mathfrak{b} \cdot \operatorname{grad})\,\mathfrak{a} - (\mathfrak{a} \cdot \operatorname{grad})\,\mathfrak{b} + \mathfrak{a}\operatorname{div}\mathfrak{b} - \mathfrak{b}\operatorname{div}\mathfrak{a}, \\ \operatorname{grad}(\mathfrak{a} \cdot \mathfrak{b}) &= (\mathfrak{a} \cdot \operatorname{grad})\,\mathfrak{b} + (\mathfrak{b} \cdot \operatorname{grad})\,\mathfrak{a} + [\mathfrak{a} \times \operatorname{curl}\mathfrak{b}] + [\mathfrak{b} \times \operatorname{curl}\mathfrak{a}]\,,\end{aligned}$$

which hold true for arbitrary $\mathfrak{a}$ and $\mathfrak{b}$, and rearranging terms, we obtain

$$\begin{aligned}[\mathfrak{a} \times \operatorname{curl}\mathfrak{b}] = \operatorname{curl}[\mathfrak{a} \times \mathfrak{b}] &+ \operatorname{grad}(\mathfrak{a} \cdot \mathfrak{b}) - 2\,(\mathfrak{b} \cdot \operatorname{grad})\,\mathfrak{a} \\ &- [\mathfrak{b} \times \operatorname{curl}\mathfrak{a}] - \mathfrak{a}\operatorname{div}\mathfrak{b} + \mathfrak{b}\operatorname{div}\mathfrak{a}.\end{aligned} \tag{76}$$

Substituting $\mathfrak{a} = \mathfrak{r}$ and $\mathfrak{b} = \operatorname{curl}\mathbf{u}$ into (76) and using the identities

$$\operatorname{curl}\mathfrak{r} = 0, \qquad \operatorname{div}\mathfrak{r} = 3, \qquad (\mathfrak{b} \cdot \operatorname{grad})\,\mathfrak{r} = \mathfrak{b},$$

we have

$$[\mathfrak{r} \times \operatorname{curl}(\operatorname{curl}\mathbf{u})] = \operatorname{curl}\left([\mathfrak{r} \times \operatorname{curl}\mathbf{u}] + \mathbf{u}\right) + \operatorname{grad}(\mathfrak{r} \cdot \operatorname{curl}\mathbf{u})\,. \tag{77}$$

Let V be the volume between the body's surface S and an arbitrary smooth surface $\widetilde{S}$ encompassing the body, and let $\mathfrak{n}$ be the outer normal to these surfaces. Then, with (77) and the relationship

$$[\mathfrak{r} \times \operatorname{grad}\wp] \equiv \wp\operatorname{curl}\mathfrak{r} - \operatorname{curl}(\mathfrak{r}\,\wp) = -\operatorname{curl}(\mathfrak{r}\,\wp)\,,$$

(75) takes the form

$$\begin{aligned}&\iiint_V \left[\mathfrak{r} \times \left(\operatorname{curl}(\operatorname{curl}\mathbf{u}) + \tfrac{1}{\mu}\ \operatorname{grad}\wp\right)\right] dV \\ &\quad= \iiint_V \left(\operatorname{curl}\left([\mathfrak{r} \times \operatorname{curl}\mathbf{u}] + \mathbf{u} - \tfrac{1}{\mu}\ \mathfrak{r}\,\wp\right) + \operatorname{grad}(\mathfrak{r} \cdot \operatorname{curl}\mathbf{u})\right) dV \\ &\quad= \iint_{S,\,\widetilde{S}} \left(\left[\mathfrak{n} \times \left([\mathfrak{r} \times \operatorname{curl}\mathbf{u}] + \mathbf{u} - \tfrac{1}{\mu}\ \mathfrak{r}\,\wp\right)\right] + (\mathfrak{r} \cdot \operatorname{curl}\mathbf{u})\,\mathfrak{n}\right) dS = 0,\end{aligned} \tag{78}$$

where the surface integral, being the difference between the integrals over the surfaces $\widetilde{S}$ and S, respectively, follows from Gauss's theorem.

Then, for the integrand in (78), we have

$$[\mathfrak{n} \times [\mathfrak{r} \times \operatorname{curl} \mathbf{u}]] + (\mathfrak{r} \cdot \operatorname{curl} \mathbf{u})\, \mathfrak{n} = [\mathfrak{r} \times [\mathfrak{n} \times \operatorname{curl} \mathbf{u}]] + (\mathfrak{n} \cdot \operatorname{curl} \mathbf{u})\, \mathfrak{r}, \tag{79}$$

and with the relationships

$$\frac{\partial r}{\partial s} = \frac{\partial z}{\partial n}, \qquad \frac{\partial r}{\partial n} = -\frac{\partial z}{\partial s},$$

and

$$\frac{\partial \mathfrak{r}}{\partial s} = \frac{\partial r}{\partial s}\, \mathbf{e}_r + \frac{\partial z}{\partial s}\, \mathbf{e}_\varphi \equiv \mathfrak{s}, \qquad \frac{\partial \mathfrak{r}}{\partial \varphi} = r\, \mathbf{e}_\varphi, \qquad \mathfrak{s} = -[\mathfrak{n} \times \mathbf{e}_\varphi],$$

we obtain

$$\begin{aligned}(\mathfrak{n} \cdot \operatorname{curl} \mathbf{u})\, \mathfrak{r} &= \frac{1}{r} \left(\frac{\partial}{\partial s} (r\, u_\varphi) + \frac{\partial}{\partial \varphi} \left(u_z \frac{\partial r}{\partial n} - u_r \frac{\partial z}{\partial n} \right) \right) \mathfrak{r} \\ &= \frac{1}{r} \left(\frac{\partial}{\partial s} \{(r\, u_\varphi)\, \mathfrak{r}\} + \frac{\partial}{\partial \varphi} \left\{ \left(u_z \frac{\partial r}{\partial n} - u_r \frac{\partial z}{\partial n} \right) \mathfrak{r} \right\} \right) \\ &\qquad - \left(u_\varphi\, \mathfrak{s} + \left(u_z \frac{\partial r}{\partial n} - u_r \frac{\partial z}{\partial n} \right) \mathbf{e}_\varphi \right) \\ &= \frac{1}{r} \left(\frac{\partial}{\partial s} \{(r\, u_\varphi)\, \mathfrak{r}\} + \frac{\partial}{\partial \varphi} \left\{ \left(u_z \frac{\partial r}{\partial n} - u_r \frac{\partial z}{\partial n} \right) \mathfrak{r} \right\} \right) + [\mathfrak{n} \times \mathbf{u}]\,.\end{aligned} \tag{80}$$

With (79) and (80), the surface integral over S in (78) takes the form

$$\begin{aligned}&\iint_S \left(\left[\mathfrak{n} \times \left([\mathfrak{r} \times \operatorname{curl} \mathbf{u}] + \mathbf{u} - \tfrac{1}{\mu}\, \mathfrak{r}\, \wp \right) \right] + (\mathfrak{r} \cdot \operatorname{curl} \mathbf{u})\, \mathfrak{n} \right) dS \\ &\qquad = \iint_S \left([\mathfrak{r} \times [\mathfrak{n} \times \operatorname{curl} \mathbf{u}]] + 2\, [\mathfrak{n} \times \mathbf{u}] + \tfrac{1}{\mu}\, [\mathfrak{r} \times \mathfrak{n}]\, \wp \right) dS \\ &\qquad = \iint_{\widetilde{S}} \left(\left[\mathfrak{n} \times \left([\mathfrak{r} \times \operatorname{curl} \mathbf{u}] + \mathbf{u} - \tfrac{1}{\mu}\, \mathfrak{r}\, \wp \right) \right] + (\mathfrak{r} \cdot \operatorname{curl} \mathbf{u})\, \mathfrak{n} \right) dS.\end{aligned} \tag{81}$$

Consequently, (22) follows from (74) and (81), and the proposition is proved.

REFERENCES

[1] H. Bateman and A. Erdelyi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[2] W. D. Collins, *A note on the axisymmetric Stokes flow of viscous fluid past a spherical cap*, Mathematika, 10 (1963), pp. 72–78.

[3] W. R. Dean and M. E. O'Neill, *A slow motion of viscous fluid caused by the rotation of a solid sphere*, Mathematika, 10 (1963), pp. 13–24.

[4] J. M. Dorrepaal, S. R. Majumdar, M. E. O'Neill, and K. B. Ranger, *A closed torus in Stokes flow*, Quart. J. Mech. Appl. Math., 29 (1976), pp. 381–397.

[5] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, New York, 1966.

[6] S. L. Goren and M. E. O'Neill, *Asymmetric creeping motion of an open torus*, J. Fluid Mech., 101 (1980), pp. 97–110.

[7] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Springer, New York, 1983.

[8] R. P. Kanwal, *Slow steady rotation of axially symmetric bodies in a viscous fluid*, J. Fluid Mech., 10 (1961), pp. 17–24.

[9] P. Krokhmal, *Exact solution of the displacement boundary-value problem of elasticity for a torus*, J. Engrg. Math., 44 (2002), pp. 345–368.

[10] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1945.

[11] S. R. Majumdar and M. E. O'Neill, *Asymmetric Stokes flows generated by the motion of a closed torus*, Z. Angew. Math. Phys., 30 (1979), pp. 967–982.

[12] H. Oberbeck, *Über stationäre Flüssigkeitsbewegungen mit Berücksichtigung der innere Reibung*, J. Reine Angew. Math., 81 (1876), pp. 62–80.

[13] L. E. Payne and W. H. Pell, *The Stokes flow problem for a class of axially symmetric bodies*, J. Fluid Mech., 7 (1960), pp. 529–549.

[14] W. H. Pell and L. E. Payne, *On Stokes flow about a torus*, Mathematika, 7 (1960), pp. 78–92.

[15] W. H. Pell and L. E. Payne, *The Stokes flow about a spindle*, Quart. Appl. Math., 18 (1960), pp. 257–262.

[16] M. Stimson and G. B. Jeffery, *The motion of two spheres in a viscous fluid*, Proc. Roy. Soc. London, Ser. A, 111 (1926), pp. 110–116.

[17] G. G. Stokes, *On the effect of the internal friction of fluids on the motion of pendulums*, Trans. Camb. Phil. Soc., 9 (1850), p. 8–106.

[18] H. Takagi, *Slow viscous flow due to the motion of a closed torus*, J. Phys. Soc. Japan, 35 (1973), pp. 1225–1227.

[19] A. F. Ulitko, *Vectorial Decompositions in the Three-Dimensional Theory of Elasticity*, Akademperiodika, Kiev, 2002.

[20] S. Wakiya, *Slow motion of a viscous fluid around two spheres*, J. Phys. Soc. Japan, 22 (1967), pp. 1101–1109.

[21] S. Wakiya, *On the exact solution of the Stokes equations for a torus*, J. Phys. Soc. Japan, 37 (1974), pp. 780–783.

[22] M. Zabarankin and A. F. Ulitko, *Hilbert formulas for r-analytic functions and Stokes flow about a biconvex lens*, Quart. Appl. Math., 64 (2006), pp. 663–693.

[23] M. Zabarankin and A. F. Ulitko, *Hilbert formulas for r-analytic functions in the domain exterior to spindle*, SIAM J. Appl. Math., 66 (2006), pp. 1270–1300.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 486–502

ON THE EXISTENCE OF SMALL PERIODIC SOLUTIONS FOR THE 2-DIMENSIONAL INVERTED PENDULUM ON A CART*

LUCA CONSOLINI[†] AND MARIO TOSQUES[‡]

Abstract. This paper studies the problem of controlling an inverted pendulum on a cart which has to track a given curve lying on a vertical plane in such a way that the pendulum rod does not overturn. The problem is reduced to finding sufficiently small T-periodic solutions for a special type of mathematical forced pendulum equation.

Key words. inverted pendulum, periodic solutions, tracking control

AMS subject classifications. 65N30, 74K20, 74S05

DOI. 10.1137/070683404

Introduction. This article studies the existence of T-periodic solutions for the equation

$$d\ddot{\theta} = g\sin\theta + \left\langle \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \ddot{\gamma}(t) \right\rangle, \tag{0.1}$$

which comes out from studying the problem of controlling an inverted pendulum on a cart which has to track a given curve lying on a vertical plane in such a way that the pendulum rod does not overturn (see Figure 1.1); in control theory, (0.1) is usually referred to as the system internal dynamics.

It is shown that there is a constant $k_0 \simeq 0.2672$, independent of the length of the pendulum rod, such that if $\|\ddot{\gamma}\| \leq k_0 g$, where g is the gravity acceleration, then there exist suitable initial conditions such that the functions $\theta(t), \dot{\theta}(t)$ are T-periodic and sufficiently small; in fact, they are uniformly bounded by the maximum norm of the acceleration of the reference trajectory.

The inverted pendulum on a cart is an important benchmark for nonlinear control techniques since it is a nonminimum phase problem and has been widely analyzed in various problems, under different points of view (see, for instance, [2], [3], [10], [6], [1], [8], [5].)

Equation (0.1) is related to the mathematical forced pendulum equation

$$\ddot{\theta} = \sin\theta + h(t),$$

which has been intensively studied in literature (see, for instance, [7] for a detailed survey). Many papers are focused on the existence and the multiplicity of periodic solutions (see, for example, [4], [9], [11]). Differently from these articles, this paper shows the existence of a specific sufficiently small periodic solution (in such a way that the pendulum does not overturn), gives precise bounds on its norm, and proposes a method for computing it. It is shown that the bounding term is directly proportional to the maximum norm of the acceleration of the reference trajectory.

*Received by the editors February 22, 2007; accepted for publication (in revised form) June 18, 2007; published electronically November 30, 2007.
http://www.siam.org/journals/siap/68-2/68340.html

[†]Dipartimento di Ingegneria dell'Informazione, Parco Area delle Scienze 181/a, 43100 Parma, Italy (luca.consolini@polirone.mn.it).

[‡]Dipartimento di Ingegneria Civile, Parco Area delle Scienze 181/a, 43100 Parma, Italy (mario.tosques@unipr.it).

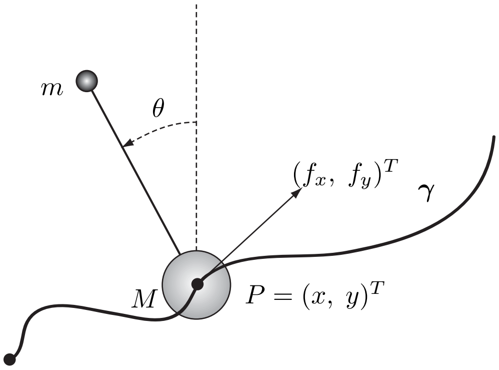


Fig. 1.1. *Inverted pendulum on a cart constrained on a curve.*

In the development of the proof, the problem of finding the right initial state is reformulated as a fixed point problem of a suitable Poincaré map associated to the internal dynamic given by (0.1). The main idea is the following: The inverted pendulum has an unstable equilibrium when the position of the cart is fixed at the origin and the pendulum rod is vertical. This unstable equilibrium corresponds to the null solution of (0.1) which is, trivially, a periodic bounded solution of the zero dynamics. Since any bounded desired reference trajectory may be prolongated to a periodic trajectory, γ may be seen as the value that the family of periodic curves $\{s\gamma\}$ assumes at $s = 1$. By prolongating the trivial null solution with a technique based on the implicit function theorem, we obtain a family of bounded periodic solutions, parameterized by $s \in [0, 1]$. The desired bounded solution is obtained by taking $s = 1$.

The paper is divided as follows. Section 1 presents the problem and section 2 the main results, section 3 contains the proof of the main theorem, and section 4 presents a computational method to drive an inverted pendulum along a given curve which is applied to a periodic spline.

The following notations will be used: For all $a, b \in \mathbb{R}$, $a \wedge b = \min\{a,\ b\}$, $a \vee b = \max\{a,\ b\}$; for all $\theta \in [0, 2\pi[$, $\tau(\theta) = (\cos\theta,\ \sin\theta)^T$; for all $x \in \mathbb{R}^2$, $\arg x = \theta$, where $\theta \in [0, 2\pi[$ is such that $x = \|x\|\tau(\theta)$; for all $x = (x_1, \dots, x_n)^T$, $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, $\langle x\,,\, y\rangle = \sum_{i=1}^n x_i y_i$, $\|x\| = \sqrt{\langle x\,,\, x\rangle}$; for all $f : [0, T] \to \mathbb{R}^2$, $\|f\|_\infty = \sup_{x\in[0,T]}\{\|f(x)\|\}$.

1. Problem formulation. Consider an inverted pendulum of mass m linked to a moving cart of mass M through a massless rod of length d; in Figure 1.1 the pendulum is represented as the smaller sphere and the cart as the bigger one. It is supposed that during the motion the control force $f(t) = \binom{f_x(t)}{f_y(t)}$ is applied on it. Let $q = (x, y, \theta) \in \mathbb{R}^3$ be the vector of generalized coordinates, where (x, y) are the coordinates of the center of mass of the moving cart and θ is the angle between the rod and the vertical axis. Given a curve γ, we want to find a control force such that, starting from $\gamma(0)$, the point P can track all of the curve γ and the rod remains close to the vertical. To find the system dynamics, let $L = T - U$ be the Lagrangian, where

$$T = \frac{1}{2}\dot{q}^T H \dot{q}$$

is the kinetic energy, the inertia matrix H is given by

$$H(q) = \begin{pmatrix} M+m & 0 & -md\cos\theta \\ 0 & M+m & -md\sin\theta \\ -md\cos\theta & -md\sin\theta & md^2 \end{pmatrix},$$

and the potential energy U is given by

$$U(q) = (M+m)gy + mgd\cos\theta.$$

The dynamic equations are derived through the Euler–Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = \tau,$$

where $\tau = (f_x,\ f_y,\ 0)^T$ is the vector of generalized forces. The resulting dynamical system has 6 states $(x,\ \dot{x},\ y,\ \dot{y},\ \theta,\ \dot{\theta})$ and is given by

$$\left\{\begin{array}{l} (M+m)\ddot{x} = md(\ddot{\theta}\cos\theta + \dot{\theta}^2\sin\theta) + f_x, \\ (M+m)\ddot{y} = md(\ddot{\theta}\sin\theta + \dot{\theta}^2\cos\theta) - (M+m)g + f_y, \\ d\ddot{\theta} = g\sin\theta + \left\langle \begin{pmatrix} \ddot{x} \\ \ddot{y} \end{pmatrix}, \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \right\rangle. \end{array}\right. \tag{1.1}$$

Then let $\gamma = (\gamma_1,\ \gamma_2) \in \mathcal{C}^2([0,T], \mathbb{R}^2)$ be a curve, and if θ is a solution of system

$$\left\{\begin{array}{l} \ddot{\theta} = d^{-1}g\sin\theta + d^{-1}\left\langle \ddot{\gamma}, \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \right\rangle, \\ \theta(0) = \theta_0, \\ \dot{\theta}(0) = \dot{\theta}_0, \end{array}\right. \tag{1.2}$$

the control force f given by

$$\left\{\begin{array}{l} f_x = (M+m)\ddot{\gamma_1} - md(\ddot{\theta}\cos\theta + \dot{\theta}^2\sin\theta), \\ f_y = (M+m)\ddot{\gamma_2} - md(\ddot{\theta}\sin\theta + \dot{\theta}^2\cos\theta) + (M+m)g \end{array}\right. \tag{1.3}$$

drives point P along all of the curve γ; that is, the solution of system (1.1), with the initial conditions $(x(0),\ \dot{x}(0))^T = \gamma(0)$, $(y(0), \dot{y}(0))^T = \dot{\gamma}(0)$, $(\theta(0), \dot{\theta}(0)) = (\theta_0, \dot{\theta}_0)$, has the property that $(x(t),\ y(t))^T = \gamma(t)$, for all $t \in [0,T]$. From these remarks, it follows that our problem is solvable if it is possible to find a suitable initial condition $(\theta_0,\ \dot{\theta}_0)$ such that the solution $(\theta,\ \dot{\theta})$ of (1.2) (which is usually referred to as the internal dynamic of the problem) remains uniformly small. Theorem 2.3 exhibits a geometric property on γ for the existence of an initial condition $(\theta_0,\ \dot{\theta}_0)$ for which there exists a periodic solution of system (1.2), which remains uniformly bounded in terms of $\|\ddot{\gamma}\|_\infty$.

2. The main results.

The following theorem gives an answer to our problem.

THEOREM 2.1 (exact tracking). *There exist positive constants $k_0 \geq 0.2672$ and $\omega_0 \leq 1.4302$ rad, independent of d and g, such that for any curve $\gamma \in C^2([0,\ T], \mathbb{R}^2)$ with*

$$\|\ddot{\gamma}\|_\infty \leq k_0 g \tag{2.1}$$

there exist an initial condition $(\theta_0,\ \dot{\theta}_0)^T$ and a control force $f \in C^0([0,T], \mathbb{R}^2)$ such that the solution of (1.1) with initial conditions $(x(0),\ y(0))^T = \gamma(0)$, $(\dot{x}(0), \dot{y}(0))^T = \dot{\gamma}(0)$, $(\theta(0), \dot{\theta}(0)) = (\theta_0, \dot{\theta}_0)$ satisfies the following properties, for all $t \in [0,T]$:

$$\begin{aligned} &(x(t),\ y(t))^T = \gamma(t), \\ &|\theta(t)| \leq \frac{\omega_0}{k_0 g}\|\ddot{\gamma}\|_\infty, \\ &|\dot{\theta}(t)| \leq \frac{(1+k_0)^2}{g(\cos\omega_0 - k)^3}\|\ddot{\gamma}\|_\infty. \end{aligned} \tag{2.2}$$

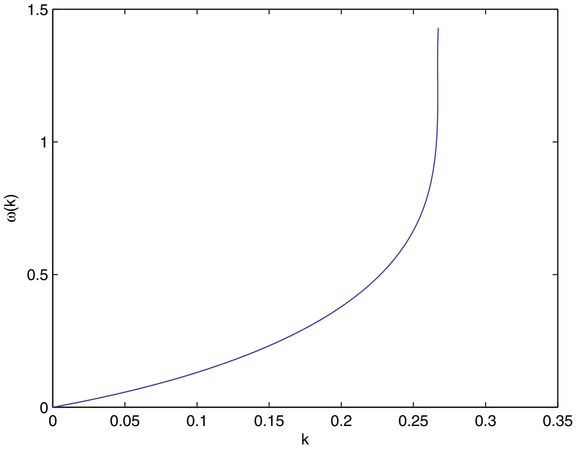


FIG. 2.1. *Function $\omega(k)$.*

In other words the curve is exactly tracked and the internal dynamics of the inverted pendulum remain bounded.

Proof. It follows directly from Theorem 2.3, by the previous considerations, since γ can be always considered the restriction on $[0, T]$ of a C^2-periodic curve. ◻

The following corollary states that it is always possible to find a reparameterization of a regular path γ that allows it to be followed with $\|(\theta, \dot{\theta})\|_\infty$ arbitrarily small.

COROLLARY 2.2 (exact path-following). *Let $\tilde{\gamma} \in C^2([0,\ \Lambda], \mathbb{R}^2)$ be an arc-length parameterized curve (that is, $\|\dot{\tilde{\gamma}}(\lambda)\| = 1$). For any $\sigma > 0$ there exist a bijection $\lambda(t) \in C^2([0,\ T],\ [0, \Lambda])$, an initial condition $(\theta_0,\ \dot{\theta}_0)^T$, and a control force $f \in C^0([0, T], \mathbb{R}^2)$ such that the solution of problem (1.1) with initial conditions $(x(0),\ y(0))^T = \gamma(0)$, $(\dot{x}(0),\ \dot{y}(0))^T = \frac{d(\tilde{\gamma}\circ\lambda)}{dt}(0)$, $(\theta(0),\ \dot{\theta}(0)) = (\theta_0,\ \dot{\theta}_0)$ satisfies the following properties for all $t \in [0,\ T]$:*

$$(x(t),\ y(t))^T = \tilde{\gamma}(\lambda(t)),\ \|(\theta(t),\ \dot{\theta}(t))^T\| \le \sigma.$$

In other words, for any arbitrary small constant σ, the exact path-following problem can always be solved by keeping the internal dynamics bounded by σ.

Proof. Set $\gamma = \tilde{\gamma} \circ \lambda$, where $\lambda(t)$ is a $C^2([0,\ T], [0,\ \Lambda])$ bijection; then

$$\|\ddot{\gamma}(t)\| = \sqrt{\ddot{\lambda}^2 + (\kappa(\lambda(t))\dot{\lambda}^2(t))^2},$$

where $\kappa(\lambda) = \frac{d}{d\lambda} \arg(\dot{\tilde{\gamma}}(\lambda))$ is the curvature of $\tilde{\gamma}(\lambda)$. Note that $\ddot{\lambda}(t)$ is the linear acceleration, and $\kappa(\lambda)\dot{\lambda}^2$ is the centripetal acceleration. Then the results follow directly from Theorem 2.1, choosing the reparameterization $\lambda(t)$ in a suitable way. ◻

THEOREM 2.3 (main theorem). *There exist $\bar{k} > 0$ and a continuous strictly increasing function $\omega : [0, \bar{k}[\to [0, \frac{\pi}{2}]$, with $\omega(0) = 0$ (see Figure 2.1), such that, for any $k \in [0, \bar{k}[$, any $d > 0$, and any periodic curve $\gamma \in C^2(\mathbb{R},\ \mathbb{R}^2)$ of period T with*

$$\|\ddot{\gamma}\|_\infty \le kg, \tag{2.3}$$

there exists an initial condition $(\theta_0,\ \dot{\theta}_0)^T$ *such that the solution of system*

$$\left\{\begin{array}{l}\ddot{\theta} = d^{-1} g \sin\theta + d^{-1}\left\langle \ddot{\gamma}\, ,\, \left(\begin{array}{c}\cos\theta\\ \sin\theta\end{array}\right)\right\rangle,\\ \theta(0)=\theta_0,\\ \dot{\theta}(0)=\dot{\theta}_0,\end{array}\right. \tag{2.4}$$

is periodic of period T*, and for all* $t \in \mathbb{R}$

$$\begin{aligned}|\theta(t)| &\le \frac{\omega(k)}{kg}\|\ddot{\gamma}\|_\infty,\\ |\dot{\theta}(t)| &\le \frac{(1+k)^2}{g(\cos\omega(k)-k)^3}\|\ddot{\gamma}\|_\infty.\end{aligned} \tag{2.5}$$

Furthermore

$$\bar{k} > 0.2672 \text{ and } \omega(0.2672) = 1.4302. \tag{2.6}$$

3. Proof of the main theorem. First of all, we define the function $\omega(k)$. For any $k \ge 0$, set $\mathcal{R}_k = \{(s,\omega) | \cos\omega - sk > 0\}$ and

$$\psi_k(s,\omega) = \frac{k(1+sk)^2}{(\cos\omega - sk)^3}\ \forall (s,\omega)\in\mathcal{R}_k.$$

Let ξ_k be the solution of the differential problem

$$\left\{\begin{array}{l}\dot{\xi}_k(s) = \psi_k(s,\xi_k(s)),\\ \xi_k(0)=0,\end{array}\right. \tag{3.1}$$

defined on the maximal interval of existence $[0,\bar{s}(k)[$. By the definition of ψ_k we get that $\xi_k : [0,\bar{s}(k)[\to [0,\frac{\pi}{2}]$ is strictly increasing, and $\bar{s}(k)$ is a strictly decreasing continuous function such that $\bar{s}(0)=+\infty$, $\bar{s}(k)\le\frac{1}{k}$, if $k>0$. Therefore there exists $\bar{k} : 0<\bar{k}\le 1$ such that $\bar{s}(\bar{k})=1$, and let $\omega : [0,\bar{k}[\to[0,\frac{\pi}{2}]$ be the function defined by $\omega(k)=\xi_k(1)$. By numerical computation it is $\bar{k}>0.2672$ with $\omega(0.2672)=1.4302$. Remark that $0\le 1\le \bar{s}(k)$ for all $k\in[0,\bar{k}[$, since $\bar{s}(k)$ is strictly decreasing and $\bar{s}(\bar{k})=1$. Set $k\in[0,\bar{k}[$, and suppose that $\|\ddot{\gamma}\|_\infty = kg$. For every $\tau, s\in\mathbb{R}$, for all $y\in\mathbb{R}^2$, let $x(t,\tau,s,y)$ be the solution of the problem

$$\left\{\begin{array}{l}\dot{x} = F(t,s,x),\ \forall t\in\mathbb{R},\\ x(\tau)=y,\end{array}\right. \tag{3.2}$$

where $F:\mathbb{R}\times\mathbb{R}\times\mathbb{R}^2\to\mathbb{R}^2$ is the C^2 map defined by

$$F(t,s,x) = \left(x_2,\ d^{-1}g\sin x_1 + s d^{-1}\left\langle \ddot{\gamma}(t)\, ,\, \left(\begin{array}{c}\cos x_1\\ \sin x_1\end{array}\right)\right\rangle\right)^T,\ \forall (t,s,x)\in\mathbb{R}\times\mathbb{R}\times\mathbb{R}^2, \tag{3.3}$$

and remark that, for every $(s,x)\in\mathbb{R}\times\mathbb{R}^2$, the map $t \rightsquigarrow F(t,s,x)$ is periodic of period T and that $(\theta,\dot{\theta})^T$ is a solution of (2.4) if and only if $x=(\theta,\dot{\theta})^T$ is a solution of (3.2) with $s=1$, $\tau=0$, and $y=(\theta_0,\dot{\theta}_0)$. Let $P:\mathbb{R}\times\mathbb{R}\times\mathbb{R}^2\to\mathbb{R}^2$ be the Poincaré map associated to problem (3.2); that is,

$$P(\tau,s,y) = x(T+\tau,\tau,s,y), \tag{3.4}$$

and set $\mathcal{P} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$ the map defined by

$$\mathcal{P}(\tau, s, y) = P(\tau, s, y) - y.$$

Set $l = d^{-1}g$, and remark that $F(t, 0, x) = (x_2,\ l \sin x_1)^T$. Therefore for all $\tau \in \mathbb{R}$, $x(t, \tau, 0, 0) = 0$, for all $t \in \mathbb{R}$, which is clearly a periodic solution of period T, and therefore

$$\mathcal{P}(\tau, 0, 0) = 0\ \forall \tau \in \mathbb{R}.$$

The following property holds:

$$\partial_y \mathcal{P}(\tau, 0, 0) = \partial_y P(\tau, 0, 0) - I \text{ is invertible } \forall \tau \in \mathbb{R}; \tag{3.5}$$

that is, $\partial_y P(\tau, 0, 0)$ has two eigenvalues different from 1. In fact for all $i = 1, 2$

$$\partial_{y_i} P(\tau, s, y) = \partial_{y_i} x(T + \tau, \tau, s, y),$$

where $\partial_{y_i} x(t, \tau, s, y)$ is the solution ϕ_i of the problem

$$\begin{cases} \dot{\phi}_i = \partial_x F(t, s, x(t, \tau, s, y))\phi_i \ \forall t \in \mathbb{R}, \\ \phi_i(\tau) = e_i \end{cases} \tag{3.6}$$

(where $e_1 = (1, 0)$, $e_2 = (0, 1)$); therefore $\partial_y P(\tau, s, y) = \Phi^y_s(T + \tau, \tau)$, where $\Phi^y_s(t,\ \tau)$ is the matrix solution of system

$$\begin{cases} \dot{\Phi} = \partial_x F(t, s, x(t, \tau, s, y))\Phi \ \forall t \in \mathbb{R}, \\ \Phi(\tau) = I, \end{cases} \tag{3.7}$$

where I is the identity matrix. Since $x(t, \tau, 0, 0) = 0$ for all $t \in \mathbb{R}$, it is $\partial_x F(t, 0, x(t, \tau, 0, 0)) = \left(\begin{smallmatrix} 0 & 1 \\ l & 0 \end{smallmatrix}\right)$ which has the eigenvalues $\sqrt{l}$ and $-\sqrt{l}$ and normalized eigenvectors:

$$W = \frac{1}{\sqrt{l + 1}} \left(\begin{pmatrix} 1 \\ \sqrt{l} \end{pmatrix}, \begin{pmatrix} -1 \\ \sqrt{l} \end{pmatrix} \right).$$

Then $\Phi^0_0(t,\ \tau) = W\left(\begin{smallmatrix} e^{\sqrt{l}t} & 0 \\ 0 & e^{-\sqrt{l}t} \end{smallmatrix}\right)W^{-1}$ for all $t, \tau \in \mathbb{R}$, which implies that for all $t, \tau \in \mathbb{R}$, $\Phi^0_0(t, \tau)$ has the eigenvalues $e^{\sqrt{l}t}$ and $e^{-\sqrt{l}t}$ and constant eigenvectors given by the columns of W; therefore

$$\partial_y \mathcal{P}(\tau, 0, 0) = \Phi^0_0(T + \tau, \tau) - I\ \forall t \in \mathbb{R},$$

and property (3.5) holds. Since $\{(\tau, 0, 0) | 0 \le \tau \le T\}$ is a compact subset of $\mathbb{R}^3$, by the implicit function theorem, we can find $\rho > 0$ and a C^1 map $y : [0, T] \times [-\rho,\ \rho] \to \overline{B((0, 0), \rho)}$ (where $\overline{B((0, 0), \rho)}$ denotes the closed ball in $\mathbb{R}^2$ of center $(0, 0)$ and radius ρ), represented in Figure 3.1, such that

$$\begin{aligned} &y(\tau, 0) = 0\ \forall \tau \in [0, T],\ \mathcal{P}(\tau, s, y(\tau, s)) = 0\ \forall (\tau, s) \in [0, T] \times [-\rho,\ \rho], \\ &\{(\tau, s, y(\tau, s)) | (\tau, s) \in [0, T] \times [-\rho,\ \rho]\} \\ &= \{(\tau, s, y) \in [0, T] \times [-\rho, \rho] \times \overline{B((0, 0), \rho)} \text{ such that } \mathcal{P}(\tau, s, y) = 0\}; \end{aligned} \tag{3.8}$$

that is (as stated in the implicit function theorem), $y(\tau, s)$ is the only solution of $\mathcal{P}(\tau, s, y) = 0$, inside $[0, T] \times [-\rho, \rho] \times \overline{B((0, 0), \rho)}$,

$$\partial_y \mathcal{P}(\tau, s, y(\tau, s)) \text{ is invertible } \forall (\tau, s) \in [0, T] \times [-\rho, \rho]. \tag{3.9}$$

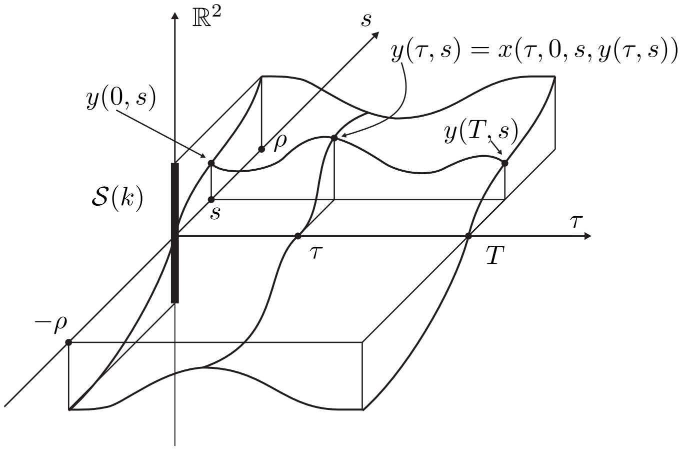


Fig. 3.1. *The implicit function method.*

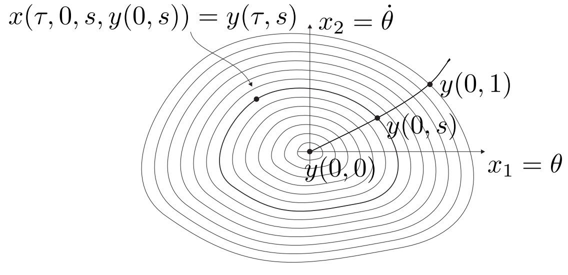


Fig. 3.2. *The family of periodic orbits* $x(t,0,s,y(0,s))$, $0\le s\le 1$.

The aim of the following is to prove that map $y(\tau,s)$ can be prolongated at least to $s=1$ for all $\tau\in[0,T]$, see Figure 3.2. It is used as a maximality procedure that requires the definition of suitable sets and functions, whose use is related to Lemmas 3.1 and 3.2.

Set $\mathcal{S}(k)=\{x=(x_1,x_2)\big||x_1|\le\omega(k)\}$. Let $\bar\rho$ be the supremum of the ρ's such that there exists a unique $y\in C^1([0,T]\times[-\rho,\rho],\ \mathcal{S}(k))$, with the properties (3.8), (3.9). Since it can be supposed that $\overline{B((0,0),\rho)}\subset\mathcal{S}(k)$ unless of decreasing ρ if necessary, we have that $\bar\rho>0$ by the previous step. Then by the definition of $\bar\rho$ we have that there exists a unique $y\in C^1([0,T]\times]-\bar\rho,\bar\rho[,\mathcal{S}(k))$ such that the properties (3.8), (3.9) are verified on $[0,T]\times]-\bar\rho,\bar\rho[$. But by the group property and the periodicity in t of $x(t,0,s,y(0,s))$

$$x(T+\tau,\tau,s,x(\tau,0,s,y(0,s)))=x(T+\tau,0,s,y(0,s))=x(\tau,0,s,y(0,s)),$$

then it is

$$y(\tau,s)=x(\tau,0,s,y(0,s))\ \forall(\tau,s)\in[0,T]\times]-\bar\rho,\bar\rho[,$$

since $y(\tau,s)$ is the only solution y in $\mathcal{S}(k)$ of the equation $y=x(\tau+T,\tau,s,y)$.

Therefore, by the group property, for all $s\in]-\bar\rho,\bar\rho[$, for all $\tau_1,\tau_2\in[0,T]$

$$\begin{aligned}x(t,\tau_1,s,y(\tau_1,s))&=x(t,\tau_1,s,x(\tau_1,0,s,y(0,s)))\\&=x(t,\tau_2,s,x(\tau_2,0,s,y(0,s)))=x(t,\tau_2,s,y(\tau_2,s));\end{aligned}$$

in other words, for all $t \in \mathbb{R}$ for all $s \in]-\bar{\rho}, \bar{\rho}[$, $x(t, \tau, s, y(\tau, s))$ is independent of $\tau \in [0, T]$, and in particular:

$$x(t, \tau, s, y(\tau, s)) = x(t, 0, s, y(0, s)) = y(t, s) \; \forall (t, \tau, s) \in [0, T]^2 \times]-\bar{\rho}, \bar{\rho}[. \tag{3.10}$$

Furthermore (3.8) imply that

$$\partial_y \mathcal{P}(\tau, s, y(\tau, s)) \partial_s y(\tau, s) + \partial_s \mathcal{P}(\tau, s, y(\tau, s)) = 0,$$

which can be rewritten in the form

$$\partial_s y(\tau, s) = \partial_y P(\tau, s, y(\tau, s)) \partial_s y(\tau, s) + \partial_s P(\tau, s, y(\tau, s)) \; \forall (\tau, s) \in [0, T] \times]-\bar{\rho}, \bar{\rho}[, \tag{3.11}$$

since for all $(\tau, s, y) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^2$, $\partial_s P(\tau, s, y) = Z_s^y(T + \tau, \tau)$, where $Z_s^y(t, \tau)$ is the solution of the following system:

$$\begin{cases} \dot{Z} = \partial_x F(t, s, x(t, \tau, s, y))Z + \partial_s F(t, s, x(t, \tau, s, y)), \\ Z(\tau) = 0. \end{cases}$$

Therefore (3.11) says that $\partial_s y(\tau, s)$ verifies the following property:

$$\partial_s y(\tau, s) = \Phi_s^{y(\tau,s)}(T + \tau, \tau) \partial_s y(\tau, s) + Z_s^{y(\tau,s)}(T + \tau, \tau) \; \forall (\tau, s) \in [0, T] \times]-\bar{\rho}, \bar{\rho}[. \tag{3.12}$$

Since $x(t, \tau, s, y(\tau, s)) = y(t, s)$ by property (3.10), we can set for all $t \in \mathbb{R}$, for all $s \in]-\bar{\rho}, \bar{\rho}[$, for all $\tau \in [0, T]$, for all $t \in \mathbb{R}$,

$$\begin{cases} \partial_x F(t, s, x(t, \tau, s, y(\tau, s))) = \begin{pmatrix} 0 & 1 \\ f(t, s, y_1(t, s)) & 0 \end{pmatrix} = A_s(t), \\ \partial_s F(t, s, x(t, \tau, s, y(\tau, s))) = \begin{pmatrix} 0 \\ b(t, y_1(t, s)) \end{pmatrix} = B_s(t), \end{cases} \tag{3.13}$$

where

$$\begin{cases} f(t, s, x_1) = d^{-1} \left[g \cos x_1 + s \left\langle \ddot{\gamma}(t), \begin{pmatrix} -\sin x_1 \\ \cos x_1 \end{pmatrix} \right\rangle \right], \\ b(t, x_1) = d^{-1} \left\langle \ddot{\gamma}(t), \begin{pmatrix} \cos x_1 \\ \sin x_1 \end{pmatrix} \right\rangle. \end{cases} \tag{3.14}$$

$A_s(t), B_s(t)$ are periodic of period T, for all $s \in]-\bar{\rho}, \; \bar{\rho}[$, $\ddot{\gamma}(t)$ and $y(t, s)$ being periodic functions of t of period T.

Since the solution of this family of periodic time-varying systems

$$\begin{cases} \dot{\xi} = A_s(t)\xi + B_s(t), \\ \xi(\tau) = z \end{cases}$$

is given by $\xi(t) = \Phi_s(t, \tau)z + \int_\tau^t \Phi_s(t, p)B_s(p)dp$, where $\Phi_s(t, \tau) \triangleq \Phi_s^{y(\tau,s)}(t, \tau)$, property (3.12) can be rewritten in the form: For all $\tau \in [0, T]$, for all $s \in]-\bar{\rho}, \; \bar{\rho}[$

$$\partial_s y(\tau, s) = \Phi_s(T + \tau, \tau) \partial_s y(\tau, s) + \int_\tau^{T+\tau} \Phi_s(T + \tau, p) B_s(p) dp. \tag{3.15}$$

Our aim is to show that $\bar{\rho} \geq 1$. Suppose, by contradiction, that $\bar{\rho} < 1$, and set, for brevity, $f_s(t) = f(t, s, y_1(t, s))$. Since $\|\ddot{\gamma}\|_\infty = kg$ and $l = d^{-1}g$ it is by (3.14) that, for all $s \in]-\bar{\rho}, \bar{\rho}[$, $f_s(t)$ is periodic of period T, and, for all $t \in \mathbb{R}$, for all $s \in]-\bar{\rho}, \; \bar{\rho}[$,

$$f_1(s, k, l) = l(\cos(\|y_1(\cdot, s)\|_\infty) - sk) \leq f_s(t) \leq l(1 + sk) = f_2(s, k, l), \tag{3.16}$$

where $\|y_1(\cdot,s)\|_\infty = \sup_{0\le t\le T}|y_1(t,s)|$. Then for every $s \in]-\bar\rho,\bar\rho[$, we can apply the corollary to Lemma 3.2, with $A(t) = A_s(t)$ and $B(t) = B_s(t)$, (3.38) being verified by (3.15).

Then it is $\|B_s(t)\| \le d^{-1}\|\ddot\gamma\|_\infty = lk$, for all $t \in \mathbb{R}$, for all $s \in]-\bar\rho,\bar\rho[$, and, by (3.39), for all $(\tau,s) \in [0,T]\times]-\bar\rho,\bar\rho[$:

$$(3.17)\qquad |\partial_s y_1(\tau,s)| \le lk\frac{\sqrt{f_2}(1+f_2)}{f_1(1+f_1)}\left(\sqrt{f_1}\wedge\left(\sqrt{f_2}\frac{1+f_1}{1+f_2}\right)\right)^{-1},$$

$$(3.18)\qquad |\partial_s y_2(\tau,s)| \le lk\frac{\sqrt{1+f_2}}{\sqrt{1+f_1}}\frac{f_2}{f_1}\left(\sqrt{f_1}\wedge\left(\sqrt{f_2}\frac{1+f_1}{1+f_2}\right)\right)^{-1}.$$

Because $\frac{1+f_2}{1+f_1} \le \frac{f_2}{f_1}$, $\frac{1+f_1}{1+f_2} \ge \frac{f_1}{f_2}$ since $f_1 \le f_2$,

$$\frac{\sqrt{f_2}(1+f_2)}{f_1(1+f_1)}\left(\sqrt{f_1}\wedge\left(\sqrt{f_2}\frac{1+f_1}{1+f_2}\right)\right)^{-1} \le \frac{f_2^2}{f_1^3},$$

which implies, by (3.17), (3.18), (3.16) that, for all $(\tau,s) \in [0,T]\times]-\bar\rho,\bar\rho[$,

$$(3.19)\qquad |\partial_s y_1(\tau,s)| \le lk\frac{f_2^2}{f_1^3} \le \frac{k(1+sk)^2}{(\cos\|y_1(\cdot,s)\|_\infty - sk)^3} = \psi_k(s,\|y_1(\cdot,s)\|_\infty).$$

Therefore by the definition of ξ_k (see (3.1)) and the comparison lemma, we obtain that

$$(3.20)\qquad \|y_1(\cdot,s)\|_\infty \le \xi_k(s)\ \forall s \in]-\bar\rho,\bar\rho[.$$

Since $\|y_1(\cdot,s)\|_\infty \le \omega(k) = \xi_k(1)$ for all $s \in [-\bar\rho,\ \bar\rho]$,

$$(3.21)\qquad |\partial_s y_1(\tau,s)| \le \frac{k(1+sk)^2}{(\cos\omega(k) - sk)^3} \le \frac{k(1+k)^2}{(\cos\omega(k)-k)^3} = \psi_k(\xi_k(1),1),$$

and

$$(3.22)\qquad |\partial_s y_2(\tau,s)| \le lk\sqrt{\frac{1+f_1}{1+f_2}}\frac{f_2^2}{f_1^3} \le lk\frac{f_2^2}{f_1^3} \le \psi_k(\xi_k(1),1).$$

Remark that $\xi_k(s)$ and $\psi_k(\xi_k(s),s)$ are well defined on 1 since $0 \le k < \bar k$ (see the definition of $\bar k$ at the beginning of the proof). Then by (3.21) and (3.22), for all $\tau \in [0,T]$ the map $s \rightsquigarrow y(\tau,s)$ may be prolongated to a Lipschitz map defined on $[-\bar\rho,\bar\rho]$ and $y(\tau,s) \in S(k)$, $(\tau,s) \in [0,T]\times[-\bar\rho,\ \bar\rho]$. Furthermore $\exists\epsilon > 0:$

$$\|y_1(\cdot,s)\|_\infty \le \omega(k) - \epsilon\ \forall s \in [-\bar\rho,\bar\rho]$$

(by (3.20), since $\xi_k(\bar\rho) < \xi(1) = \omega(k)$, ξ being strictly increasing and $\bar\rho < 1$ by the absurd hypothesis). Moreover, by continuity, $\mathcal{P}(\tau,\pm\bar\rho,y(\tau,\pm\bar\rho)) = 0$ for all $\tau \in [0,T]$, and by (3.27) of Lemma 3.1, $\partial_y\mathcal{P}(\tau,\pm\bar\rho,y(\tau,\pm\bar\rho))$ is invertible for all $\tau \in [0,T]$. Therefore, by applying the implicit function theorem to the curves $\tau \rightsquigarrow y(\tau,\pm\bar\rho)$, the definition of $\bar\rho$ is contradicted.

Then $\bar\rho \ge 1$, and, taking $s = 1$, $\mathcal{P}(0,1,y(0,1)) = 0$; that is, there exists an initial condition $y(0,1)$ such that the solution $x(t,0,1,y(0,1))$ is periodic of period T, with the properties for all $t \in [0,T]$:

$$|x_1(t,0,1,y(0,1))| = |y_1(t,1)| \le \omega(k),$$

$$|x_2(t,0,1,y(0,1))| = |y_2(t,1)| \le \frac{k(1+k)^2}{(\cos\omega(k)-k)^3}.$$

This completes the proof if $\|\ddot{\gamma}\|_\infty = kg$. Suppose now that $\|\ddot{\gamma}\|_\infty < kg$. Remark that for all $\alpha : 0 \leq \alpha \leq 1$, $\omega(\alpha k) \leq \alpha\omega(k)$, and $\frac{\alpha k(1+\alpha k)^2}{(\cos\omega(\alpha k)-\alpha k)^3} \leq \alpha\frac{k(1+k)^2}{(\cos\omega(k)-k)^3}$, which implies (2.5) if $\alpha = \frac{\|\ddot{\gamma}\|_\infty}{kg}$, and the theorem has been proved. □

The following lemma studies the properties of the eigenvalue and eigenvectors of the fundamental matrix Φ associated to an hyperbolic matrix having the form (3.23) with the suitable control (3.24).

LEMMA 3.1. *Let $A(t)$ be the following (2×2) continuous matrix:*

$$A(t) = \begin{pmatrix} 0 & 1 \\ f(t) & 0 \end{pmatrix}, \tag{3.23}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a continuous function such that

$$0 < f_1 \leq f(t) \leq f_2 < +\infty, \tag{3.24}$$

where f_1 and f_2 are given constants.

Set $\tau \geq 0$, and let $\Phi(t)$ be the solution to

$$\begin{cases} \dot{\Phi} = A(t)\Phi \ \forall t \in [0, +\infty[, \\ \Phi(\tau) = I, \end{cases} \tag{3.25}$$

and let, for all $i = 1, 2$, $\mu_i(t)$ and $w_i(t)$ be, respectively, the eigenvalues and the corresponding normalized eigenvectors of $\Phi(t)$ (i.e., for all $i = 1, 2$, $\Phi(t)w_i(t) = \mu_i(t)w_i(t)$ and $\|w_i(t)\| = 1$). Then

$$w_i(t) \in \Sigma_i \ \forall t \geq \tau, \tag{3.26}$$

$$\mu_2(t) \leq e^{-\lambda_0(t-\tau)} < e^{\lambda_0(t-\tau)} \leq \mu_1(t) \ \forall t \geq \tau, \tag{3.27}$$

where Σ_1 and Σ_2 are the cones defined as follows:

$$\Sigma_1 = \left\{ x \in \mathbb{R}^2 | \left\langle x\,, \begin{pmatrix} 1 \\ \sqrt{f_1} \end{pmatrix}^\perp \right\rangle \geq 0, \ \left\langle x\,, \begin{pmatrix} 1 \\ \sqrt{f_2} \end{pmatrix}^\perp \right\rangle \leq 0 \right\},$$

$$\Sigma_2 = \left\{ x \in \mathbb{R}^2 | \left\langle x\,, \begin{pmatrix} 1 \\ -\sqrt{f_1} \end{pmatrix}^\perp \right\rangle \leq 0, \ \left\langle x\,, \begin{pmatrix} 1 \\ -\sqrt{f_2} \end{pmatrix}^\perp \right\rangle \geq 0 \right\},$$

and $\lambda_0(f_1, f_2) = \sqrt{f_1} \wedge (\sqrt{f_2}\frac{1+f_1}{1+f_2})$.

Proof. Clearly it can be assumed that $\tau = 0$. First of all, it is shown that Σ_1 is positively invariant with respect to the vector field $A(t)x$. To this end it is sufficient to show that

$$\left\langle A(t)x\,, \begin{pmatrix} 1 \\ \sqrt{f_1} \end{pmatrix}^\perp \right\rangle \geq 0 \ \forall x \in P_1, \forall t \geq 0,$$

$$\left\langle A(t)x\,, \begin{pmatrix} 1 \\ \sqrt{f_2} \end{pmatrix}^\perp \right\rangle \leq 0 \ \forall x \in P_2, \forall t \leq 0,$$

where $P_1 = \{x \in \mathbb{R}^2 | x_1 \geq 0, \ x_2 \leq x_1\sqrt{f_1}\}$, $P_2 = \{x \in \mathbb{R}^2 | x_1 \geq 0, \ x_2 \geq x_1\sqrt{f_2}\}$.

In fact if $x \in P_1$, then

$$\left\langle A(t)x\,,\begin{pmatrix}1\\ \sqrt{f_1}\end{pmatrix}^{\perp}\right\rangle = \left\langle x\,,\, A^T(t)\begin{pmatrix}-\sqrt{f_1}\\ 1\end{pmatrix}\right\rangle$$
$$= \left\langle\begin{pmatrix}x_1\\ x_2\end{pmatrix}\,,\begin{pmatrix}f(t)\\ -\sqrt{f_1}\end{pmatrix}\right\rangle = x_1 f(t) - x_2\sqrt{f_1} \geq x_1(f(t)-f_1) \geq 0.$$

If $x \in P_2$, then

$$\left\langle A(t)x\,,\begin{pmatrix}1\\ \sqrt{f_2}\end{pmatrix}^{\perp}\right\rangle = x_1 f(t) - x_2\sqrt{f_2} \leq x_1(f(t)-f_2(t)) \leq 0.$$

Therefore Σ_1 is positively invariant with respect to $A(t)x$, and analogously it is proved that Σ_2 is positively invariant with respect to the vector field $-A(-t)x$.

Moreover, for all $i = 1, 2$, the eigenvectors $w_i(t)$ of $\Phi(t)$ satisfy $w_i(t) \in \Sigma_i$ for all $t \geq 0$. In fact set $\alpha_i = \arcsin\frac{\sqrt{f_i}}{\sqrt{1+f_i}}$, $i = 1, 2$, and define the continuous map $G_t : [\alpha_1,\ \alpha_2] \to \mathbb{R}$ by

$$G_t(\alpha) = \arg(\Phi(t)\tau(\alpha)) - \alpha$$

(recall that $\tau(\alpha) = (\cos(\alpha), \sin(\alpha))^T$). Because of the positive invariance of Σ_1, $G_t(\alpha_1) \geq 0$ and $G_t(\alpha_2) \leq 0$; therefore, by Bolzano's theorem there exists $\bar{\alpha} \in [\alpha_1,\ \alpha_2]$ such that $G_t(\bar{\alpha}) = 0$; that is, there exists $\mu(t)$ such that

$$\Phi(t)\tau(\bar{\alpha}(t)) = \mu(t)\tau(\bar{\alpha}(t)),$$

and therefore $w_1(t) = \tau(\bar{\alpha}(t))$ is a normalized eigenvector. Analogously, by using the positive invariance of Σ_2 with respect to $-A(-t)$, it is verified that the other eigenvector $w_2(t)$ of $\Phi(t)$ belongs to Σ_2.

Given any $\bar{x} \in \Sigma_1\backslash\{0\}$, let $x(t) = \Phi(t)\bar{x}$ for all $t \geq 0$. Since $x(t) \in \Sigma_1$, by the positive invariance of Σ_1, there exists functions $\rho(t)$ and $\beta(t)$ such that

$$x(t) = \rho(t)\begin{pmatrix}1\\ \beta(t)\end{pmatrix},\ \rho(t) \geq 0,\ \sqrt{f_1} \leq \beta(t) \leq \sqrt{f_2}\ \forall t \geq 0.$$

Since $\|x(t)\| = \rho(t)\sqrt{1+\beta^2(t)}$ for all t, with $\|x(t)\| > 0$,

$$\frac{d\|x(t)\|}{dt} = \frac{\langle \dot{x}(t)\,,\, x(t)\rangle}{\|x(t)\|} = \frac{\langle A(t)x(t)\,,\, x(t)\rangle}{\rho(t)\sqrt{1+\beta^2(t)}}$$
$$= \frac{\rho(t)\beta(t)}{\sqrt{1+\beta^2(t)}}(1+f(t)) = \|x(t)\|\frac{\beta(t)}{1+\beta(t)^2}(1+f(t)).$$

Since $\sqrt{f_1} \leq \beta(t) \leq \sqrt{f_2}$, for all $t \geq 0$

$$\frac{\beta(t)}{1+\beta^2(t)} \geq \frac{\sqrt{f_1}}{1+f_1} \wedge \frac{\sqrt{f_2}}{1+f_2},$$

and

$$\frac{d\|x(t)\|}{dt} \geq \left(\frac{\sqrt{f_1}}{1+f_1} \wedge \left(\frac{\sqrt{f_2}}{1+f_2}(1+f_1)\right)\right)\|x(t)\|$$
$$= \left(\sqrt{f_1} \wedge \left(\sqrt{f_2}\frac{1+f_1}{1+f_2}\right)\right)\|x\| = \lambda_0(f_1, f_2)\|x\|,$$

and therefore $\|x(t)\| \geq e^{\lambda_0 t}\|x(0)\|$. Finally let $x(t) = \Phi(t)w_1(\tau)$, where $w_1(\tau)$ is the eigenvector of $\Phi(t)$ belonging to Σ_1. Since $w_1(t) \in \Sigma_1$ for all $t \geq 0$, by the previous reasoning,

$$\|x(t)\| \geq e^{\lambda_0 t}\|w(\tau)\|,$$

which implies, since $x(\tau) = \mu_1(\tau)w(\tau)$ for all $\tau \geq 0$ that:

$$\mu_1(t) \geq e^{\lambda_0 t}\ \forall t \geq 0.$$

Therefore (3.27) holds since $\mu_2(t) = \mu_1(t)^{-1}$ is the trace of $A(t) = 0$. □

The following lemma gives an estimate on the fixed points of the solution for a nonhomogeneous time-dependent linear system associated to an hyperbolic matrix $A(t)$ having the form (3.23) with controls (3.24) on the coefficients.

LEMMA 3.2. *Let $A(t)$ be a continuous 2×2 matrix defined on $[0, +\infty[$, satisfying the hypotheses of Lemma 3.1, and let $B(t) \in \mathbb{R}^2$ be a continuous vector on $[0, +\infty[$. Let $\xi \in \mathbb{R}^2$ be such that there exists $\tau \geq 0$ and $T > 0$ for which*

$$\xi = \Phi(T + \tau, \tau)\xi + \int_\tau^{T+\tau} \Phi(T + \tau, p)B(p)dp,$$

where $\Phi(t, \tau)$ is a solution of (3.25). Then

$$(3.28)\qquad \begin{aligned} |\xi_1| &\leq c_1 \frac{1 + e^{-\lambda_0 T}}{\lambda_0}\|B\|_{[\tau,\ T+\tau]}, \\ |\xi_2| &\leq c_2 \frac{1 + e^{-\lambda_0 T}}{\lambda_0}\|B\|_{[\tau,\ T+\tau]}, \end{aligned}$$

where

$$(3.29)\qquad \begin{cases} \lambda_0(f_1, f_2) = \sqrt{f_1} \wedge \left(\sqrt{f_2}\dfrac{1 + f_1}{1 + f_2}\right), \\ c_1(f_1, f_2) = \dfrac{\sqrt{f_2}}{f_1}\dfrac{1 + f_2}{1 + f_1}, \\ c_2(f_1, f_2) = \dfrac{f_2}{f_1}\dfrac{\sqrt{1 + f_2}}{\sqrt{1 + f_1}}, \end{cases}$$

and $\|B\|_{[\tau,T+\tau]} = \max\{\|B(t)\| | \tau \leq t \leq T + \tau\}$.

Proof. Without loss of generality, assume $\tau = 0$, and therefore suppose that ξ satisfies

$$(3.30)\qquad \xi = \Phi(T, 0)\xi + \int_0^T \Phi(T, \tau)B(\tau)d\tau.$$

Set $W(t, \tau) = (w_1(t, \tau), w_2(t, \tau))$, where $w_i(t, \tau)$ are the normalized eigenvectors and $\mu_i(t)$ the eigenvalues of $\Phi(t, \tau)$. Remark that, for any invertible 2×2 matrix $W = (w_1,\ w_2)$,

$$(3.31)\qquad \forall z \in \mathbb{R}^2,\ z = (W^{-1}z)_1 w_1 + (W^{-1}z)_2 w_2,$$

where $(W^{-1}z)_i$ are the components of vector $W^{-1}z$. Therefore

$$\begin{aligned} (I - \Phi(T, 0))\xi &= (I - \Phi(T, 0)) \sum_{i=1}^{2} (W^{-1}(T, 0)\xi)_i w_i(T, 0) \\ &= \sum_{i=1}^{2} (1 - \mu_i(T, 0))(W^{-1}(T, 0)\xi)_i w_i(T, 0), \end{aligned}$$

and moreover

$$\begin{aligned}\int_0^T \Phi(T,\tau)B(\tau)d\tau &= \int_0^T \Phi(T,\tau)\sum_{j=1}^{2}(W^{-1}(T,\tau)B(\tau))_j w_j(T,\tau)d\tau\\ &= \sum_{j=1}^{2}\int_0^T (W^{-1}(T,\tau)B(\tau))_j \mu_j(T,\tau)w_j(T,\tau)d\tau\\ &= \sum_{i=1}^{2}\left(\sum_{j=1}^{2}\int_0^T \mu_j(T,\tau)(W^{-1}(T,\tau)B(\tau))_j\right.\\ &\qquad\left.\cdot\ (W^{-1}(T,0)w_j(T,\tau))_i d\tau\right) w_i(T,0).\end{aligned}$$

Therefore, by (3.30) it follows that

(3.32)

$$(W^{-1}(T,0)\xi)_1 = \sum_{j=1}^{2}\frac{1}{1-\mu_1(T,0)}\int_0^T \mu_j(T,\tau)(W^{-1}(T,\tau)B(\tau))_j(W^{-1}(T,0)w_j(T,\tau))_1 d\tau.$$

Moreover (3.30) can also be written in the form

$$(\Phi(T,0)^{-1}-I)\xi = \int_0^T \Phi(\tau,0)^{-1}B(\tau)d\tau, \tag{3.33}$$

and then we obtain

$$(\Phi(T,0)^{-1}-I)\xi = \sum_{i=1}^{2}(\mu_i(T,0)^{-1}-1)(W^{-1}(T,0)\xi)_i w_i(T,0),$$

$$\begin{aligned}&\int_0^T \Phi(\tau,0)^{-1}B(\tau)d\tau\\ &\quad= \sum_{i=1}^{2}\left(\sum_{j=1}^{2}\int_0^T \mu_j(\tau,0)^{-1}(W^{-1}(\tau,0)B(\tau))_j(W^{-1}(T,0)w_j(\tau,0))_i d\tau\right) w_i(T,0).\end{aligned}$$

Moreover, by (3.33)

(3.34)

$$(W^{-1}(T,0)\xi)_2 = \sum_{j=1}^{2}\frac{\mu_2(T,0)}{1-\mu_2(T,0)}\int_0^T \mu_j(\tau,0)^{-1}(W^{-1}(\tau,0)B(\tau))_j(W^{-1}(T,0)w_j(\tau,0))_2 d\tau.$$

Note that if $w_1 \in \Sigma_1$ and $w_2 \in \Sigma_2$,

$$|(W^{-1}x)_1| \vee |(W^{-1}x)_2| \leq \sqrt{\frac{f_2(1+f_2)}{f_1(1+f_1)}}\|x\|. \tag{3.35}$$

In fact, first of all, we can suppose that $\|x\| = 1$, and then W and x can be written in the following way:

$$W = \begin{pmatrix} \frac{1}{\sqrt{1+a}} & -\frac{1}{\sqrt{1+b}} \\ \frac{\sqrt{a}}{\sqrt{1+a}} & \frac{\sqrt{b}}{\sqrt{1+b}} \end{pmatrix}, \quad x = \begin{pmatrix} \frac{1}{\sqrt{1+c}} \\ \frac{\sqrt{c}}{\sqrt{1+c}} \end{pmatrix},$$

where $f_1 \leq a, b, c \leq f_2$. Then

$$W^{-1}x = \begin{pmatrix} \frac{\sqrt{1+a}}{\sqrt{1+b}} \frac{\sqrt{b}+\sqrt{c}}{\sqrt{a}+\sqrt{b}} \\ \\ \frac{\sqrt{1+b}}{\sqrt{1+c}} \frac{\sqrt{c}-\sqrt{a}}{\sqrt{a}+\sqrt{b}} \end{pmatrix},$$

which implies (3.35), since

$$\frac{\sqrt{1+a}}{\sqrt{1+c}} \vee \frac{\sqrt{1+b}}{\sqrt{1+c}} \leq \frac{\sqrt{1+f_2}}{\sqrt{1+f_1}}, \quad \frac{\sqrt{b}+\sqrt{c}}{\sqrt{a}+\sqrt{b}} \vee \frac{\sqrt{c}-\sqrt{a}}{\sqrt{a}+\sqrt{b}} \leq \frac{\sqrt{f_2}}{\sqrt{f_1}}.$$

Moreover

$$|(W^{-1}B(t))_i| \leq \frac{\sqrt{1+f_2}}{2\sqrt{f_1}} \|B(t)\|, i = 1, 2, \tag{3.36}$$

and for all $z \in \mathbb{R}^2$, for all $i = 1, 2$

$$\begin{aligned} |z_1| &\leq \frac{1}{\sqrt{1+f_1}} (|(W^{-1}z)_1| + |(W^{-1}z)_2|), \\ |z_2| &\leq \frac{\sqrt{f_2}}{\sqrt{1+f_2}} (|(W^{-1}z)_1| + |(W^{-1}z)_2|). \end{aligned} \tag{3.37}$$

Therefore from (3.32), (3.34), (3.35), (3.36), and Lemma 3.1, setting $c = \frac{1}{2} \frac{1+f_2}{f_1} \sqrt{\frac{f_2}{1+f_1}}$,

$$\begin{aligned} &|(W^{-1}(T,0)\xi)_1| \vee (W^{-1}(T,0)\xi)_2| \\ &\leq c\|B\|_{[0,T]} \left\{ \frac{1}{\mu_1(T,0)-1} \left[\mu_1(T,0) \int_0^T \mu_1(\tau,0)^{-1} d\tau \right.\right. \\ &\left.\left. + \int_0^T \mu_2(T,\tau) d\tau \right] \vee \frac{1}{1-\mu_2(T,0)} \left[\mu_2(T,0) \int_0^T \mu_1(\tau,0)^{-1} d\tau + \int_0^T \mu_2(T,\tau) d\tau \right] \right\} \\ &\leq c\|B\|_{[0,T]} \left\{ \frac{1}{\mu_1(T,0)-1} \left[\mu_1(T,0) \int_0^T e^{-\lambda_0 \tau} d\tau \right.\right. \\ &\left.\left. + \int_0^T e^{-\lambda_0 (T-\tau)} d\tau \right] \vee \frac{1}{1-\mu_2(T,0)} \left[\mu_2(T,0) \int_0^T e^{-\lambda_0 \tau} d\tau + \int_0^T e^{-\lambda_0 (T-\tau)} d\tau \right] \right\} \\ &= c\|B\|_{[0,T]} \left(\frac{\mu_1(T,0)+1}{\mu_1(T,0)-1} \vee \frac{\mu_2(T,0)+1}{1-\mu_2(T,0)} \right) (1-e^{-\lambda_0 T}) \\ &\leq c\|B\|_{[0,T]} \frac{1+e^{-\lambda_0 T}}{1-e^{-\lambda_0 T}} \frac{1-e^{-\lambda_0 T}}{\lambda_0}, \end{aligned}$$

which implies the thesis by (3.37). □

COROLLARY 3.3. *Assume that $A(t)$ and $B(t)$ satisfy the hypotheses of Lemma 3.2 and that $A(t)$ and $B(t)$ are periodic of period T. If $\xi \in \mathbb{R}^2$ is such that there exists $\tau \geq 0$ and $T \geq 0$ for which*

$$\xi = \Phi(T+\tau, \tau)\xi + \int_\tau^{T+\tau} \Phi(T+\tau, p) B(p) dp, \tag{3.38}$$

where $\Phi(t,\tau)$ *is a solution of* (3.25), *then*

$$|\xi_i| \le \frac{c_i}{\lambda_0}\|B\|_{[\tau,T+\tau]},\ i=1,2, \tag{3.39}$$

where λ_0 *and* c_i *are given by* (3.29).

Proof. If $A(t)$ is periodic, there exists a periodic and continuous matrix $P(t)$ and a constant matrix E such that $\Phi(t) = P(t)e^{Et}$, which implies that, for all $n \in \mathbb{N}$, $\Phi(T+\tau,\tau)^n = \Phi(nT+\tau,\tau)$ and $\Phi(T+\tau,\tau) = \Phi((n+1)T+\tau, nT+\tau)$. Therefore from (3.38) it follows by induction that:

$$\xi = \Phi(nT+\tau,\tau)\xi + \int_\tau^{nT+\tau} \Phi(T+\tau,p)B(p)dp\ \forall n \in \mathbb{N},$$

which implies (3.39), by (3.28). □

4. A computational method. As can be deduced from section 1 and the proof of Theorem 2.3, the following computational method for finding the control forces to drive the inverted pendulum on a cart along an assigned curve can be stated.

Given a curve $\gamma \in \mathcal{C}^2(\mathbb{R},\ \mathbb{R}^2)$, *the control force* f *is given by* (1.3), *where* θ *is the solution of system* (2.4) *with initial condition* $(\theta_0,\ \dot\theta_0) = y(1)$, $y(s)$ *being the solution of the following differential equation:*

$$\dot y(s) = (I - \Phi(T,s,y(s)))^{-1}Z(T,s,y(s)), \quad y(0)=0,$$

where $\Phi(t,s,y)$ *and* $Z(t,s,y)$ *are, respectively, the solutions of the systems*

$$\begin{cases} \dot\Phi = \partial_x F(t,s,x(t,s,y))\Phi, \\ \Phi(0) = I \end{cases} \quad \textit{and} \quad \begin{cases} \dot Z = \partial_x F(t,s,x(t,s,y))Z + \partial_s F(t,s,x(t,s,y)), \\ Z(0) = 0, \end{cases}$$

$x(t,s,y)$ *being the solution of the system* $\dot x = F(t,s,x)$, *with* $x(0) = y$ *and where* $F(t,s,x) = (x_2,\ d^{-1}(g\sin x_1 + s\left\langle (\cos x_1, \sin x_1)^T,\ \ddot\gamma(t)\right\rangle))^T$.

As an example we apply this method to a periodic curve γ given by a fifth-order spline of period $T = 5s$. This function has a continuous third-order derivative.

The spline γ satisfies the properties

$$\begin{aligned} &\gamma(0) = (0,0),\ \dot\gamma(0) = (0,-1),\ \ddot\gamma(0) = (-1,0),\ \dddot\gamma(0) = (0,0), \\ &\gamma(1) = (-2-2),\ \gamma(2) = (1,-1),\ \gamma(3) = (2,0),\ \gamma(4) = (-2,2), \end{aligned} \tag{4.1}$$

and $\frac{d^i\gamma(5)}{dt^i} = \frac{d^i\gamma(0)}{dt^i}$, for $i = 0,\ldots,3$, and is represented in Figure 4.1; it is $\|\ddot\gamma\|_\infty = 2.2m/s^2$. For numerical computation of γ use, for instance, Matlab Spline Toolbox.

Assume $d = 1$, $g = 9.8\frac{m}{s^2}$. By applying the method outlined above, we can find a control force that drives the pendulum along the spline without overturning. In particular the initial condition for system (2.4) is given by $(\theta_0,\ \dot\theta_0) = (0.0283,\ 0.0030)$, and the couple $(\theta,\ \dot\theta)$ associated to the solution is shown in Figure 4.2; it is $|\theta(t)| \le 0.085$, $|\dot\theta(t)| \le 0.24$; in fact, the pendulum rod remains almost vertical (see Figure 4.3). Remark that the bounds given by Theorem 2.3 are $|\theta(t)| \le 0.4861$, $|\dot\theta(t)| \le 1.1727$. In this example, these bounds may appear very large, but this is justified by the fact that they must apply to any trajectory whose acceleration is bounded by $2.2m/s^2$.

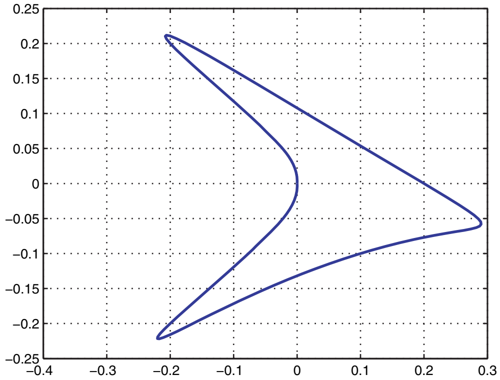


FIG. 4.1. *The fifth-order spline* γ.

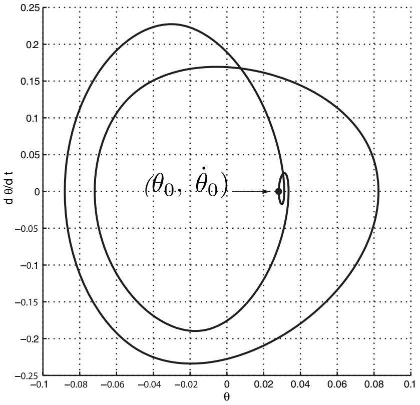


FIG. 4.2. *Solution of* (2.4) *for spline* γ.

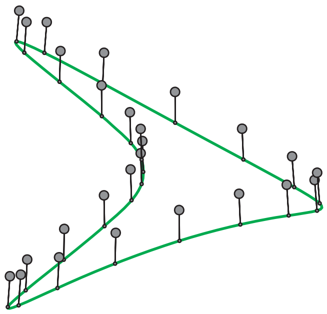


FIG. 4.3. *Control simulation for spline* γ.

REFERENCES

[1] D. ANGELI, *Almost global stabilization of the inverted pendulum via continuous state feedback*, Automatica, 37 (2001), pp. 1103–1108.

[2] K. J. ASTROM AND K. FURUTA, *Swinging up a pendulum by energy control*, Automatica, 36 (2000), pp. 287–295.

[3] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* I: *The first matching theorem*, IEEE Trans. Automat. Control, 45 (2000), pp. 2253–2270.

[4] T. R. Ding, R. Iannacci, and F. Zanolin, *Existence and multiplicity results for periodic solutions of semilinear Duffing equations*, J. Differential Equations, 105 (1993), pp. 364–409.

[5] J. Hauser, A. Saccon, and R. Frezza, *On the driven inverted pendulum*, in Proceedings of the 44th IEEE Conference on Decision and Control, European Control Conference, 2005, pp. 6176–6180.

[6] T. Holzhuter, *Optimal regulator for the inverted pendulum via Euler-Lagrange backward integration*, Automatica, 40 (2004), pp. 1613–1620.

[7] J. Mawhin, *Seventy-five years of global analysis around the forced pendulum equation*, in Proceedings of the Conference on Differential Equations and Their Applications, Equadiff 9 Brno, 1997, pp. 115–145.

[8] F. Mazenc and S. Bowong, *Tracking trajectories of the cart-pendulum system*, Automatica, 39 (2003), pp. 677–684.

[9] R. Ortega, *Counting periodic solutions of the forced pendulum equation*, Nonlinear Analysis, 42 (2000), pp. 1055–1062.

[10] A. Shiriaev, A. Pogromsky, H. Ludvigsen, and O. Egeland, *On global properties of passivity-based control of an inverted pendulum*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 283–300.

[11] G. Tarantello, *On the number of solutions for the forced pendulum equation*, J. Differential Equations, 80 (1989), pp. 79–93.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 503–522

DYNAMICS OF STOICHIOMETRIC BACTERIA-ALGAE INTERACTIONS IN THE EPILIMNION[*]

HAO WANG[†], HAL L. SMITH[‡], YANG KUANG[‡], AND JAMES J. ELSER[§]

Abstract. Bacteria-algae interaction in the epilimnion is modeled with the explicit consideration of carbon (energy) and phosphorus (nutrient). Global qualitative analysis and bifurcation diagrams of this model are presented. We hypothesize that there are three dynamical scenarios determined by the basic reproductive numbers of bacteria and algae. Effects of key environmental conditions are examined through these scenarios and from systematic and extensive simulations. It is also shown that excessive sunlight will destroy bacterial communities. Bifurcation diagrams for the depth of epilimnion mimic the profile of Lake Biwa, Japan. Competition of bacterial strains are modeled to examine Nishimura's hypothesis that in severely P-limited environments such as Lake Biwa, P-limitation exerts more severe constraints on the growth of bacterial groups with higher nucleic acid contents, which allows low nucleic acid bacteria to be competitive.

Key words. stoichiometry, bacteria, cell quota, persistence, competitive system

AMS subject classifications. Primary, 92B05, 92D40, 92D25; Secondary, 34A34, 34D05, 34D23, 34D40

DOI. 10.1137/060665919

1. Introduction.

Stoichiometry is the accounting behind chemistry. It deals with the balance of multiple chemical elements in chemical reactions. Many chemical processes are effectively studied and modeled with the applications of some simple yet powerful stoichiometric constraints. Since biomass growth is a biochemical process, ubiquitous and natural stoichiometric constraints may also be useful for modeling species growth and interactions [15, 23, 27, 28]. This concept forms the framework of the newly emerging research area of ecological stoichiometry, the study of the balance of energy and multiple chemical elements in ecological interactions [37].

It is observed that plant quality can dramatically affect the growth rate of herbivorous grazers and may even lead to their extinction. Specifically, if the quantity of an essential element in plant biomass is lower than the minimum threshold for its consumer, then the consumer's growth rate may suffer. This has been shown for both aquatic [30, 37] and terrestrial systems [31]. Stoichiometry-based population models explicitly model the highly varying nutritional quality of plant resources for consumer-resource dynamics.

Solar energy (for producing organic carbon) and nutrients (phosphorus, nitrogen, etc.) are important factors regulating ecosystem characteristics and species density. Phosphorus is often a limiting nutrient for algal production in lakes [11]. For example,

[*] Received by the editors July 24, 2006; accepted for publication (in revised form) June 26, 2007; published electronically November 30, 2007.
http://www.siam.org/journals/siap/68-2/66591.html

[†] Corresponding author. Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804. Current address: School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30329 (wanghao@math.gatech.edu).

[‡] Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (halsmith@asu.edu, kuang@asu.edu). The second author's research was partially supported by NSF grant DMS 0414270. The third author's research was partially supported by DMS-0436341, DMS/NIGMS-0342388, and DEB-0516494.

[§] School of Life Sciences, Arizona State University, Tempe, AZ 85287-15011 (j.elser@asu.edu). The fourth author's research was partially supported by DMS-0436341, DMS/NIGMS-0342388, and DEB-0516494.

in Lake Biwa, Japan, phosphorus is an extremely limiting element for both algal and bacterial growth. Lake Biwa is a large (surface area, $674km^2$) and deep (maximum depth, $104m$) lake located in the central part of Honshu Island, Japan. Nishimura, Kim, and Nagata [32] used flow cytometry to examine seasonal variations in vertical distributions of bacterioplankton in Lake Biwa. They hypothesized that in severely phosphorus (P)-limited environments such as Lake Biwa, P-limitation exerts more severe constraints on the growth of bacterial groups with higher nucleic acid (HNA) contents, which allows low nucleic acid (LNA) bacteria to be competitive and become an important component of the microbial community. A main purpose of this paper is to examine this hypothesis theoretically.

The interaction between bacteria and algae in pelagic ecosystems is complex [6]. Bacteria are nutrient-rich organisms whose growth is easily limited by nutrient supply and organic matter produced by plants and algae, which have very flexible stoichiometry compared to bacteria [29]. Suspended algae, also called phytoplankton, live in almost all types of aquatic environments. Algae grow in open water by taking up nutrients such as phosphorus and nitrogen from the water and capturing energy from sunlight. Extra energy in the form of organic carbon can be exuded from algae during photosynthesis. Bacteria require dissolved organic carbon (DOC) as a source of energy and carbon. Hence, algae are an important source of DOC to bacteria. However, bacteria and algae compete with each other for phosphorus if bacteria are limited by phosphorus [18].

In temperate lakes, the water column is seasonally separated by a thermocline into two parts, epilimnion and hypolimnion (Figure 1.1). The epilimnion is the upper warmer layer overlying the thermocline. It is usually well mixed. The hypolimnion is the bottom colder layer. The absorption and attenuation of sunlight by the water itself, by dissolved substances, and by algae are major factors controlling the potential photosynthesis and temperature. Solar energy, essential for algae, decreases rapidly with depth. Nutrients are redistributed from epilimnion to hypolimnion as the plankton detritus gradually sinks to lower depths and decomposes; the redistribution is partially offset by the active vertical migration of the plankton and by eddy diffusion across the thermocline [19]. In many lakes, algal DOC exudation is a prime energy source for bacterial growth. To simplify the study of algal stimulation of bacterial growth, we assume below that algal DOC exudation is the only source for bacterial subsistence.

Algae dynamics in a lake system have been modeled by many researchers [2, 8, 9, 20, 21, 22]. Chemostat theory and experiments have been applied frequently to nutrient competition of bacteria [4, 14, 34, 35]. For example, bacteria-algae interaction was modeled by Bratbak and Thingstad [3]. Their work provides a useful framework to develop a more realistic model. In recent years, modeling stoichiometric food web systems has gained much attention [1, 9, 15, 17, 24, 25, 26, 27]. However, these models are not directly applicable to the phytoplankton-bacteria interaction. Our models, motivated by the experiments and hypotheses of Nishimura, Kim, and Nagata [32], can be viewed as an extension as well as a variation of the work of Diehl, Berger, and Wöhrl [9] where they modeled algal growth experiments subject to varying light and nutrient availability, but without bacteria.

In the following, we will model the ecological stoichiometry of bacteria-algae interactions in the epilimnion under the "well mixed" assumption [2, 20, 21]. We perform a global qualitative analysis and present bifurcation diagrams illustrating model behavior. We discuss the implications of these bifurcation diagrams and the basic reproductive numbers of bacteria and algae. Proofs of mathematical results are placed

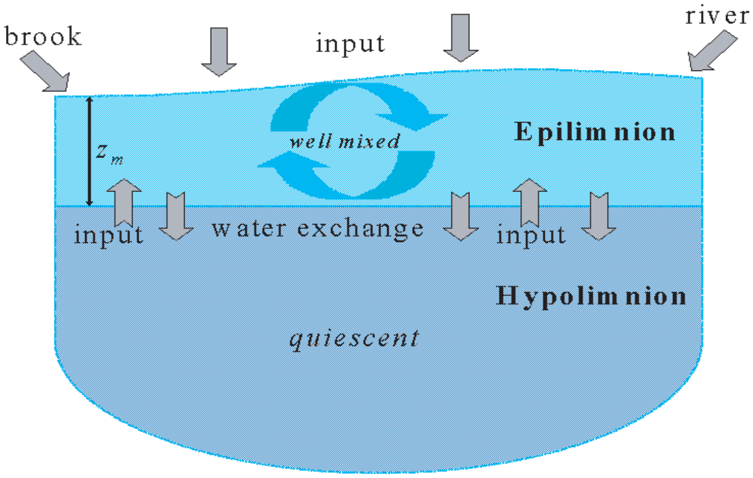


FIG. 1.1. *The cartoon lake system for our mathematical modeling.*

in the appendix. Competing bacterial strains are modeled to test some hypotheses of Nishimura, Kim, and Nagata [32]. A brief discussion section concludes the paper.

2. Modeling bacteria-algae interactions. Our model consists of five highly interconnected nonlinear differential equations, tracking the rates of change for algae, algal cell quota, dissolved mineral phosphorus, heterotrophic bacteria, and dissolved organic carbon (see Table 2.1). The algal growth is assumed to depend on the light intensity and phosphorus availability. This will be modeled according to the Lambert-Beer law and the Droop equation. The rates of change for Q, P, B are modeled according to standard approaches. In addition, algal sinking and water exchange between epilimnion and hypolimnion are included in the model. The challenge is to model the algal exudation of DOC, which is needed in the DOC equation.

According to the Lambert–Beer law, the light intensity at the depth s of a water column with algal abundance A is [20]

$$I(s, A) = I_{in} \exp[-(kA + K_{bg})s].$$

The algal carbon uptake function takes the Monod form $\frac{I(s,A)}{I(s,A)+H}$ [9].

The epilimnion is well mixed overnight [9, 20]. The depth-averaged algal growth function contains the factor (for carbon) [2, 20]

$$h(A) \equiv \frac{1}{z_m} \int_0^{z_m} \frac{I(s, A)}{I(s, A) + H} ds = \frac{1}{z_m(kA + K_{bg})} \ln\left(\frac{H + I_{in}}{H + I(z_m, A)}\right) \tag{2.1}$$

and the Droop term (for phosphorus) $1 - \frac{Q_m}{Q}$, where Q_m is the minimum algal phosphorus cell quota and Q is the actual algal phosphorus cell quota.

Algal sinking takes place at the interface between epilimnion and hypolimnion, and its rate is negatively related to the volume of epilimnion, because with a larger volume there is relatively less proportion of total species abundances or element concentrations for sinking. For convenience, we assume that the algal sinking rate is inversely proportional to the mixing layer depth z_m [2, 9]. D is the water exchange rate across the interface between epilimnion and hypolimnion and between the epilimnion and the inflow and outflow (Figure 1.1). We assume that there is a constant

Table 2.1
Variables in the bacteria-algae system (2.2).

Var.	Meaning	Unit
A	algal carbon density	mgC/m^3
Q	algal cell quota (P:C)	gP/gC $(= mgP/mgC)$
P	dissolved mineral phosphorus concentration	mgP/m^3
B	heterotrophic bacterial abundance	mgC/m^3
C	DOC concentration	mgC/m^3

phosphorus concentration, P_{in}, in the hypolimnion and in the inflow. Using the same reasoning as for algal sinking, we assume the water exchange is inversely proportional to z_m. We assume that bacteria have a fixed stoichiometry, since compared to algae, their elemental composition is relatively constant [29]. We assume that bacterial growth functions for carbon and phosphorus take the Monod form: $f(P) = \frac{P}{K_P+P}$ and $g(C) = \frac{C}{K_C+C}$, where K_P, K_C are half-saturation constants, respectively.

The exudation rate of DOC by algae is the difference between the potential growth rate attained when growth is not P-limited, $\mu_A A \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds$, and the actual growth rate, $\mu_A A(1 - \frac{Q_m}{Q}) \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds$, which gives us the form $\mu_A A \frac{Q_m}{Q} \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds$. This actually assumes that algae always fix carbon at rate $\mu_A A \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds$ and then have to dispose of excessive carbon. As in [9], we assume that the algal phosphorus uptake rate is $\rho(Q,P) = \rho_m(\frac{Q_M-Q}{Q_M-Q_m}) \frac{P}{M+P}$. At the minimum cell quota, the specific phosphorus uptake rate is just a saturating function of P. At the maximum cell quota, there is no uptake. The algal cell quota dilution rate is proportional to the algal growth rate [2].

The above assumptions yield the following bacteria-algae interaction system:

$$\begin{aligned}
\frac{dA}{dt} &= \underbrace{\mu_A A \left(1 - \frac{Q_m}{Q}\right) \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds}_{\text{algal growth limited by nutrient and energy}} - \underbrace{l_m A}_{\text{respiration}} - \underbrace{\frac{\nu + D}{z_m} A}_{\text{sinking and exchange}}, \\
\frac{dQ}{dt} &= \underbrace{\rho(Q,P)}_{\text{replenishment}} \underbrace{- \mu_A Q \left(1 - \frac{Q_m}{Q}\right) \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds}_{\text{dilution due to growth}}, \\
\frac{dP}{dt} &= \underbrace{\frac{D}{z_m}(P_{in} - P)}_{\text{P input and exchange}} - \underbrace{\rho(Q,P)A}_{\text{P consumption by algae}} - \underbrace{\theta \mu_B B f(P) g(C)}_{\text{P consumption by bacteria}}, \\
\frac{dB}{dt} &= \underbrace{\mu_B B f(P) g(C)}_{\text{bacterial growth}} - \underbrace{(\mu_r + \mu_g) B}_{\text{respiration and grazing}} - \underbrace{\frac{D}{z_m} B}_{\text{exchange}}, \\
\frac{dC}{dt} &= \underbrace{\mu_A A \frac{Q_m}{Q} \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds}_{\text{DOC exudation from algae}} - \underbrace{\frac{1}{r} \mu_B B f(P) g(C)}_{\text{DOC consumption by bacteria}} - \underbrace{\frac{D}{z_m} C}_{\text{exchange}}.
\end{aligned} \tag{2.2}$$

In the rest of this paper, we assume the following parameter values (with units and sources given in Table 2.2) for numerical simulations: $I_{in} = 300$, $k = 0.0004$, $K_{bg} = 0.3$, $H = 120$, $z_m = 30$, $Q_m = 0.004$, $Q_M = 0.04$, $\rho_m = 0.2$, $M = 1.5$, $\mu_A = 1$,

Table 2.2
Parameters in bacteria-algae system (2.2).

Par.	Meaning	Value	Ref.
I_{in}	light intensity at surface	$300\mu mol(photons)/(m^2 \cdot s)$	[9]
k	specific light attenuation coeff. of algal biomass	0.0003–$0.0004 m^2/mgC$	[2, 9]
K_{bg}	background light attenuation coefficient	0.3–$0.9/m$	[2, 9]
H	h.s.c.[1] for light-dependent algal production	$120\mu mol(photons)/(m^2 \cdot s)$	[9]
z_m	depth of epilimnion	$> 0m$, $30m$ in Lake Biwa	[32]
Q_m	algal cell quota at which growth ceases	$0.004gP/gC$	[9]
Q_M	algal cell quota at which nutrient uptake ceases	$0.04gP/gC$	[9]
ρ_m	maximum specific algal nutrient uptake rate	0.2–$1gP/gC/day$	[2, 9]
M	h.s.c. for algal nutrient uptake	$1.5mgP/m^3$	[9]
μ_A	maximum algal specific production rate	$1.0/day$	[9]
l_m	algal specific maintenance respiration loss	0.05–$0.13/day$	[2, 9]
ν	algal sinking velocity	0.05–$0.25m/day$	[2, 9]
D	water exchange rate	$0.02m/day$	[2]
P_{in}	phosphorus input	0–$150mgP/m^3$	[2]
K_P	P-dependent h.s.c. for bacterial growth	0.06–$0.4mgP/m^3$	[4]
K_C	C-dependent h.s.c. for bacterial growth	100–$400mgC/m^3$	[5]
μ_B	maximum bacterial growth rate	1.5–$4.0/day$	[4, 5]
θ	bacterial fixed cell quota	0.0063–$0.1585mgP/mgC$	[7, 16]
μ_r	bacterial respiration loss	0.1–$2.5/day$	[5, 13]
μ_g	grazing mortality rate of bacteria	0.06–$0.36/day$	[32]
r	C-dependent yield constant for bacterial growth	0.31–0.75	[10, 13]

[1] "h.s.c." stands for half-saturation constant.

$l_m = 0.1$, $\nu = 0.25$, $D = 0.02$, $P_{in} = 120$, $K_P = 0.06$, $K_C = 100$, $\mu_B = 3$, $\theta = 0.1$, $\mu_r = 0.2$, $\mu_g = 0.1$, $r = 0.5$. These specific values are taken from [2, 9] or selected from within the reasonable ranges (see Table 2.2).

Our first theorem states that there is a bounded set which all solutions of the system (2.2) eventually enter.

Theorem 1. *The system* (2.2) *is dissipative.*

3. Algae dynamics. In order to have a comprehensive understanding of the model (2.2), we study first the algae dynamics without bacteria ($B = 0$):

$$\begin{aligned} \frac{dA}{dt} &= \mu_A A \left(1 - \frac{Q_m}{Q}\right) \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds - l_m A - \frac{\nu + D}{z_m} A \equiv A\Psi(A,Q), \\ \frac{dQ}{dt} &= \rho(Q,P) - \mu_A Q \left(1 - \frac{Q_m}{Q}\right) \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds, \\ \frac{dP}{dt} &= \frac{D}{z_m}(P_{in} - P) - \rho(Q,P)A. \end{aligned} \tag{3.1}$$

From (2.1), we recall that

$$h(A) = \frac{1}{z_m} \int_0^{z_m} \frac{I(s,A)}{I(s,A)+H} ds. \tag{3.2}$$

$h(A)$ is decreasing in A, and $0 < h(A) < 1$. Furthermore, $Ah(A)$ is increasing in A. Biologically meaningful initial conditions are given by $A(0) > 0$, $\;Q_m \le Q(0) \le Q_M$,

$P(0) \geq 0$. We analyze this system on the positively invariant set

$$\Omega = \{(A, Q, P) \in \mathbb{R}^3_+ \mid A \geq 0,\, Q_m \leq Q \leq Q_M,\, P \geq 0\}.$$

Obviously the set where $A = 0$ is invariant for the system. It is easy to see that $Q_m < Q < Q_M$ whenever $Q_m < Q(0) < Q_M$; that is, the cell quota stays within the biologically confined interval.

There can be two types of steady state solutions for system (3.1): the algae extinction steady state $E_0 = (0, \hat{Q}, P_{in})$, where

$$\hat{Q} = \frac{\beta(P_{in})Q_M + \mu_A Q_m h(0)}{\beta(P_{in}) + \mu_A h(0)} > 0 \quad \text{with} \quad \beta(P) = \frac{\rho_m}{Q_M - Q_m}\frac{P}{M + P},$$

and positive steady state(s) $E^* = (\bar{A}, \bar{Q}, \bar{P})$ with $\Psi(\bar{A}, \bar{Q}) = 0$.

The standard computation shows that the basic reproductive number for algae is

$$R_0 = \frac{\mu_A \beta(P_{in})(Q_M - Q_m)h(0)}{(\beta(P_{in})Q_M + \mu_A Q_m h(0))(l_m + \frac{\nu + D}{z_m})} = \frac{\mu_A h(0)(1 - Q_m/\hat{Q})}{l_m + \frac{\nu + D}{z_m}}.$$

Here $h(0) = \frac{1}{z_m K_{bg}} \ln(\frac{H + I_{in}}{H + I_{in}\exp(-z_m K_{bg})})$ is the potential average sunlight intensity in the epilimnion without algal shading. Indeed, R_0 is calculated from $\Psi(0, \hat{Q})$ so that $R_0 > 1 \Leftrightarrow \Psi(0, \hat{Q}) > 0$. R_0 is the average amount of new algae produced by one unit of algae (measured in carbon content) during the algal life span in the epilimnion. It is an indicator of algal viability. Part of Theorem 2 states that R_0 is an indicator for the local stability of E_0.

We observe that increasing sunlight input or phosphorus input enhances algal viability, since $\frac{\partial R_0}{\partial I_{in}} = \frac{\partial R_0}{\partial h(0)}\frac{\partial h(0)}{\partial I_{in}} > 0$ and $\frac{\partial R_0}{\partial P_{in}} = \frac{\partial R_0}{\partial \beta(P_{in})}\frac{\partial \beta(P_{in})}{\partial P_{in}} > 0$. Weakening water exchange enhances algal viability, since $\frac{\partial R_0}{\partial D} < 0$.

Theorem 2 is our main mathematical result. When $R_0 < 1$, we establish the local and global stability of E_0, which is equivalent to saying that algae will die out. It can be shown that there is no positive equilibrium E^* when $R_0 < 1$, in which case the existing results of general competitive systems can be applied to prove the global stability of E_0. When $R_0 > 1$, we prove that E_0 is unstable, algae are uniformly persistent, and there is a unique positive steady state E^*.

Theorem 2. *If $R_0 < 1$, E_0 is locally asymptotically and globally asymptotically stable. $R_0 > 1$ implies that E_0 is unstable, there exists a unique positive equilibrium E^*, and algae uniformly persist: there exists $\epsilon > 0$ such that*

$$\liminf_{t \to \infty} A(t) > \epsilon$$

for all solutions with $A(0) > 0$.

In the following, we show that the global stability of E^* is true in two special cases. It is known that the algal cell quota changes on a much faster timescale than the algal (carbon) biomass and the nutrient [23]. Additionally, since dQ/dt is linear in Q, there is a unique solution to $dQ/dt = 0$. Hence, the fast-slow approximation is achieved by setting $dQ/dt = 0$ and substituting the solution of $dQ/dt = 0$ into the other equations. Then the following theorem holds.

Theorem 3. *E^* is globally asymptotically stable for the planar system obtained from the system* (3.1) *by setting $dQ/dt = 0$, when $R_0 > 1$.*

The next theorem gives a partial result of global stability of E^*.

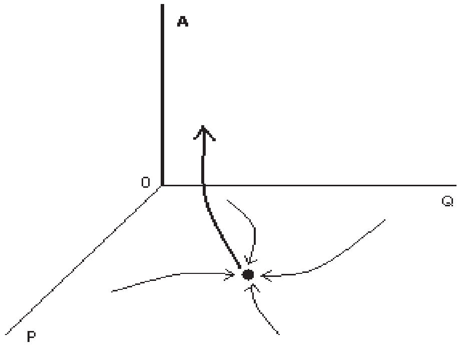


FIG. 3.1. *Algae dynamics phase space when $R_0 > 1$. The algae extinction equilibrium $E_0 = (0, \hat{Q}, P_{in})$ is globally attracting on the subspace $\Omega_2 = \{x \in \Omega \mid A = 0\}$, but is a uniform weak repeller for $\Omega_1 = \{x \in \Omega \mid A \neq 0\}$, and A is persistent in this case.*

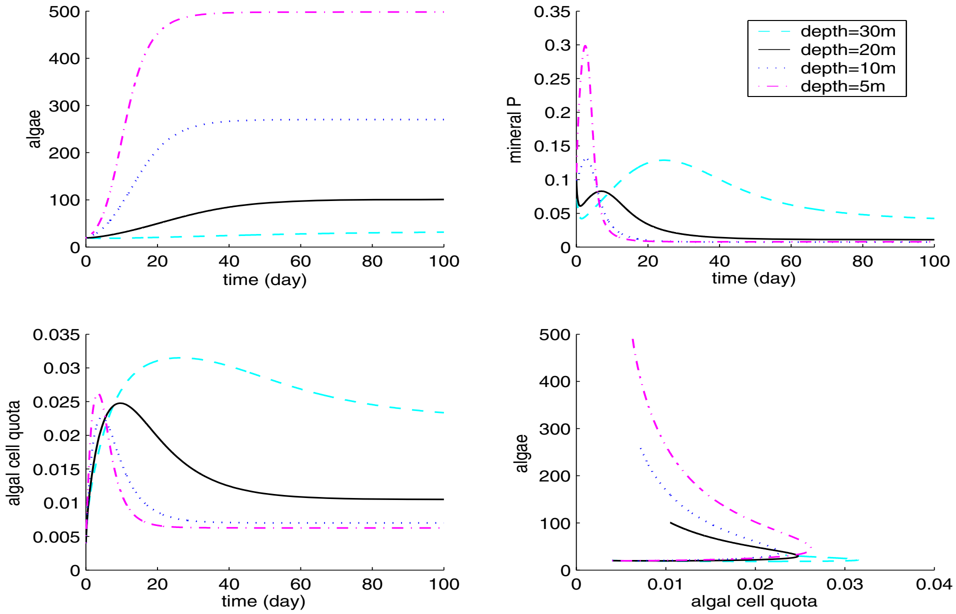


FIG. 3.2. *Algae dynamics without bacteria with respect to different depths of the mixing layer: All the variables approach the positive equilibrium in about two months. The simulation benefited from the explicit expression of $h(A)$ given in* (2.1).

THEOREM 4. *E^* is globally asymptotically stable when $R_0 > 1$ and $l_m = \nu = 0$.*

The property that a locally asymptotically stable (in linear approximation) steady state is globally attracting is an open condition in parameters [36]. Hence, E^* is still globally asymptotically stable for small positive l_m and ν.

Our main mathematical results for the system (3.1) are briefly expressed by the phase space diagram (Figure 3.1) for the case $R_0 > 1$. Typical solutions are simulated in Figure 3.2 for different depths of epilimnion. Algal abundance is negatively related to the depth because the average sunlight intensity in the epilimnion is lower when the epilimnion is deeper. The eventual P concentration is relatively large when the epilimnion is deep, whereas the eventual concentration is small when it is shallow.

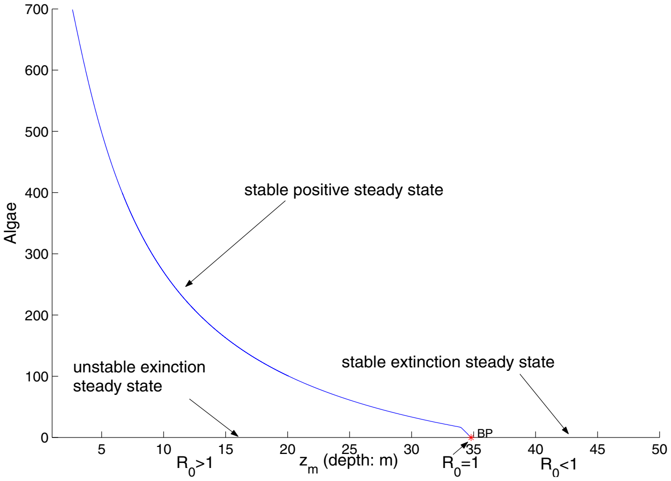


Fig. 3.3. *Bifurcation diagram for algae dynamics without bacteria. The shallower the better for algae in the algae system* (3.1)*. This bifurcation diagram confirms our mathematical findings. When $R_0 > 1$, the algae extinction equilibrium is unstable, and the only positive equilibrium appears to be globally attractive. The branching point occurs at $R_0 = 1$. When $R_0 < 1$, there is no positive equilibrium, and the algae extinction equilibrium is globally attracting. This numerical result is generated by the continuation software "MatCont" in MATLAB.*

Eventual concentrations at the depths $5m, 10m, 20m$ are similar and low, which indicates that P becomes limiting in a shallow epilimnion. In contrast, the algal cell quota is positively related to epilimnion depth. This is due to the fact that the algal cell quota is positively related to the P concentration. Hence, epilimnion depth has two influences on algae in our simulation: It is positively related to P (at least in Figure 3.2) and negatively related to the average sunlight intensity through the average light uptake integral term. In the case of Figure 3.2, if C has a larger effect than P, then algal abundance is negatively related to the depth. It is not clear whether or not algal abundance can be positively related to depth when P is more limited than C in some lakes. The algae-quota phase plane shows that algae and their cell quotas are positively related in the very beginning, but they are negatively related eventually, demonstrating a general phenomenon of "larger quantity leads to lower quality." The bifurcation diagram with respect to the mixing layer depth (Figure 3.3) illustrates that algae love shallower epilimnions and also illustrates our mathematical findings.

4. Bacteria-algae interaction dynamics.

We return to the original bacteria-algae system (2.2). We analyze this system on the positively invariant set

$$\Omega = \{(A, Q, P, B, C) \in \mathbb{R}^5_+ \mid A \geq 0,\ \ Q_m \leq Q \leq Q_M,\ \ P \geq 0, B \geq 0, C \geq 0\}.$$

The system (2.2) may have three types of equilibria: the extinction steady state $e_0 = (0, \hat{Q}, P_{in}, 0, 0)$, the bacteria extinction only steady state $e_1 = (\bar{A}, \bar{Q}, \bar{P}, 0, \bar{C})$, and the coexistence steady state(s) e^* with all components positive (see Figure 4.1). We can calculate the basic reproductive number for bacteria, R_1, by linearizing about e_1. We obtain

$$R_1 = \frac{\mu_B f(\bar{P}) g(\bar{C})}{\mu_r + \mu_g + \frac{D}{z_m}},$$

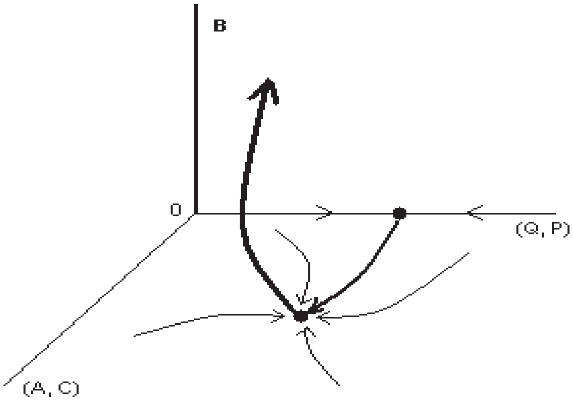


Fig. 4.1. *An abstract phase space diagram for system* (2.2) *when* $R_0 > 1$ *and* $R_1 > 1$. Q, P *are placed on one axis (say x-axis),* A, C *are placed on another axis (say y-axis), and* B *is on the vertical axis (z-axis). Extinction equilibrium* $e_0 = (0, \hat{Q}, P_{in}, 0, 0)$ *is globally attracting on the subspace* $\{x \in \Omega \mid A = B = C = 0\}$*, but is a repeller for* $\Omega_2 = \{x \in \Omega \mid B = 0\}$*. Bacteria extinction only equilibrium* $e_1 = (\bar{A}, \bar{Q}, \bar{P}, 0, \bar{C})$ *is globally attracting on the subspace* Ω_2*, but is a repeller for* $\Omega_1 = \{x \in \Omega \mid B \neq 0\}$*. Bacteria persist, and at least one coexistence equilibrium exists.*

where $\bar{C} = \frac{\mu \bar{A} z_m}{D} \frac{Q_m}{\bar{Q}} h(\bar{A})$ and $\bar{P}$, $\bar{A}$, $\bar{Q}$ are components of E^* in the system (3.1). This number is defined under the assumption $R_0 > 1$. When $R_0 < 1$, we have proved that there is no positive equilibrium in system (3.1), which means at least one of $\bar{P}$, $\bar{A}$, $\bar{Q}$ is undefined or out of the region of interest. Biologically, R_1 is the average biomass of new bacteria produced by one unit of bacterial biomass during the bacterial life span in epilimnion. R_1 should be an indicator for the local stability of e_1; hence, R_1 is an indicator for the bacterial viability when $R_0 > 1$.

A simple sufficient condition for the extinction of both algae and bacteria is given in the next theorem.

Theorem 5. *When* $\mu_A h(0) < \frac{D}{z_m}$ *(* $\Leftrightarrow \frac{\mu_A}{K_{bg}} \ln(\frac{H+I_{in}}{H+I_{in}\exp(-z_m K_{bg})}) < D$ *), both algae and bacteria will die out; i.e.,* $\lim_{t\to\infty} A(t) = \lim_{t\to\infty} B(t) = 0$ *for all nonnegative initial conditions.*

It is easy to observe that $R_0 < \frac{\mu_A h(0)}{D/z_m}$. Hence $\mu_A h(0) < \frac{D}{z_m}$ implies $R_0 < 1$.

Figure 4.2 confirms that both species go extinct when $R_0 < 1$, a weaker condition than the condition $\mu_A h(0) < \frac{D}{z_m}$ in Theorem 5. The line-filled region expands rapidly when the sunlight increases past $800 \mu mol(photons)/(m^2 \cdot s)$. This suggests that high light intensity can negatively affect bacteria, even driving them to extinction due to competition with algae. Hence, the balance of light and nutrient is significant for the lake system, which is in agreement with the “light:nutrient” hypothesis [38].

Branching points in Figures 3.3 and 6.1 are identical, since all of them are determined by the same condition $R_0 = 1$. R_1 does not affect this branching point, since R_1 is defined only if $R_0 > 1$. $R_0 > 1$ implies $R_1 > 1$ in the white region of Figure 4.2. Upon existent mathematical results, we hypothesize that there are three types of dynamics: (1) $R_0 > 1, R_1 > 1$ ensure the persistence of species (white region in Figure 4.2); (2) $R_0 > 1, R_1 < 1$ enable the persistence of algae but the extinction of bacteria (line-filled region in Figure 4.2); (3) $R_0 < 1$ leads to the extinction of all species (grey region in Figure 4.2).

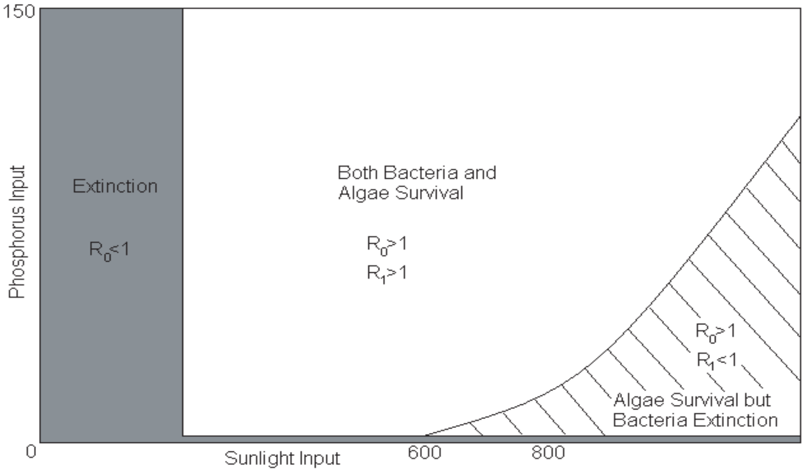
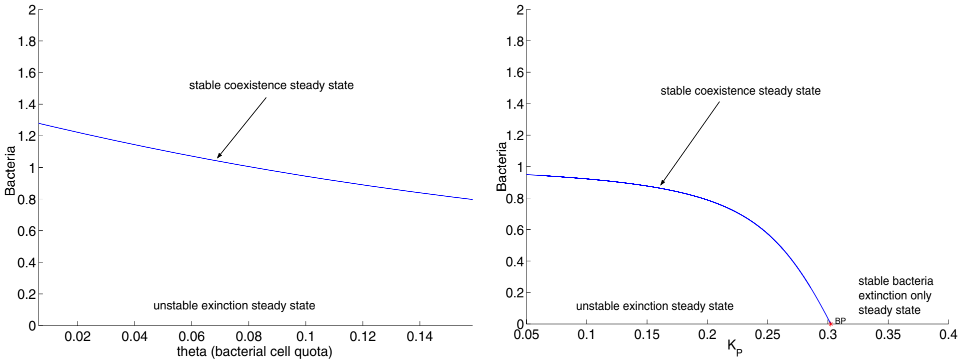


FIG. 4.2. *Regions of P_{in} versus I_{in} for survival and extinction of bacteria and algae. Both algae and bacteria go extinct ($R_0 < 1$) in the grey region. Both algae and bacteria survive ($R_0 > 1, R_1 > 1$) in the white region. Algae survive, but bacteria go extinct ($R_0 > 1, R_1 < 1$) in the line-filled region. We run simulations of the system* (2.2) *for each pair of (I_{in}, P_{in}) and then put the point in the grey region if both A and B go to zero, in the white region if both persist, and in the line-filled region if A persists, but B goes to zero.*

5. Competing bacterial strains. In lake ecosystems, bacteria comprise the most important trophic level for processing dissolved organic matter (DOM) and consume almost half of the primary production [32]. Most existing studies have treated bulk bacterial communities as a homogeneous pool, even though they consist of diverse subgroups that differ in metabolic state, DOM use, growth rate, susceptibility to grazing, and phylogenetic affiliations. One of the challenges for aquatic microbial ecology is to clarify variations and regulation of different bacterial subgroups in order to better understand the internal dynamics of the bacterioplankton "black box."

Growth characteristics and ecological roles of LNA bacteria are controversial. Some previous studies have claimed that LNA bacteria represent less active, dormant, or even dead cells. However, Nishimura, Kim, and Nagata [32] found that the growth rates of LNA bacteria were comparable to or even exceeded HNA bacteria in Lake Biwa. This is probably because LNA bacteria have higher nutrient uptake efficiencies (this means bacteria take up nutrients efficiently even at very low external concentrations, i.e., have a low half-saturation constant for P) and lower requirements for P (this means less P per unit carbon is needed, or a smaller cell quota). An important implication of this scenario is that LNA bacteria, under severe P-limitation conditions, represent an "active" subgroup that outcompetes HNA bacteria and hence may play an important role in the functioning of the microbial loop [32]. In fact, both of these seemingly contradictory statements can be correct under different situations. One of our main motivations for this work is to examine these statements theoretically.

To examine the statement that "LNA bacteria have lower requirements for P," we plot the bifurcation diagram of the bacterial variable with respect to the cell quota parameter θ for the system (2.2). This is done in Figure 5.1(a). Clearly, this figure supports the "P requirement" hypothesis, since a lower bacterial cell quota gives higher bacterial abundance at equilibrium. The second statement, "LNA bacteria have higher nutrient uptake efficiencies," is supported by Figure 5.1(b), which has



(a) Bacteria vs θ: supporting 'P requirement' hypothesis. θ is the bacterial homeostasis cell quota.

(b) Bacteria vs K_P: supporting 'nutrient uptake efficiency' hypothesis. K_P is the P-dependent half-saturation constant for the bacterial growth.

FIG. 5.1. *We examine Nishimura's hypotheses by system* (2.2).

higher sensitivity than Figure 5.1(a). This seems to suggest that "nutrient uptake efficiency" is probably the key factor for LNA bacteria to dominate HNA bacteria near the surface since this is where P is most limiting.

To examine Nishimura's hypothesis, we model the competition of two bacterial strains, HNA bacteria (B_1) and LNA bacteria (B_2), and assume these two strains are heterogeneous in P usage and the maximum growth rate, but homogeneous in C usage. Elser et al. [12], and Sterner and Elser [37] have proposed the "growth rate hypothesis" to explain variation among organisms in biomass C:P and N:P ratios. The growth rate hypothesis states that differences in organismal C:N:P ratios are caused by differential allocations to RNA necessary to meet the protein synthesis demands of rapid rates of biomass growth and development [37, p. 144]. Due to the growth rate hypothesis, the bacterial cell quota is strongly correlated to the maximum growth rate; that is, $\theta_1/\theta_2 = \iota\mu_1/\mu_2$, where ι is a positive constant. For convenience, we assume $\theta_1/\theta_2 = \mu_1/\mu_2$. With these assumptions, the competition system takes the form

$$
\begin{aligned}
\frac{dA}{dt} &= \mu_A A\left(1-\frac{Q_m}{Q}\right)h(A) - l_m A - \frac{\nu + D}{z_m}A,\\
\frac{dQ}{dt} &= \rho(Q,P) - \mu_A Q\left(1-\frac{Q_m}{Q}\right)h(A),\\
\frac{dP}{dt} &= \frac{D}{z_m}(P_{in}-P) - \rho(Q,P)A - [\theta_1\mu_1 B_1 f_1(P) + \theta_2\mu_2 B_2 f_2(P)]g(C),\\
\frac{dB_1}{dt} &= \mu_1 B_1 f_1(P) g(C) - (\mu_r + \mu_g)B_1 - \frac{D}{z_m}B_1,\\
\frac{dB_2}{dt} &= \mu_2 B_2 f_2(P) g(C) - (\mu_r + \mu_g)B_2 - \frac{D}{z_m}B_2,\\
\frac{dC}{dt} &= \mu_A A\frac{Q_m}{Q}h(A) - \frac{1}{r}[\mu_1 B_1 f_1(P) + \mu_2 B_2 f_2(P)]g(C) - \frac{D}{z_m}C,
\end{aligned}
\tag{5.1}
$$

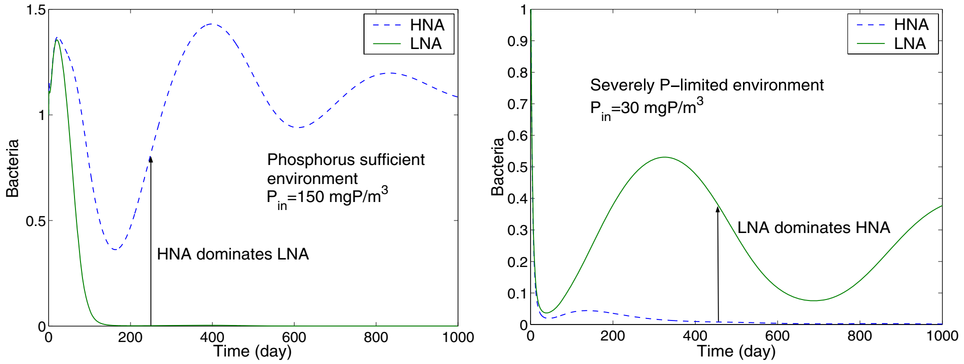


FIG. 5.2. *Under different lake environments, a different bacterial strain dominates.* $\mu_1 = 4, \mu_2 = 2, K_1 = 0.15, K_2 = 0.06, \theta_1 = 0.1, \theta_2 = 0.05$ *with the same units as in Table* 2.2.

where $f_i(P) = \frac{P}{K_i+P}$, $i = 1, 2$. Since the maximum LNA bacteria growth rate is lower and LNA bacteria have higher nutrient uptake efficiencies, it is biologically reasonable to assume that $\mu_1 > \mu_2$, $K_1 > K_2$.

The positivity and dissipativity of the system (5.1) obviously hold, and the proof can be formulated in a fashion similar to that of Theorem 1.

As we can see from Figure 5.2, HNA bacteria grow faster than LNA bacteria whenever P is sufficient, simply because in such situations the maximum HNA bacteria growth rate is greater than that of the LNA bacteria. But LNA bacteria grow faster than HNA bacteria whenever P is severely limited (Figure 5.2), because LNA bacteria have higher nutrient uptake efficiencies and lower requirements for P. Therefore, these seemingly conflicting phenomena can happen under distinct nutrient conditions.

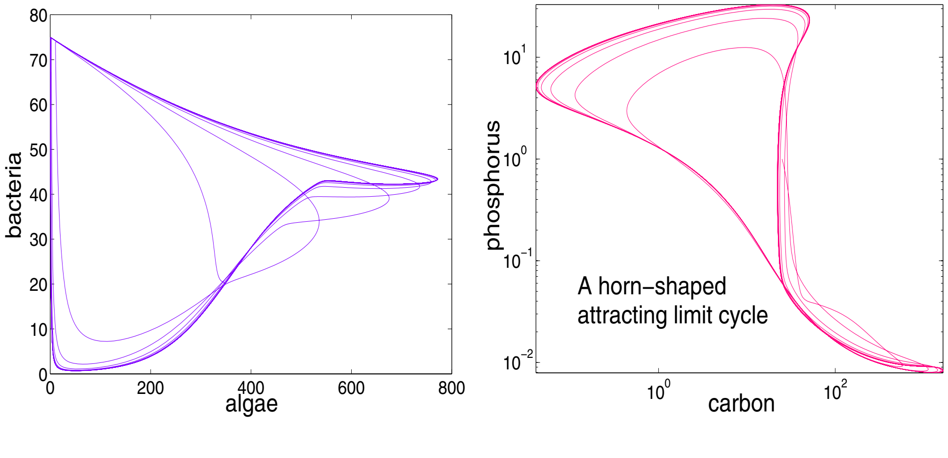
We can seek the expression of the potential positive steady state of the system (5.1). From the bacterial equations, we have

$$\mu_i f_i(P) g(C) = (\mu_r + \mu_g) + \frac{D}{z_m}$$

for the potential positive steady state. Solving it for P, we obtain

$$P = \frac{aK_i}{\mu_i g(C) - a},$$

where $a = (\mu_r + \mu_g) + \frac{D}{z_m}$ is a constant. Assume the system (5.1) has a positive steady state; then $\mu_i g(C) > a$ holds for $i = 1, 2$. For a fixed C level, the P level for the potential positive steady state of that bacterial strain is increasing in K_i, but decreasing in μ_i. Since K_2 is smaller, LNA bacteria have more chance to survive because of the lower level of P required to reach its potential positive steady state level. However, μ_2 is also smaller, which can reduce the LNA bacteria's chance to survive because of the higher level of P required to obtain its potential positive steady state level. In other words, K_i and μ_i work together in a nonlinear fashion. These arguments are only true for the case when a single steady state is globally attractive, in which case only one bacterial strain persists (see Figure 5.2). These bacterial strains may coexist in the form of limit cycles, as periodic solutions are possible even for system (2.2) (for example, Figure 5.3).



(a) Bacteria vs algae

(b) P vs C

FIG. 5.3. *With shallow mixing layer* $z_m = 1$ *and default values for other parameters, system* (2.2) *exhibits complex dynamics.*

6. Discussion. Mechanistically formulated mathematical models of population dynamics are sought after since they often have advantages over phenomenologically derived ad hoc models in generating plausible and verifiable dynamics. However, the challenge of developing a mechanistic and predictive theory for biological systems is daunting. Exciting progress in understanding and modeling ecological systems in the last decade has been achieved through the application of the theory of ecological stoichiometry [37] and the consideration of interactions between nutrient and light availability [8, 9, 20, 21, 22]. Our models (2.2) and (5.1), hybrids of mechanistic and phenomenological derivations and motivated by the experiments and hypotheses of Nishimura, Kim, and Nagata [32], continue this newly established tradition. They can be viewed as an extension as well as a variation of the work of Diehl, Berger, and Wöhrl [9] who modeled algal growth experiments subject to varying light and nutrient availability.

Our preliminary analytical results on system (2.2) demonstrate that it is mathematically interesting, and our extensive bifurcation and numerical simulation work suggests that it is biologically sound.

We leave many mathematical questions open, including the global qualitative result below.

Conjecture. E^* is globally asymptotically stable when $R_0 > 1$.

Theorems 3 and 4, together with the bifurcation diagram (Figure 3.3) and the uniqueness of E^*, support the conjecture. This limiting case global stability result (Theorem 3) for the positive equilibrium suggests that the conjecture is true when the cell quota evolves on a much faster timescale than other variables. Theorem 4 and its extension are pure mathematical results that give more credence to the conjecture.

Obviously, algae are favored by shallow epilimnia, sufficient sunlight, and P inputs, while bacteria are favored by medium depths of epilimnion and sunlight and sufficient P input. With a larger P input, the ecosystem can thrive with more intensive sunlight input. Alternatively, with more intensive sunlight input, the algae-bacteria ecosystem may need more P input to be viable. When the epilimnion is very shallow

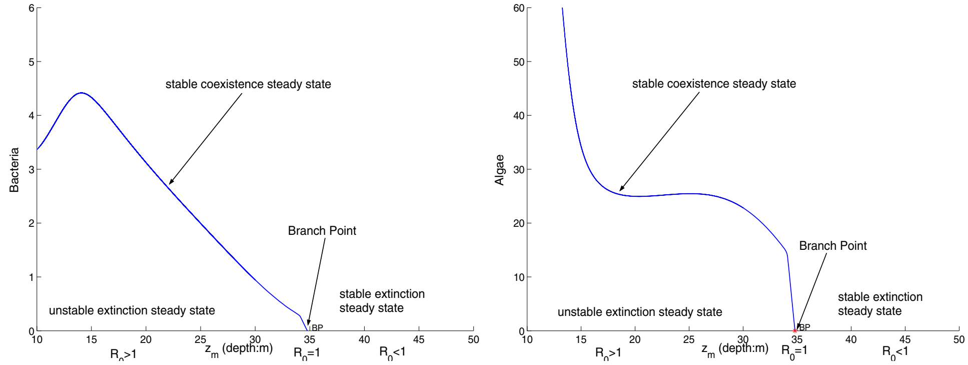


FIG. 6.1. *Bifurcation diagrams of system* (2.2) *with respect to depth of epilimnion.*

($z_m = 1$), the system (2.2) may generate complicated attractors such as that shown in Figure 5.3. It can be said that a shallower epilimnion tends to be more transient and fragile than a deep one.

Bifurcation diagrams of bacteria and algae versus depth are shown in Figure 6.1. From Figures 6.1(a) and (b), we observe that neither algae nor bacteria may survive in a very deep mixing layer ($> 35m$). In Lake Biwa, the northern lake with mean depth $43m$ is much deeper than the southern lake with mean depth $4m$. Our bifurcation diagrams (Figure 6.1) try to mimic bacterial and algal abundances in Lake Biwa from the southern site to the northern site. According to these diagrams, bacteria-to-algae ratios in the south should be smaller than in the north. *This numerical observation may be tested in the field.*

The mathematical study of the more involved system (5.1) is even more complex, and we thus opted to perform only numerical simulations to examine the hypotheses of Nishimura, Kim, and Nagata [32]. Our bifurcation diagrams (Figure 5.1) suggest that higher nutrient uptake efficiencies of LNA bacteria are the key factor for LNA bacteria to dominate HNA bacteria in severely P-limited lakes.

Appendix.

Proof of Theorem 1. Positivity obviously holds for the system. Let $R = AQ + P + \theta B$, which is the total phosphorus in system (2.2). Then

$$\frac{dR}{dt} = \frac{D}{z_m}(P_{in} - R) - \left(l_m + \frac{\nu}{z_m}\right) AQ - \theta(\mu_r + \mu_g)B \leq \frac{D}{z_m}(P_{in} - R),$$

which implies

$$R^\infty = \limsup_{t\to\infty} R(t) \leq P_{in} \qquad \text{and} \qquad R(t) \leq \max\{P_{in}, R(0)\}.$$

Since all the variables are positive and $Q_m \leq Q \leq Q_M$, we have

$$A^\infty = \limsup_{t\to\infty} A(t) \leq \frac{P_{in}}{Q_m} \qquad \text{and} \qquad A(t) \leq \frac{1}{Q_m} \max\{P_{in}, R(0)\}.$$

Noting that

$$\frac{dC}{dt} \leq \mu_A A - \frac{D}{z_m} C \leq \frac{\mu_A}{Q_m} \max\{P_{in}, R(0)\} - \frac{D}{z_m} C,$$

we have

$$C(t) \leq \max\left\{\frac{z_m\mu_A}{DQ_m}\max\{P_{in}, R(0)\}, C(0)\right\} = \max\left\{\frac{z_m\mu_A P_{in}}{DQ_m}, \frac{z_m\mu_A R(0)}{DQ_m}, C(0)\right\}.$$

Hence, for given initial conditions, $C(t)$ is bounded. Therefore, all the variables are bounded. It is easy to show that

$$\limsup_{t\to\infty} C(t) = C^\infty \leq \frac{z_m\mu_A}{D}A^\infty \leq \frac{z_m\mu_A}{D}\frac{P_{in}}{Q_m} = \frac{z_m\mu_A P_{in}}{DQ_m}.$$

Consequently, the bacteria-algae system (2.2) is dissipative, and

$$\wp = \left\{(A, Q, P, B, C) \in \Omega \mid AQ + P + \theta B \leq P_{in}, C \leq \frac{z_m\mu_A P_{in}}{DQ_m}\right\}$$

is a global attracting region for the system. □

Proof of Theorem 2. At E_0, the Jacobian matrix is

$$J(E_0) = \begin{pmatrix} \Psi(0, \hat{Q}) & 0 & 0 \\ + & \lambda_1 & + \\ - & 0 & \lambda_2 \end{pmatrix},$$

where λ_1 and λ_2 are negative numbers. It is easy to see that the eigenvalues of $J(E_0)$ are $\Psi(0, \hat{Q}), \lambda_1$, and λ_2. $R_0 < 1$ implies $\Psi(0, \hat{Q}) < 0$. Hence E_0 is locally asymptotically stable. $R_0 > 1$ implies $\Psi(0, \hat{Q}) > 0$, which implies that E_0 is unstable.

For the case $R_0 > 1$, let $x = (A, Q, P)$ and $x' = F(x)$; then $F : \mathbb{R}^3_+ \longrightarrow \mathbb{R}^3$ is locally Lipschitzian. Let $\Omega_1 = \{(A, Q, P) \in \Omega \mid A \neq 0\}$; $\Omega_2 = \{(A, Q, P) \in \Omega \mid A = 0\}$; then $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, with Ω_2 being a closed invariant subset of $\mathbb{R}^3_+$ and Ω_1 positively invariant. E_0 is the only equilibrium in Ω_2. It is easy to show that the solution that starts in Ω_2 converges to $\{E_0\}$.

The singleton set $\{E_0\}$ is a uniform weak repeller for Ω_1 when $R_0 > 1$ and an isolated invariant set in Ω [39]. It is acyclic in Ω_2. Hence Ω_2 is a uniform strong repeller for Ω_1, and there exists an equilibrium $x^* \in \Omega_1$, $F(x^*) = 0$ [40]. The first conclusion implies that A is uniformly persistent.

We are now ready to establish the uniqueness of E^*. E^* satisfies

$$\text{(A.1)} \qquad \mu_A\left(1 - \frac{Q_m}{Q}\right)h(A) - l_m - \frac{\nu + D}{z_m} = 0,$$

$$\text{(A.2)} \qquad \rho(Q, P) - \mu_A Q\left(1 - \frac{Q_m}{Q}\right)h(A) = 0,$$

$$\text{(A.3)} \qquad \frac{D}{z_m}(P_{in} - P) - \rho(Q, P)A = 0.$$

By simple eliminations, we see that

$$P = P_{in} - \frac{z_m}{D}\left(l_m + \frac{\nu + D}{z_m}\right)AQ = P_{in} - \left(\frac{z_m}{D}l_m + \frac{\nu + D}{D}\right)AQ.$$

By substituting this into (A.1), we have

$$\text{(A.4)} \qquad A = \frac{M + P_{in} - \frac{\rho_m}{(l_m + \frac{\nu+D}{z_m})Q}\frac{Q_M - Q}{Q_M - Q_m}P_{in}}{(\frac{z_m}{D}(l_m + \frac{\nu+D}{z_m})Q - \rho_m\frac{z_m}{D}\frac{Q_M - Q}{Q_M - Q_m}} \equiv F(Q).$$

Therefore

$$\Phi(Q) \equiv \Psi(F(Q), Q) = \mu_A \left(1 - \frac{Q_m}{Q}\right) h(F(Q)) - l_m - \frac{\nu + D}{z_m} = 0.$$

We will show that $\Phi(Q) = 0$ has a unique positive solution. To this end, we show that $F(Q)$ is decreasing in Q. Notice that

$$F(x) = \frac{a - (\frac{b}{x} - c)}{dx - (e - fx)} = \frac{(a+c)x - b}{(d+f)x^2 - ex},$$

where

$$a = M + P_{in}, b = \frac{\rho_m}{l_m + \frac{\nu+D}{z_m}} \frac{Q_M}{Q_M - Q_m} P_{in}, c = \frac{\rho_m}{l_m + \frac{\nu+D}{z_m}} \frac{1}{Q_M - Q_m} P_{in},$$

$$d = \frac{z_m}{D}\left(l_m + \frac{\nu + D}{z_m}\right), e = \rho_m \frac{z_m}{D} \frac{Q_M}{Q_M - Q_m}, f = \rho_m \frac{z_m}{D} \frac{1}{Q_M - Q_m}.$$

We have

$$F'(x) = \frac{-(a+c)(d+f)x^2 + 2b(d+f)x - be}{((d+f)x^2 - ex)^2} \equiv \frac{G(x)}{((d+f)x^2 - ex)^2}.$$

Observe that

$$\Delta = [2b(d+f)]^2 - 4(a+c)(d+f)be = 4b(d+f)[b(d+f) - e(a+c)] < 0$$

since $b(d+f) - e(a+c) = -\rho_m \frac{z_m}{D} \frac{Q_M}{Q_M - Q_m} M < 0$. Therefore $G(x) = -(a+c)(d+f)x^2 + 2b(d+f)x - be < 0$. Therefore $F(x)$ is strictly decreasing in x. As a result, the uniqueness of E^* holds.

Assume now that $R_0 < 1$. If we have a positive equilibrium, then $\bar{Q} > \hat{Q}$ since $\Phi(\hat{Q}) = \Psi(0, \hat{Q}) < 0$. Observe that since $\bar{P} < M + \bar{P}$, we need only

$$\left(l_m + \frac{\nu + D}{z_m}\right) \bar{Q} < \rho_m \frac{Q_M - \bar{Q}}{Q_M - Q_m}$$

to guarantee that $\bar{P} > 0$. To ensure $\bar{A} > 0$, due to (A.4), we need the more restrictive condition

$$\left(\frac{z_m}{D} l_m + \frac{\nu + D}{D}\right) \bar{Q} - \rho_m \frac{z_m}{D} \frac{Q_M - \bar{Q}}{Q_M - Q_m} < -M \left(\frac{z_m}{D} l_m + \frac{\nu + D}{D}\right) \bar{Q}/P_{in}.$$

Hence $\bar{A} > 0$ will ensure $\bar{P} > 0$. The previous inequality implies that

$$\bar{Q} < \frac{\beta(P_{in}) Q_M}{\beta(P_{in}) + (l_m + \frac{\nu+D}{z_m})}.$$

Recall that $R_0 < 1$ implies that $\mu_A(1 - Q_m/\hat{Q})h(0) < l_m + \frac{\nu+D}{z_m}$ and

$$\hat{Q} = \frac{\beta(P_{in}) Q_M + \mu_A Q_m h(0)}{\beta(P_{in}) + \mu_A h(0)}.$$

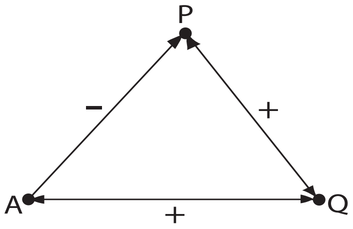


FIG. A.1. *Graph of algae system* (3.1) *to check that the system is competitive. This is observed from the Jacobian matrix of system* (3.1) *in the proof of Theorem* 2.

Hence

$$\bar{Q} - \hat{Q} < \frac{\beta(P_{in})Q_M}{\beta(P_{in}) + (l_m + \frac{\nu+D}{z_m})} - \frac{\beta(P_{in})Q_M + \mu_A Q_m h(0)}{\beta(P_{in}) + \mu_A h(0)}.$$

Simple computation shows that

$$\bar{Q} - \hat{Q} < -\frac{\mu_A h(0) Q_m(\beta(P_{in}) + \mu_A h(0))}{(\beta(P_{in}) + (l_m + \frac{\nu+D}{z_m}))(\beta(P_{in}) + \mu_A h(0))} < 0,$$

a contradiction to $\bar{Q} > \hat{Q}$.

We now proceed to show that E_0 is globally asymptotically stable when $R_0 < 1$. The Jacobian matrix of system (3.1) has the structure

$$J = \begin{pmatrix} * & + & 0 \\ + & * & + \\ - & + & * \end{pmatrix}$$

which is sign-stable for the off-diagonal elements. According to the graph in Figure A.1, every closed loop has an even number of edges with + signs; thus the system (3.1) is *monotone* ([33, p. 50–51]) in Ω with respect to the order defined by

$$K_m = \{(A, Q, P) \in \mathbb{R}^3 \mid A \geq 0, Q \leq 0, P \geq 0\}.$$

An application of monotone dynamical system theory ([33, Prop. 4.3, p. 44]) yields the statement that if system (3.1) has a positive periodic solution in Ω, then it contains an equilibrium in Ω. However, we have shown that there is no positive equilibrium E^* when $R_0 < 1$. Hence system (3.1) has no positive periodic solution in Ω. By the Poincaré–Bendixson theory for the monotone algae system and the local stability of E_0, we see that E_0 is globally asymptotically stable. □

Proof of Theorem 3. An application of the quasi–steady state approximation for the cell quota equation yields

$$\tilde{Q} = \frac{Q_M \beta(P) + Q_m \mu_A h(A)}{\beta(P) + \mu_A h(A)} \equiv \gamma(A, P)$$

which is increasing in both A and P, with $\gamma(A, P) \in (Q_m, Q_M)$. The system (3.1) is then reduced to a two-dimensional system:

$$\text{(A.5)} \qquad \begin{cases} \dfrac{dA}{dt} = \mu_A A \left(1 - \dfrac{Q_m}{\gamma(A, P)}\right) h(A) - l_m A - \dfrac{\nu + D}{z_m} A \equiv F_1(A, P), \\ \dfrac{dP}{dt} = \dfrac{D}{z_m}(P_{in} - P) - \mu_A(\gamma(A, P) - Q_m) h(A) A \equiv F_2(A, P). \end{cases}$$

There are still two equilibria: $\tilde{E}_0 = (0, P_{in})$ and $\tilde{E}^* = (\bar{A}, \bar{P})$. All the theorems above for system (3.1) hold for system (A.5). Choose the Dulac function $\delta(A, P) = 1/A$. Then

$$\begin{aligned}\frac{\partial(\delta F_1)}{\partial A} &= \partial\left[\mu_A\left(1 - \frac{Q_m}{\gamma(A,P)}\right)h(A) - l_m - \frac{\nu + D}{z_m}\right]/\partial A\\ &= \partial\left[\mu_A\frac{(Q_M - Q_m)\beta(P)h(A)}{Q_M\beta(P) + Q_m\mu_A h(A)}\right]/\partial A < 0,\\ \frac{\partial(\delta F_2)}{\partial P} &= \partial\left[\frac{D}{z_m}(P_{in} - P)/A - \mu_A(\gamma(A,P) - Q_m)h(A)\right]/\partial P < 0.\end{aligned}$$

Therefore $\frac{\partial(\delta F_1)}{\partial A} + \frac{\partial(\delta F_2)}{\partial P} < 0$. By the Poincaré–Bendixson theory, the positive equilibrium $\tilde{E}^*$ is globally asymptotically stable for the system (A.5) when $R_0 > 1$. □

Proof of Theorem 4. The system (3.1) satisfies the conservation principle as follows:

$$\begin{aligned}\frac{d(P + AQ)}{dt} &= \frac{D}{z_m}(P_{in} - P) - \left(l_m + \frac{\nu + D}{z_m}\right)AQ\\ &= \frac{D}{z_m}(P_{in} - P) - \frac{D}{z_m}AQ = \frac{D}{z_m}[P_{in} - (P + AQ)];\end{aligned}$$

then, all solutions of system (3.1) asymptotically approach the surface $P + AQ = P_{in}$ as $t \to \infty$. We need only show that E^* is globally asymptotically stable on the surface $P + AQ = P_{in}$, which is the limiting case of system (3.1). The whole system is reduced to be a planar system on the surface; then, we can prove global stability on the surface as Theorem 3 when $R_0 > 1$. According to Smith and Waltman [35], E^* is also globally asymptotically stable for the system (3.1) when $R_0 > 1$. □

For convenience, in the following proofs we use the same notations Ω, Ω_1, Ω_2, $\wp$, F, M, etc. for system (2.2) as we did for system (3.1). Although they are different from those used for system (3.1), they play similar roles for system (2.2).

Proof of Theorem 5. For system (2.2), consider the total carbon $T = A + B/r + C$. Then $\frac{dT}{dt} = \mu_A Ah(A) - \frac{D}{z_m}T - (l_m + \frac{\nu}{z_m})A - \frac{\mu_r + \mu_g}{r}B \leq \mu_A Ah(A) - \frac{D}{z_m}T$, which gives us $\frac{dT}{dt} \leq \mu_A Th(T) - \frac{D}{z_m}T$, since $Ah(A)$ is increasing in A. By the condition $\mu_A h(0) < \frac{D}{z_m}$ and because $h(A)$ is decreasing in A, we have $\mu_A h(T) - \frac{D}{z_m} \leq \mu_A h(0) - \frac{D}{z_m} < 0$, which implies $T \to 0$ as $t \to \infty$. Together with positivity of all the variables, we have $\lim_{t\to\infty} A(t) = \lim_{t\to\infty} B(t) = 0$ for all nonnegative initial conditions; that is, both algae and bacteria go extinct. □

Acknowledgments. We are grateful to the referees for their many helpful and insightful suggestions. We would like to thank Val Smith for introducing the paper of Nishimura, Kim, and Nagata [32] to us, Irakli Loladze for an initial discussion, and James Cotner for comments on the bacterial cell quota. We also thank Sebastian Diehl for his helpful suggestions, and James Grover for his valuable comments on parameter estimations and conclusions.

REFERENCES

[1] T. Andersen, *Pelagic Nutrient Cycles: Herbivores as Sources and Sinks for Nutrients*, Springer-Verlag, Berlin, 1997.

[2] S. A. Berger, S. Diehl, T. J. Kunz, D. Albrecht, A. M. Oucible, and S. Ritzer, *Light supply, plankton biomass, and seston stoichiometry in a gradient of lake mixing depths*, Limnol. Oceanogr., 51 (2006), pp. 1898–1905.

[3] G. Bratbak and T. F. Thingstad, *Phytoplankton-bacteria interactions: An apparent paradox? Analysis of a model system with both competition and commensalism*, Mar. Ecol. Prog. Ser., 25 (1985), pp. 23–30.

[4] C. T. Codeço and J. P. Grover, *Competition along a spatial gradient of resource supply: A microbial experimental model*, Am. Nat., 157 (2001), pp. 300–315.

[5] J. P. Connolly, R. B. Coffin, and R. E. Landeck, *Modeling carbon utilization by bacteria in natural water systems*, in Modeling the Metabolic and Physiologic Activities of Microorganisms, C. Hurst, ed., John Wiley, New York, 1992, pp. 249–276.

[6] J. B. Cotner and B. A. Biddanda, *Small players, large role: Microbial influence on autoheterotrophic coupling and biogeochemical processes in aquatic ecosystems*, Ecosystems, 5 (2002), pp. 105–121.

[7] J. B. Cotner, *private communication*, 2006.

[8] S. Diehl, *Phytoplankton, light, and nutrients in a gradient of mixing depths: Theory*, Ecology, 83 (2002), pp. 386–398.

[9] S. Diehl, S. A. Berger, and R. Wöhrl, *Flexible algal nutrient stoichiometry mediates environmental influences on phytoplankton and its abiotic resources*, Ecology, 86 (2005), pp. 2931–2945.

[10] S. Diehl, *private communication*, 2006.

[11] W. T. Edmondson, *The Uses of Ecology: Lake Washington and Beyond*, University of Washington Press, Seattle, 1991.

[12] J. J. Elser, D. Dobberfuhl, N. A. MacKay, and J. H. Schampel, *Organism size, life history, and N:P stoichiometry: Towards a unified view of cellular and ecosystem processes*, BioScience, 46 (1996), pp. 674–684.

[13] P. A. del Giorgio and J. J. Cole, *Bacterial growth efficiency in natural aquatic systems*, Annu. Rev. Ecol. Syst., 29 (1998), pp. 503–541.

[14] J. P. Grover, *Resource Competition*, Population and Community Biology Series, Chapman & Hall, London, 1997.

[15] J. P. Grover, *Stoichiometry, herbivory and competition for nutrients: Simple models based on planktonic ecosystems*, J. Theoret. Biol., 214 (2002), pp. 599–618.

[16] J. P. Grover, *private communication*, 2006.

[17] D. O. Hessen and B. Bjerkeng, *A model approach to planktonic stoichiometry and consumer-resource stability*, Freshwater Biol., 38 (1997), pp. 447–472.

[18] D. O. Hessen, K. Nygaard, K. Salonen, and A. Vähätalo, *The effect of substrate stoichiometry on microbial activity and carbon degradation in humic lakes*, Environ. Int., 20 (1994), pp. 67–76.

[19] A. J. Horne and C. R. Goldman, *Limnology*, 2nd ed., McGraw–Hill, New York, 1994.

[20] J. Huisman and F. J. Weissing, *Light-limited growth and competition for light in well-mixed aquatic environments: An elementary model*, Ecology, 75 (1994), pp. 507–520.

[21] J. Huisman and F. J. Weissing, *Competition for nutrients and light in a mixed water column: A theoretical analysis*, Am. Nat., 146 (1995), pp. 536–564.

[22] C. A. Klausmeier and E. Litchman, *Algal games: The vertical distribution of phytoplankton in poorly mixed water columns*, Limnol. Oceanogr., 46 (2001), pp. 1998–2007.

[23] C. A. Klausmeier, E. Litchman, and S. A. Levin, *Phytoplankton growth and stoichiometry under multiple nutrient limitation*, Limnol. Oceanogr., 49 (2004), pp. 1463–1470.

[24] Y. Kuang, J. Huisman, and J. J. Elser, *Stoichiometric plant-herbivore models and their interpretation*, Math. Biosc. Eng., 1 (2004), pp. 215–222.

[25] L. D. J. Kuijper, B. W. Kooi, T. R. Anderson, and S. A. L. M. Kooijman, *Stoichiometry and food-chain dynamics*, Theor. Popul. Biol., 66 (2004), pp. 323–339.

[26] J. D. Logan, A. Joern, and W. Wolesensky, *Mathematical model of consumer homeostasis control in plant-herbivore dynamics*, Math. Comput. Modelling, 40 (2004), pp. 447–456.

[27] I. Loladze, Y. Kuang, and J. J. Elser, *Stoichiometry in producer-grazer systems: Linking energy flow with element cycling*, Bull. Math. Biol., 62 (2000), pp. 1137–1162.

[28] I. Loladze, Y. Kuang, J. J. Elser, and W. F. Fagan, *Coexistence of two predators on one prey mediated by stoichiometry*, Theor. Popul. Biol., 65 (2004), pp. 1–15.

[29] W. Makino, J. B. Cotner, R. W. Sterner, and J. J. Elser, *Are bacteria more like plants or animals? Growth rate and resource dependence of bacterial C:N:P stoichiometry*, Funct. Ecol., 17 (2003), pp. 121–130.

[30] W. A. Nelson, E. McCauley, and F. J. Wrona, *Multiple dynamics in a single predator-prey system: Experimental effects of food quality,* Proc. R. Soc. Lond. Ser. B., 268 (2001), pp. 1223–1230.

[31] J. A. Newman, D. J. Gibson, E. Hickam, M. Lorenz, E. Adams, L. Bybee, and R. Thompson, *Elevated carbon dioxide results in smaller populations of the bird cherry-oat aphid* Rhopalosiphum padi, Ecol. Entom., 24 (1999), pp. 486–489.

[32] Y. Nishimura, C. Kim, and T. Nagata, *Vertical and seasonal variations of bacterioplankton subgroups with different nucleic acid contents: Possible regulation by phosphorus*, Appl. Environ. Microbiol., 71 (2005), pp. 5828–5836.

[33] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[34] H. L. Smith and B. Li, *Competition for essential resources: A brief review*, Fields Inst. Commun., 36 (2003), pp. 213–227.

[35] H. L. Smith and P. Waltman, *The Theory of the Chemostat: Dynamics of Microbial Competition*, Cambridge University Press, Cambridge, UK, 1994.

[36] H. L. Smith and P. Waltman, *Perturbation of a globally stable steady state*, Proc. Amer. Math. Soc., 127 (1999), pp. 447–453.

[37] R. W. Sterner and J. J. Elser, *Ecological Stoichiometry—The Biology of Elements from Molecules to the Biosphere*, Princeton University Press, Princeton, NJ, 2002.

[38] R. W. Sterner, J. J. Elser, E. J. Fee, S. J. Guildford, and T. H. Chrzanowski, *The light: nutrient ratio in lakes: The balance of energy and materials affects ecosystem structure and process*, Am. Nat., 150 (1997), pp. 663–684.

[39] H. R. Thieme, *Mathematics in Population Biology*, Princeton Ser. Theor. Comput. Biol., Princeton University Press, Princeton, NJ, 2003.

[40] H. R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model)*, SIAM J. Math. Anal., 24 (1993), pp. 407–435.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 523–543

A DIFFUSION ANALYSIS APPROACH TO TE MODE PROPAGATION IN RANDOMLY PERTURBED OPTICAL WAVEGUIDES*

EMMANUEL PERREY-DEBAIN[†] AND I. DAVID ABRAHAMS[‡]

Abstract. The aim of this work is to model the evolution of the modal distribution of the electromagnetic field as it propagates along a randomly deformed multimode optical waveguide. When the number of guided modes becomes large we can regard the discrete set of modes as a quasi continuum. In some cases, nearest neighbor coupling predominates over other power transfer mechanisms and the coupling process can be ideally described in terms of a diffusion equation. The theory is applied to the propagation of guided transverse electric (TE) field waves in a slab waveguide with parabolic refractive index profile. Numerical simulations are in good agreement with theoretical results, and the error is shown to behave as the inverse of the number of guided modes. The technique allows the prediction of the long-distance modal distribution for a very large number of guided modes within fixed computational resources.

Key words. random waveguide, modal diffusion, optical fiber

AMS subject classifications. 78A45, 78A48, 78A50

DOI. 10.1137/060673874

1. Introduction.

While more and more sophisticated methods for the manufacture and control of graded index multimode fibers are being implemented, the random variations of the optical and geometrical properties of fibers from the ideal model are impossible to avoid. These small imperfections influence the signal propagation as a result of mode coupling, and their cumulative effects may become important after a long propagation length. Thus, they need to be taken into account when calculating the power attenuation, the signal distortion, and the bandwidth of the fiber [1, 2, 3]. The statistical treatment of wave propagation in random waveguides has been the topic of numerous papers, and a complete survey would merit a separate article; some interesting references can be found in [4].

The most common approach consists in deriving and solving the coupled power equations describing the evolution of the average power carried by the propagating modes. The earliest investigations of mode coupling in optical waveguides were concerned with the excess losses which result from the coupling of guided modes to radiation modes [5, 6]. Rowe and Young [7] showed that when random perturbations are present in a two-mode waveguide, one can derive coupled power equations for the power in each mode. Marcuse [8] generalized this result to any number of guided modes, and an excellent summary of this work can be found in his textbook [9]. When the number of guided modes becomes too large, a direct algebraic treatment of the coupled system is ruled out because of the computational overhead. In some cases,

[*] Received by the editors November 1, 2006; accepted for publication (in revised form) August 7, 2007; published electronically November 30, 2007. This research was carried out as part of the activities of the Smith Institute Knowledge Transfer Network (http://www.smithinst.ac.uk).

http://www.siam.org/journals/siap/68-2/67387.html

[†] Laboratoire Roberval, Université de Technologie de Compiègne, BP 60319-60203, Compiègne Cedex, France (emmanuel.perrey-debain@utc.fr). This author's research was supported by a post-doctoral grant from the UK Engineering and Physical Sciences Research Council (EPSRC) and Photon Design (Oxford, UK).

[‡] School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, England (i.d.abrahams@maths.manchester.ac.uk).

however, nearest neighbor coupling predominates over other power transfer mechanisms and, within appropriate limits, the coupling process can be ideally described in terms of a diffusion equation in which the mode number is treated as a continuous variable. This idea originated in the mid-seventies for dealing with the specific problem of random bends [10, 11, 12].

The diffusion analysis approach to mode propagation in optical fibers relies on many simplifications and assumptions which render the theory's validity difficult to estimate. Curiously enough, no progress has been made since the mid-seventies and, until recently, Gloge's diffusion theory [10] has been the starting point for evaluating mode conversion in step-index multimode fibers [13, 14]. In this paper (and in a forthcoming article discussing the three-dimensional waveguide [15]), we aim at offering a new contribution to the diffusion approach [16] by treating the problem in a much more rigorous manner. It is found that, for the specific case of a slab waveguide with a parabolic index profile, the coupled power equations system can be approximated as a diffusion equation with an approximation error of order $\mathcal{O}(N^{-1})$, where N is the number of modes. In practice, the theory leads to a numerically tractable problem for predicting the long-distance modal distribution of the transverse electric (TE) field for any waveguide supporting a sufficiently large number of modes. Furthermore, it allows one to identify nondiffusive regimes in which the modal power distribution is not the solution of a diffusion equation and which exhibits irregular behavior.

The structure of the paper is as follows. The statement of the problem is presented in section 2. In section 3, the standard coupled power equations for the slab waveguide are stated, and a continuous model is derived in section 4. We finally compare the theoretical results with numerical solutions of the diffusion equation for various cases in section 5.

2. Problem statement. We aim to study the propagation of a monochromatic TE field $E_Y = E(X,Z)e^{-i\omega t}$ in a weakly guiding two-dimensional dielectric waveguide whose parabolic graded-index profile n is affected by a small random perturbation, say, δn. The field is governed by the time-harmonic wave equation

$$\frac{\partial^2 E}{\partial X^2} + \frac{\partial^2 E}{\partial Z^2} + \kappa^2 n^2(X) E = \kappa^2 \delta n^2(X,Z) E, \tag{2.1}$$

where κ is the vacuum wavenumber, Z is the guide axis, and X is the transverse coordinate. The refractive index of the unperturbed waveguide has the parabolic profile

$$n^2(X) = n_0^2(1 - 2\Delta(X/a)^2) \tag{2.2}$$

in the waveguide region, $|X| \leq a$, and $n^2(X) = n_c^2$ in the infinite cladding, $|X| > a$. The profile height parameter $\Delta = (n_0^2 - n_c^2)/2n_0^2$ is assumed to be small, and backscattering is ignored so that, under appropriate scaling, the problem can be conveniently formulated [4] as the following Schrödinger-type equation for the amplitude $\Psi = Ee^{-i\kappa n_0 Z}$:

$$2i\frac{\partial \Psi}{\partial z} = -\frac{\partial^2 \Psi}{\partial x^2} + v(x)\Psi + \delta v(x,z)\Psi, \tag{2.3}$$

where $z = \sqrt{2\Delta}Z/a$, $x = \sqrt{V}X/a$, and $\delta v = V(n_0^2 - n_c^2)^{-1}\delta n^2$. Here V denotes the usual waveguide parameter $V = \kappa n_0 a\sqrt{2\Delta}$ and v stands for the quadratic potential with finite depth: $v(x) = Vf(\bar{x})$, where $\bar{x} = x/\sqrt{V} = X/a$ and f is the normalized

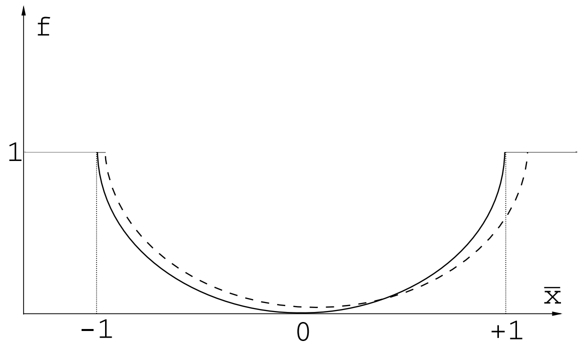


FIG. 2.1. *Profile of the normalized index of refraction; solid line indicates the unperturbed waveguide; dashed line indicates the waveguide under small deformation.*

quadratic profile (see Figure 2.1)

$$f(\bar{x}) = \left\{ \begin{array}{ll} \bar{x}^2, & |\bar{x}| \le 1 \\ 1, & |\bar{x}| > 1 \end{array} \right\}. \tag{2.4}$$

Let us now introduce ϵ, the relative amplitude of the perturbation of the slab profile. We can simulate small deviations from the ideal profile (2.4) as follows:

$$\tilde{f}(\bar{x}, z) = f(\bar{x}) + \delta f(\bar{x}, z) = (1 + \epsilon g_2(z))(\bar{x} + \epsilon g_1(z)/2)^2 + \epsilon g_3(z), \quad a^-(z) \le \bar{x} \le a^+(z).$$

The functions $g_q(z)$'s are random processes with amplitudes that do not exceed unity. In practice, g_1 simulates random oscillations of the center of the waveguide around the optical axis (microbending), g_2 is a random change of waveguide width, and g_3 is a random variation of the average refractive index. Note that the core-cladding interface is perturbed slightly from $\bar{x} = \pm 1$ to $\bar{x} = a^{\pm}(z)$. However, we will be interested in the behavior of the wave field away from the interface and so, as is usual, we will ignore the effects of the core-cladding interface on the mode intercoupling in this article. To first order, we find that the nondimensionalized perturbation to the refractive index is

$$\tilde{f}(\bar{x}, z) - f(\bar{x}) = \delta f(\bar{x}, z) = \epsilon(g_1(z)\bar{x} + g_2(z)\bar{x}^2 + g_3(z)) + O(\epsilon^2). \tag{2.5}$$

We can be more general, and henceforth take $\delta f = \epsilon g$, where the normalized random perturbation g is assumed to have the separable form

$$g(\bar{x}, z) = \sum_{q=1}^{Q} g_q(z)\phi_q(\bar{x}), \tag{2.6}$$

in which the g_q's are real-valued zero-mean, *independent*, stationary, and ergodic processes with respect to the waveguide axis coordinate z, and $\phi_q(\bar{x})$ are *deterministic* functions that can be referred to as the "perturbation modes." These functions are assumed to be sufficiently regular that they can be conveniently formulated as the truncated Taylor series

$$\phi_q(\bar{x}) = \sum_{n=0}^{\eta_q} b_{q,n}\bar{x}^n, \quad \text{where} \quad b_{q,n} = \frac{1}{n!}\left(\frac{d^n\phi_q}{d\bar{x}^n}\right)_{\bar{x}=0}. \tag{2.7}$$

3. Coupled mode theory. We wish to solve the parabolic equation (2.3), given the initial conditions $\Psi(x,0)$ at the input of the waveguide. For sufficiently small perturbations, it is possible to express the field distribution in the waveguide by using standard perturbation theory [17, 18, 19], i.e., Ψ is expanded in the eigenfunction basis of the unperturbed waveguide as

$$\Psi(x,z) = \sum_{\nu} a_{\nu}(z)\psi_{\nu}(x)e^{-i\beta_{\nu}z} + \sum \int a(\beta,z)\psi(\beta,x)e^{-i\beta z}d\beta. \tag{3.1}$$

The first summation extends over all the discrete spectrum of *guided modes* satisfying the eigenmode problem

$$-\frac{d^2\psi_{\nu}}{dx^2} + v(x)\psi_{\nu} = 2\beta_{\nu}\psi_{\nu}, \tag{3.2}$$

with the boundary condition $\psi_{\nu}(x) \to 0$ as $|x| \to \infty$. Eigenvalues of (3.2) are real positive quantities lying in the range $0 < \beta_{\nu} < \beta_c$ ($\beta_c = V/2$ is the cut-off wavenumber). They characterize the number of oscillations of the eigenfunctions along the transverse section of the waveguide. For a very large number of modes ($V \gg 1$), the highly oscillating wavefunction ψ_{ν} are conveniently described by the WKB approximation, and the largest permitted eigenvalue below cut-off can be shown to be approximately given by the upper bound, $\max_{\nu} \beta_{\nu} \approx \beta_c$. The integral in (3.1) extends over modes of the continuum (radiation modes), and the summation sign in front of the integral indicates summation over even and odd modes. These modes are oscillatory solutions of (3.2) and do not have the evanescent behavior (in x) of the guided mode fields. To be consistent with the forward scattering approximation introduced earlier, we restrict the integration domain to small propagation constants: $\beta_c \leq \beta \ll \beta_c/\Delta$. The expansion coefficients $a_{\nu}(z)$ and $a(\beta,z)$ are unknown functions of z. Using orthogonality properties of both guided and radiation modes, (2.3) is transformed into the system of ordinary differential equations

$$\begin{aligned}\frac{da_{\nu}}{dz} &= -i\epsilon\beta_c \sum_{\mu} C_{\nu,\mu}(z)e^{i(\beta_{\nu}-\beta_{\mu})z}a_{\mu}(z)\\ &\quad - i\epsilon\beta_c \sum \int a(\beta,z)C_{\nu}(\beta,z)e^{i(\beta_{\nu}-\beta)z}d\beta,\end{aligned} \tag{3.3}$$

where coupling coefficients are given by the overlap integrals

$$C_{\nu,\mu}(z) = \int_{-\infty}^{\infty} \psi_{\nu}(x)g(\bar{x},z)\psi_{\mu}(x)dx \tag{3.4}$$

and

$$C_{\nu}(\beta,z) = \int_{-\infty}^{\infty} \psi_{\nu}(x)g(\bar{x},z)\psi(\beta,x)dx. \tag{3.5}$$

Though feasible, a numerical solution of the coupled mode system (3.3) can be obtained at a heavy price, which could be well above standard computational resources. In fact, the solution of (3.3) contains more information (i.e., regarding the phase) than is required. It is now well established [9] that, under some additional assumptions, system (3.3) can be averaged over an ensemble of N_w similar waveguide realizations. More precisely, if the guided modes are weakly coupled over a distance which is large compared to the correlation length of the random process g, then the average power $A_\nu(z) = \langle |a_\nu(z)|^2 \rangle = \lim_{N_w \to \infty} \frac{1}{N_w} \sum_{N_w} |a_\nu(z)|^2$ carried by mode ν can be shown to satisfy the system of master equations

$$\frac{1}{\epsilon^2} \frac{dA_\nu}{dz} = \sum_\mu W_{\nu\to\mu}(A_\mu - A_\nu) - \alpha_\nu A_\nu, \tag{3.6}$$

where the transition probability matrix coefficients $W_{\nu\to\mu}$ are given from the spectral density of $C_{\nu,\mu}(z)$ evaluated at the wavenumber spacing $|\beta_\nu - \beta_\mu|$, i.e.,

$$W_{\nu\to\mu} = 2\beta_c^2 \int_0^\infty \langle C_{\nu,\mu}(0) C_{\nu,\mu}(z) \rangle \cos[(\beta_\nu - \beta_\mu)z] dz. \tag{3.7}$$

These are, by definition, positive quantities, and thus $W_{\nu\to\mu} \geq 0$. Note that a similar derivation can be found in the context of quantum mechanics [20] and in acoustics [21]. The power loss coefficients α_ν are positive quantities taking into account the coupling between mode ν to the continuum of radiation modes. A rigorous analysis of the radiation loss is a very difficult task as it requires both an accurate description of the perturbation in the vicinity of the core-cladding interface as well as precise knowledge of the mathematical form for the guided and radiation modes close to cut-off. Nevertheless, for waveguides supporting a sufficiently large number of modes, only highest order modes near cut-off carry nonnegligible energy near the interface and therefore suffer from very high losses. To simplify the analysis we will assume that $\alpha_\nu = \infty$ when $\nu \geq \beta_c$, which means that high order modes carry no power: $A_{\nu \geq \beta_c} = 0$. These assumptions were introduced by Marcuse [11] for the parabolic index fiber and were recently found to be in agreement with measurements carried out by Golowich et al. [3]. Due to the symmetry of the matrix coefficients $W_{\nu\to\mu}$, the solution of (3.6) is given explicitly by

$$\mathbf{A}(z) = \mathbf{U} \exp\bigl(\epsilon^2 \mathbf{\Lambda}\, z\bigr)\, \mathbf{U}^T\, \mathbf{A}(0), \tag{3.8}$$

where vector $\mathbf{A}(z) = (A_1(z), A_2(z), \ldots)^T$, and $\mathbf{A}(0)$ contains the initial conditions at $z = 0$, i.e.,

$$A_\nu(0) = |a_\nu(0)|^2 = \left| \int_{-\infty}^{\infty} \Psi(x, 0) \psi_\nu(x) dx \right|^2. \tag{3.9}$$

The diagonal matrix $(\mathbf{\Lambda})_{i,i} = \lambda_i$, $i = 1, 2, \ldots$, contains the real eigenvalues in descending order, and the column vectors $\mathbf{U}^{(i)} = (U_1^{(i)}, U_2^{(i)}, \ldots)^T$ are the eigenmodes of the real symmetric system with eigenvalues λ_i:

$$\mathbf{W}\mathbf{U}^{(i)} = \lambda_i \mathbf{U}^{(i)}, \tag{3.10}$$

with the cut-off condition that $U_\nu^{(i)} = 0$ when $\nu \geq \beta_c$. Note that the transition probability operator $\mathbf{W}$ is defined, from (3.6), as

$$(\mathbf{W}\mathbf{U}^{(i)})_\nu = \sum_\mu W_{\nu\to\mu} \left(U_\mu^{(i)} - U_\nu^{(i)} \right) - \alpha_\nu U_\nu^{(i)}. \tag{3.11}$$

Due to the special structure of (3.11) and the positivity of the off-diagonal terms, Dozier and Tappert [21] showed that Gerschgorin discs with radius $R_\nu = \sum_{\mu\neq\nu} W_{\nu\to\mu}$

lie in the left part of the complex plane, and hence all eigenvalues, λ_i, are negative. The special case, $\lambda_1 = 0$, can be referred to as the adiabatic case and corresponds to the long-distance solutions $\lim_{z\to\infty} A_\nu(z) = (\sum_\mu 1)^{-1} \sum_\mu A_\mu(0)$, i.e., an equipartition of energy is achieved whatever the initial conditions. This is a consequence of neglecting the radiation loss. This ideal scenario was considered in [21] but is obviously unrealistic in our context as losses from the highest order modes are unavoidable.

4. The continuous model.

4.1. Simplification when $V \gg 1$. Numerical diagonalization of the transition probability operator $\mathbf{W}$ (3.11) becomes impractical for very large V. Nevertheless, progress can be made if we are interested only in the lower $|\lambda_i|$ corresponding to long-distance solutions. To achieve this, we need to find a continuum analogue of (3.10). Let us first observe that in the limit of large V, the set of orthonormal functions ψ_ν satisfying (3.2) are the classical harmonic oscillator bases [18, 20]:

$$\psi_\nu(x) = \frac{1}{\sqrt{\pi^{1/2} 2^\nu \nu!}} H_\nu(x) e^{-x^2/2}, \quad \nu = 0, 1, 2, \ldots, \quad \text{with} \quad \beta_\nu = \nu + 1/2, \tag{4.1}$$

where H_ν denotes the usual Hermite polynomials. These are good approximations to the exact solutions, except for modes near cut-off, $\beta_\nu \approx \beta_c$. We assume that (4.1) is valid for all modes below cut-off; these modes are unaffected by the interface core-cladding as the power carried in this region is negligible and the evaluation of the coupling coefficients can be greatly simplified by extending the perturbation (2.6) over the whole real line as

$$C_{\nu,\mu}(z) = \sum_{q=1}^{Q} \sum_{n=0}^{\eta_q} g_q(z) b_{q,n} \int_{-\infty}^{\infty} \psi_\nu(x) \bar{x}^n \psi_\mu(x)\, dx. \tag{4.2}$$

To make some progress, we can observe that Hermite polynomials fall into the class of orthogonal polynomials satisfying a three-term recurrence relation which, in terms of the normalized function ψ_ν, reads

$$x\psi_\nu(x) = \frac{1}{\sqrt{2}} \left(\sqrt{\nu}\psi_{\nu-1}(x) + \sqrt{\nu+1}\psi_{\nu+1}(x) \right). \tag{4.3}$$

Using purely algebraic arguments, the nth power of the two-term recurrence operator (4.3) is established in [22]. This leads to following result.

LEMMA 4.1. *Given positive integers $(\zeta, n) \in \mathbb{N}^2$, the following integration formula holds:*

$$\int_{-\infty}^{\infty} \psi_\nu(x) x^n \psi_{\nu+\zeta}(x) dx = 2^{-\frac{n}{2}} F_\zeta(\nu) G_{\zeta,n}(\nu), \tag{4.4}$$

where

$$F_\zeta(\nu) = \prod_{l=1}^{\zeta} \sqrt{\nu + l} \quad \text{and} \quad G_{\zeta,n}(\nu) = \sum_{\sigma \in \frac{n-\zeta}{2} \cap \mathbb{N}} \; \sum_{\underline{i}_\sigma \in \mathcal{I}_\sigma^n} \prod_{l=1}^{\sigma} (\nu + 1 + i_l - l) \tag{4.5}$$

and $\mathcal{I}_\sigma^n$ is the set of indices $\underline{i}_\sigma = (i_1, \ldots, i_\sigma) \in \mathbb{N}^\sigma$ associated with the nested sum

$$\sum_{\underline{i}_\sigma \in \mathcal{I}_\sigma^n} = \sum_{i_\sigma=0}^{n-\sigma} \sum_{i_{\sigma-1}=0}^{i_\sigma} \cdots \sum_{i_2=0}^{i_3} \sum_{i_1=0}^{i_2}. \tag{4.6}$$

By common convention, the products above take the value unity when the lower limit exceeds the upper, and the notation $\sum_{\sigma\in\xi\cap\mathbb{N}}$ indicates that σ takes the value of ξ when ξ is an integer, or else the sum is zero. Note that the parameter ν has been written as an argument of F and G because we will soon generalize it to take noninteger values. Several other quantities will soon be defined which will also use this convention.

The result (4.4) shows that, for the ideal modes just described, the coupling coefficients between modes ν and $\nu+\zeta$ (ζ positive) take the form

$$C_{\nu,\nu+\zeta}(z) = \sum_{q=1}^{Q} g_q(z)\Phi_{q,\zeta}(\nu), \tag{4.7}$$

where

$$\Phi_{q,\zeta}(\nu) = F_\zeta(\nu)\sum_{n\geq 0} b_{q,n}(2V)^{-\frac{n}{2}}G_{\zeta,n}(\nu), \tag{4.8}$$

and $b_{q,n}$ is written in (2.7). The factorization of the quantity $F_\zeta(\nu)$ in (4.8) is a key result since it allows us to define the polynomial series w_ζ defined over the real line $\tilde{\nu}\in\mathbb{R}$ as

$$w_\zeta(\tilde{\nu}) = \sum_{q=1}^{Q} \Phi^2_{q,|\zeta|}(\tilde{\nu}-|\zeta|/2)\,\Gamma_q(\zeta), \tag{4.9}$$

where ζ now belongs to the whole integer set, $\zeta\in\mathbb{Z}$, and Γ_q stands for the spectral density function

$$\Gamma_q(\zeta) = 2\int_0^\infty \langle g_q(0)g_q(z)\rangle\cos(\zeta z)dz. \tag{4.10}$$

Finally the transition probability matrix coefficients $W_{\nu\to\nu+\zeta}$ are given from the regular function w_ζ evaluated at the midpoint $\nu+\zeta/2$, i.e.,

$$W_{\nu\to\nu+\zeta} = \beta_c^2\, w_\zeta(\nu+\zeta/2). \tag{4.11}$$

This is a key result of this article; it relates the transition probability matrix to a regular function over continuous arguments. This fact will be used shortly in obtaining a Taylor series expansion.

Let us now introduce η as the maximum exponent in the truncated Taylor expansion (2.7), i.e., $\eta=\max_{1\leq q\leq Q}\{\eta_q\}$. By virtue of (4.5), the transition matrix has a band-diagonal structure: $W_{\nu\to\nu+\zeta}=0$ when $|\zeta|>\eta$, and furthermore the roots of w_ζ are such that

$$W_{\nu\to -1} = \cdots = W_{\nu\to\nu-\eta} = 0. \tag{4.12}$$

Thus, there is no coupling with negative indices and (4.11) is exact for all guided modes. Let us now introduce a real analytic function $\tilde{U}_i$ which interpolates the discrete values of the elements of the column vector $\mathbf{U}^{(i)}$,

$$\tilde{U}_i(\nu) = U_\nu^{(i)} \quad \text{in the interval} \quad 0\leq\nu<\beta_c, \tag{4.13}$$

with the cut-off condition $\tilde{U}_i(\nu)=0$ in the interval $\beta_c\leq\nu\leq\beta_c+\eta$. This last condition does not need to be satisfied for $\nu>\beta_c+\eta$ due to the band-diagonal structure of the transition probability matrix. Similarly, we introduce a regular function $\tilde{\gamma}$ such that $\tilde{\gamma}(\nu) = \alpha_\nu$ (the attenuation factor) for all modes below cut-off $0 \leq \nu < \beta_c$. Note that such interpolating functions exist, and it suffices to consider the Lagrange interpolation polynomial, for instance. We can now exploit the equality (4.11) to give the continuum analogue of (3.11). Direct application of Taylor's theorem for real analytic functions yields the following result.

Lemma 4.2. *Let $\tilde{U}_i(\tilde{\nu})$ be the real analytic function introduced earlier. Then we have*

$$
\begin{aligned}
(\mathbf{W}\mathbf{U}^{(i)})_\nu &= \sum_{\zeta=-\eta}^{\eta} W_{\nu\to\nu+\zeta}\left(\tilde{U}_i(\nu+\zeta) - \tilde{U}_i(\nu)\right) - \tilde{\gamma}(\nu)\tilde{U}_i(\nu) \\
&= \beta_c^2 \frac{d}{d\tilde{\nu}}\left(D(\tilde{\nu})\frac{d\tilde{U}_i}{d\tilde{\nu}}\right)\Bigg|_{\tilde{\nu}=\nu} - \tilde{\gamma}(\nu)\tilde{U}_i(\nu) + \mathcal{R}(\tilde{U}_i)(\nu), \qquad (4.14)
\end{aligned}
$$

where $D(\tilde{\nu}) = \sum_{\zeta=1}^{\eta} \zeta^2 w_\zeta(\tilde{\nu})$ and the residual term $\mathcal{R}(\tilde{U}_i)$ is given by the Taylor series

$$
\mathcal{R}(\tilde{U}_i) = 4\beta_c^2 \sum_{(n,m)\in\mathbb{N}_*^2} \sum_{\zeta=1}^{\eta} \frac{(\zeta/2)^{2(n+m+1)}}{(2n+1)!(2m+1)!} \frac{d^{2m+1}}{d\tilde{\nu}^{2m+1}}\left(w_\zeta \frac{d^{2n+1}\tilde{U}_i}{d\tilde{\nu}^{2n+1}}\right),
$$

where $\mathbb{N}_*^2 = \mathbb{N}^2 \backslash \{(0,0)\}$.

Proof. First, we have the following equality, taking an expansion about the midpoint $\nu + \zeta/2$:

$$
\tilde{U}_i(\nu+\zeta) - \tilde{U}_i(\nu) = \zeta \frac{d\tilde{U}_i}{d\tilde{\nu}}(\nu+\zeta/2) + 2\sum_{n=1}^{\infty} \frac{(\zeta/2)^{2n+1}}{(2n+1)!}\frac{d^{2n+1}\tilde{U}_i}{d\tilde{\nu}^{2n+1}}(\nu+\zeta/2). \qquad (4.15)
$$

Now, functions w_ζ are polynomial series and so, given (4.11), we find

$$
\begin{aligned}
\left(w_\zeta \frac{d\tilde{U}_i}{d\tilde{\nu}}\right)(\nu+\zeta/2) - \left(w_\zeta \frac{d\tilde{U}_i}{d\tilde{\nu}}\right)(\nu-\zeta/2) &= \zeta \frac{d}{d\tilde{\nu}}\left(w_\zeta \frac{d\tilde{U}_i}{d\tilde{\nu}}\right)(\nu) \\
&\quad + 2\sum_{n=1}^{\infty} \frac{(\zeta/2)^{2n+1}}{(2n+1)!}\frac{d^{2n+1}}{d\tilde{\nu}^{2n+1}}\left(w_\zeta \frac{d\tilde{U}_i}{d\tilde{\nu}}\right)(\nu).
\end{aligned}
$$

Repeating this operation for the residual term on the right-hand side of (4.15) yields the expected result. □

The reader's attention is drawn to the fact that the leading term on the right-hand side of (4.14) is just the diffusion operator acting on $\tilde{U}_i$. We comment on this further below.

4.2. Asymptotic series expansion. We are interested in the eigenmode solutions of (3.10) when the number of guided modes is sufficiently large. To do this, we introduce the small parameter $\varepsilon = 1/V$. The asymptotic approach starts by writing every quantity as a power series of ε, and equating coefficients of like powers to yield a hierarchy of equations. Let us first observe that by introducing the normalized variables $u = \beta_c^{-1}\tilde{\nu}$, where $\beta_c^{-1} = 2/V = 2\varepsilon \ll 1$, function w_ζ admits the regular series expansion

$$
w_\zeta(\tilde{\nu}) = W_\zeta^0(u) + \varepsilon W_\zeta^1(u) + \varepsilon^2 W_\zeta^2(u) + \cdots + \varepsilon^\eta W_\zeta^\eta(u), \qquad (4.16)
$$

where

$$W^{\alpha}_{\zeta}(u) = \lim_{\varepsilon \to 0} \frac{1}{\alpha!} \frac{\partial^{\alpha} w_{\zeta}(u\,(2\varepsilon)^{-1})}{\partial \varepsilon^{\alpha}}, \quad \alpha = 0, 1, 2, \dots, \eta, \tag{4.17}$$

are all polynomial series with respect to the continuous variable u. The first two terms are given explicitly by

$$W^{0}_{\zeta}(u) = u^{\zeta} \sum_{q=1}^{Q} \Gamma_q(\zeta)[P^{I}_{q,\zeta}(u)]^2 \tag{4.18}$$

and

$$W^{1}_{\zeta}(u) = u^{\zeta-1} \sum_{q=1}^{Q} \Gamma_q(\zeta) P^{I}_{q,\zeta}(u)[\zeta P^{I}_{q,\zeta}(u) + 4P^{II}_{q,\zeta}(u)], \tag{4.19}$$

where

$$P^{I}_{q,\zeta}(u) = \sum_{\sigma=0}^{\lfloor(\eta_q-\zeta)/2\rfloor} 2^{-n} b_{q,n} C^{n}_{\sigma} u^{\sigma}|_{n=\zeta+2\sigma} \tag{4.20}$$

and

$$P^{II}_{q,\zeta}(u) = \sum_{\sigma=1}^{\lfloor(\eta_q-\zeta)/2\rfloor} 2^{-n} b_{q,n} A^{n}_{\sigma} u^{\sigma}|_{n=\zeta+2\sigma} \tag{4.21}$$

with

$$C^{n}_{\sigma} = \mathrm{card}(\mathcal{I}^{n}_{\sigma}) = \frac{n!}{\sigma!(n-\sigma)!} \quad \text{and} \quad A^{n}_{\sigma} = \sum_{\underline{\mathrm{i}}_{\sigma} \in \mathcal{I}^{n}_{\sigma}} \sum_{l=1}^{\sigma} (1 + i_l - l - n/2 + \sigma).$$

Note that the quantity $\lfloor x \rfloor$ in these equations indicates the floor of x, i.e., the largest integer less than or equal to the real number x, and η_q (from (2.7)) is the maximum value of n with nonzero $b_{q,n}$. The expansion (4.16) suggests writing a solution $\tilde{U}_i(\tilde{\nu})$ in the form of an asymptotic series expansion

$$\tilde{U}_i(\tilde{\nu}) = U_{i,0}(u) + \varepsilon U_{i,1}(u) + \varepsilon^2 U_{i,2}(u) + \cdots \tag{4.22}$$

and

$$\lambda_i = \lambda_{i,0} + \varepsilon \lambda_{i,1} + \varepsilon^2 \lambda_{i,2} + \cdots . \tag{4.23}$$

Similarly, we may assume that

$$\tilde{\gamma}(\tilde{\nu}) = \gamma_0(u) + \varepsilon \gamma_1(u) + \varepsilon^2 \gamma_2(u) + \cdots . \tag{4.24}$$

Substituting these expansions into (3.10) yields a series of diffusion equations (for brevity we restrict ourselves to writing just the leading order and the first order corrections):

$$\frac{d}{du}\left(D_0(u)\frac{dU_{i,0}}{du}\right) - \gamma_0\, U_{i,0} = \lambda_{i,0} U_{i,0}, \tag{4.25}$$

$$\begin{aligned}\frac{d}{du}\left(D_0(u)\frac{dU_{i,1}}{du}\right) - \gamma_0\, U_{i,1} &= \lambda_{i,0} U_{i,1} + \lambda_{i,1} U_{i,0} \\ &\quad + \gamma_1\, U_{i,0} - \frac{d}{du}\left(D_1(u)\frac{dU_{i,0}}{du}\right),\end{aligned} \tag{4.26}$$

where functions

$$D_0(u) = \sum_{\zeta=1}^{\eta} \zeta^2 W_\zeta^0(u) \quad \text{and} \quad D_1(u) = \sum_{\zeta=1}^{\eta} \zeta^2 W_\zeta^1(u) \tag{4.27}$$

can be interpreted as diffusion coefficients controlling the average transfer of modal power at the mode "number" $\tilde{\nu} = \beta_c u$. At subsequent orders, formulae are more complicated due to the presence of the residual term $\mathcal{R}(\tilde{U}_i)$. Each diffusion equation in the family must be solved over the unit interval $[0,1]$ and all have boundary data $U_{i,\alpha}(1) = 0$, $\alpha \in \mathbb{N}$. The boundary condition at the origin emerges naturally after realizing that there is no transfer of modal energy from negative indices; this implies that

$$\left(D_0 \frac{dU_{i,\alpha}}{du} \right)_{u=0} = 0. \tag{4.28}$$

This boundary condition is in fact the continuous analogue of the no-coupling condition (4.12) and is automatically satisfied for any *regular* solution since, by construction (see (4.18)), we have $D_0(0) = 0$.

4.3. Nature of the leading order solution. In this section, we are interested in *regular* solutions of the leading order eigenmode satisfying

$$\frac{d}{du}\left(D_0 \frac{d\varphi}{du} \right) - \gamma_0\, \varphi = \lambda \varphi, \quad \varphi(1) = 0. \tag{4.29}$$

From the remark below (3.11) we also require that $\lambda < 0$. Since $D_0(u)$ is a strictly positive polynomial in $(0,1]$ with $D_0(0) = 0$, (4.29) is a singular Sturm–Liouville eigenvalue problem and $u = 0$ is a singular endpoint. The regularity of the eigensolution therefore depends upon the behavior of $D_0(u)$ as u tends to zero. Fortunately, (4.29) admits exact analytical solutions for monomial perturbations $g(\bar{x}, z) = g_n(z)\bar{x}^n$ (recall that $\bar{x}$ is the scaled transverse coordinate given in section 2 by $\bar{x} = X/a$) because in these cases

$$D_0(u) = d_n u^n \quad \text{with} \quad d_n = 2^{-2n} \sum_{\sigma=0}^{[\frac{n-1}{2}]} (n - 2\sigma)^2 \Gamma_n(n - 2\sigma)(C_\sigma^n)^2 > 0; \tag{4.30}$$

by neglecting the loss term γ_0, it can be shown that the general family of solutions (up to the normalization constant) is

$$\varphi(u) = u^{(1-n)/2} J_{\pm\vartheta}(\omega_{n,i}^{\pm} u^{(2-n)/2}), \quad \vartheta = \frac{n-1}{n-2} \quad (n \geq 3), \tag{4.31}$$

where $\omega_{n,i}^{\pm}$ satisfies $J_{\pm\vartheta}(\omega_{n,i}^{\pm}) = 0$ and $J_{\pm\vartheta}$ is the Bessel function of the first kind of order $\pm\vartheta$. The particular case $n = 3$ yields $\vartheta = 2$ so an additional independent solution is given by $\varphi(u) = u^{-1} Y_2(y_{2,i} u^{-1/2})$, where Y_2 denotes the usual Bessel function of the second kind of order 2 and $y_{2,i}$ are the zeros of Y_2. When $n = 2$, (4.29) is the classical Euler–Cauchy equation with general solution $\varphi(u) = u^r$, where r satisfies the associated characteristic equation $r^2 + r - \lambda/d_2 = 0$. A quick inspection reveals that, as λ is negative, $\mathrm{Re}(r) < 0$. To summarize, for the specific case where the diffusion coefficient has the simple form $D_0(u) = d_n u^n$ and losses are neglected

($\gamma_0 = 0$), solutions of (4.29) are all singular at the origin except when $n = 1$, for which there exists a unique regular solution of the form

$$\varphi(u) = J_0(j_{0,i}\sqrt{u}), \tag{4.32}$$

where $j_{0,i}$ are the zeros of J_0 in ascending order. The other independent solution is $\varphi(u) = Y_0(y_{0,i}\sqrt{u})$, where Y_0 denotes the usual Bessel function of the second kind of order 0 and $y_{0,i}$ are the zeros of Y_0. This result suggests that regular solutions are expected, provided the diffusion coefficient has *linear* behavior as $u \to 0$. This is confirmed by the following proposition.

Lemma 4.3. *There exists a unique power series solution to the Sturm–Liouville problem* (4.29)*, provided $D_0(u) \sim u$ as $u \to 0$, which is equivalent to*

$$\Gamma_q(1)\, b_{q,1}^2 \neq 0, \quad q = 1, \dots, Q. \tag{4.33}$$

Proof. The diffusion coefficient $D_0(u)$ has the general polynomial form

$$D_0(u) = \sum_{n=1}^{\eta} d_n u^n, \tag{4.34}$$

where the first coefficient is explicitly given by $d_1 = \frac{1}{4}\sum_{q=1}^{Q} \Gamma_q(1) b_{q,1}^2$. Without loss of generality, we may assume that $\gamma_0(u)$ has a power series expansion. Substituting the Frobenius–Fuchs series

$$\varphi(u) = u^c \sum_{j=0}^{\infty} a_j u^j, \quad a_0 \neq 0, \tag{4.35}$$

in (4.29) leads to the indicial equation: $d_1 a_0 c^2 = 0$ and the existence of a power series solution is guaranteed if $d_1 \neq 0$, which is equivalent to (4.33) since the Γ_q's are positive functions. Furthermore, the series is unique due to Fuchs' theorem. Note the associated eigenvalues can be checked to be real negative since the regularity of φ implies that

$$\lambda \int_0^1 \varphi^2\, du = -\int_0^1 D_0 \left(\frac{d\varphi}{du}\right)^2 du - \int_0^1 \gamma_0 \varphi^2\, du, \tag{4.36}$$

where all the integrals are positive; the result is apparent by inspection. This is consistent with that found for the discrete eigenvalue problem (3.10).

If $d_1 = 0$ and $d_2 \neq 0$, then the indicial equation becomes

$$c^2 + c - \frac{\lambda + \gamma_0(0)}{d_2} = 0. \tag{4.37}$$

In the limit of large V, the energy carried by the fundamental mode $\nu = 0$ is vanishingly small at the core-cladding interface, so $\lim_{V\to\infty} \tilde{\gamma}(0) = 0$ and therefore $\gamma_0(0) = 0$. Now, given the fact that $\lambda < 0$ and $d_2 > 0$, roots of the quadratic form have strictly negative real part, and this leads to singular solutions. When $d_1 = d_2 = 0$, $u = 0$ is an irregular singular point and there is no series solution. □

In order to give a physical explanation of the condition (4.33), for simplicity let us assume that the waveguide is affected by a single perturbation mode, i.e.,

$g(\bar{x},z) = g_1(z)\phi_1(\bar{x})$. As the number of guided modes tends to infinity, the modal distribution tends to the solution of a diffusion equation provided that (see (4.10))

$$\Gamma_1(1) = 2\int_0^\infty \langle g_1(0)g_1(z)\rangle \cos(z)dz \neq 0 \quad \text{and} \quad b_{1,1} = \left(\frac{d\phi_1}{d\bar{x}}\right)_{\bar{x}=0} \neq 0. \tag{4.38}$$

The first inequality is nothing other than the well-known "resonance" condition [9] to ensure the coupling between two adjacent modes with equal spacing in β space; in other words, the perturbation g_1 must have spatial frequency support at $\beta_{\nu+1}-\beta_\nu = 1$. The second condition means that the perturbation cannot be locally flat in the vicinity of the waveguide axis. If it is, then as $V \to \infty$ the lowest order modes, localized very near the axis, will not "see" any perturbation at all and there will be no coupling and therefore no modal diffusion. This behavior is illustrated numerically in the last section of this paper.

4.4. Regular solution and first order correction. Given a random perturbation satisfying (4.33), we call $\{U_{i,0}\}_{i=1}^\infty$ the set of regular solutions of the self-adjoint eigenvalue problem (4.25). Assuming that the associated eigenvalues are all distinct, the following orthogonality property holds:

$$\int_0^1 U_{i,0}U_{k,0}\, du = \|U_{i,0}\|^2\delta_{i,k}, \tag{4.39}$$

where $\|\cdot\|$ stands for the usual energy norm of $L^2([0,1])$. The orthogonality of the eigenfunctions is in line with the orthogonality of the eigenvectors

$$\mathbf{U}^{(i)} \cdot \mathbf{U}^{(k)} = \delta_{i,k}. \tag{4.40}$$

Let us define vectors $\mathbf{V}_\alpha^{(i)}$, $\alpha = 0,1,2,\ldots$, as the discrete versions of their continuous counterpart: $(\mathbf{V}_\alpha^{(i)})_\nu = U_{i,\alpha}(\nu/\beta_c)$ for all guided modes. Rewriting the perturbation expansion (4.22) in its vectorial form gives

$$\mathbf{U}^{(i)} = \mathbf{V}_0^{(i)} + \varepsilon\mathbf{V}_1^{(i)} + \varepsilon^2\mathbf{V}_2^{(i)} + \cdots. \tag{4.41}$$

According to (4.40), the norm of the leading order solution $\|U_{i,0}\|$ must be chosen such that $\|\mathbf{V}_0^{(i)}\|_2^2 = 1 + \mathcal{O}(\varepsilon)$. This can easily be shown to be satisfied by simply taking

$$\|U_{i,0}\| = \beta_c^{-\frac{1}{2}} = \sqrt{2}\varepsilon^{\frac{1}{2}}. \tag{4.42}$$

The first order correction is explicitly obtained by expanding $U_{i,1}$ in the leading order orthogonal basis $U_{i,0}$, i.e.,

$$U_{i,1}(u) = \sum_{k=1}^\infty v_{i,k}\, U_{k,0}(u). \tag{4.43}$$

Substituting (4.43) into (4.26) and using orthogonality properties yields

$$v_{i,k} = \frac{(2\varepsilon)^{-1}}{\lambda_{i,0} - \lambda_{k,0}} \int_0^1 \left[\frac{d}{du}\left(D_1\frac{dU_{i,0}}{du}\right)U_{k,0} - \gamma_1 U_{i,0}U_{k,0}\right] du, \quad k \neq i, \tag{4.44}$$

and

$$\lambda_{i,1} = (2\varepsilon)^{-1} \int_0^1 \left[\frac{d}{du}\left(D_1 \frac{dU_{i,0}}{du} \right) U_{i,0} - \gamma_1 U_{i,0} U_{i,0} \right] du. \tag{4.45}$$

The diagonal correction terms $v_{i,i}$ stem from the discrete normalization (4.40). To first order, this is equivalent to the condition

$$\|\mathbf{V}_0^{(i)} + \varepsilon \mathbf{V}_1^{(i)}\|_2^2 = 1 + \mathcal{O}(\varepsilon^2). \tag{4.46}$$

The correspondence between the discrete and continuous norms is given by the composite trapezoidal rule:

$$\int_0^{\beta_c} (U_{i,0}(\tilde{\nu}/\beta_c) + \varepsilon U_{i,1}(\tilde{\nu}/\beta_c))^2 \, d\tilde{\nu} = \|\mathbf{V}_0^{(i)} + \varepsilon \mathbf{V}_1^{(i)}\|_2^2 - \frac{1}{2}(U_{i,0}(0) + \varepsilon U_{i,1}(0))^2 + T^{(i)},$$

where the quadrature error $T^{(i)}$ is bounded by

$$|T^{(i)}| \le \frac{\varepsilon}{6} \max_{u \in [0,1]} \left| \frac{d^2 (U_{i,0} + \varepsilon U_{i,1})^2}{du^2}(u) \right|.$$

Thus, due to the normalization (4.42), $T^{(i)} \sim \mathcal{O}(\varepsilon^2)$. Moreover, by construction

$$\int_0^{\beta_c} (U_{i,0}\,(\tilde{\nu}/\beta_c) + \varepsilon U_{i,1}(\tilde{\nu}/\beta_c))^2 \, d\tilde{\nu} = 1 + \varepsilon 2 v_{i,i} + \mathcal{O}(\varepsilon^2),$$

and so the normalization condition (4.46) is satisfied if

$$v_{i,i} = -\frac{\varepsilon^{-1}}{4} U_{i,0}^2(0). \tag{4.47}$$

Note that the orthogonality of the first order eigenvectors is checked in Appendix A, confirming the above analysis.

5. Numerical experiments.

5.1. Linear perturbation. In this section we shall focus on the linear perturbation $g(\bar{x}, z) = g_1(z)\bar{x}$. This arises from random changes in the direction of the waveguide axis. From (4.20)–(4.21), $P_{1,1}^I(u) = 2^{-1}$ and $P_{1,1}^{II}(u) = 0$, which, taking $Q = \eta = 1$ in (4.18), (4.19), and (4.27), leads to the simple form for the diffusion coefficients: $D_0(u) = u\Gamma_1(1)/4$ and $D_1(u) = \Gamma_1(1)/4$. To simplify the analysis, the spectral density of g_1 is chosen so that (see (4.10)) $\Gamma_1(1) = 4$. This choice gives

$$D_0(u) = u \quad \text{and} \quad D_1(u) = 1. \tag{5.1}$$

By neglecting the radiation losses, we get the leading order solution $U_{i,0} \equiv \varphi_i$ (see Appendix B). The leading order eigenvalue is $\lambda_{i,0} = -j_{0,i}^2/4$, and we show in the appendix that the integral (4.45) can be evaluated analytically to yield $\lambda_{i,1} = -\lambda_{i,0}$. However, an analytical form for (4.44) could not be found and so numerical integration has to be performed. To summarize, the eigenvector solution of the original eigenvalue problem (3.10) is, to first order,

$$\mathbf{U}^{(i)} = \mathbf{V}_0^{(i)} + \varepsilon \sum_{k=1}^{\infty} v_{i,k} \mathbf{V}_0^{(k)} + \mathcal{O}(\varepsilon^2), \tag{5.2}$$

Table 5.1
Evolution of the first eigenvalue.

V	Discrete system	Continuous model
10	−1.317801	−1.301216
100	−1.431525	−1.431338
500	−1.442912	−1.442904
2500	−1.445218	−1.445218

Table 5.2
Evolution of the second eigenvalue.

V	Discrete system	Continuous model
10	−7.067015	−6.856034
100	−7.544146	−7.541637
500	−7.602682	−7.602579
2500	−7.614772	−7.614768

where $(\mathbf{V}_0^{(i)})_\nu = (\varphi_i(\nu/\beta_c))$. The off-diagonal terms are given explicitly by

$$v_{i,k} = \frac{1 - \frac{j_{0,k}}{j_{0,i}} \int_0^1 u^{-1} J_1(\sqrt{u} j_{0,i}) J_1(\sqrt{u} j_{0,k})\, du}{\left(\frac{j_{0,k}^2}{j_{0,i}^2} - 1\right) |J_1(j_{0,i}) J_1(j_{0,k})|} \tag{5.3}$$

and $v_{i,i} = -1/(2J_1^2(j_{0,i}))$. The eigenvalues are, from (4.23), found to be

$$\lambda_i = -\frac{j_{0,i}^2}{4}(1-\varepsilon) + \mathcal{O}(\varepsilon^2). \tag{5.4}$$

In this example, the associated original discrete system (3.10) may be written explicitly as

$$(\nu+1)(U_{\nu+1}^{(i)} - U_\nu^{(i)}) + \nu(U_{\nu-1}^{(i)} - U_\nu^{(i)}) = \lambda_i U_\nu^{(i)}. \tag{5.5}$$

So, mode coupling occurs only between adjacent modes, and the description of the power coupling process in terms of a diffusion equation is clearly validated since (5.5) is nothing but the finite difference discretization of (B.1).

Tables 5.1 and 5.2 display the values of the first two eigenvalues calculated from the original discrete system (5.5) and using the first order approximation (5.4). The number of digits of accuracy given by the continuous model is in agreement with the expected V^{-2} law; recall that the above expressions are correct to $\mathcal{O}(\varepsilon^2)$, $\varepsilon \to 0$, which is equivalent to $\mathcal{O}(V^{-2})$, $V \to \infty$. This is clearly confirmed in Figure 5.1, where the evolution of the quadratic errors (in percentages) for the leading order solutions

$$E_0^{(i)} = 100 \times \|\mathbf{U}^{(i)} - \mathbf{V}_0^{(i)}\|_2$$

and for the first order solutions

$$E_1^{(i)} = 100 \times \|\mathbf{U}^{(i)} - \mathbf{V}_0^{(i)} - \varepsilon \mathbf{V}_1^{(i)}\|_2$$

are plotted against the waveguide parameter V. Note that the first order correction vector $\mathbf{V}_1^{(i)}$ is computed with only the first 30 terms in the infinite sum in (5.2).

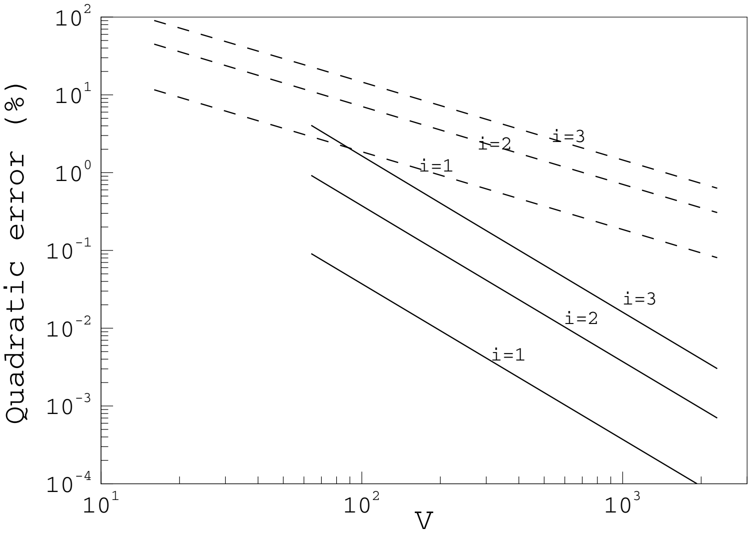


FIG. 5.1. *Error behavior with respect to V for the first three eigenmodes; dashed lines denote leading order solution, and straight lines denote first order solution.*

5.2. General case: The Rayleigh–Ritz method. We consider a random perturbation satisfying (4.33). Regular solutions of the Sturm–Liouville eigenvalue problem (4.25) can be numerically recovered using the Rayleigh–Ritz method [23]. For the sake of notational simplicity, we define the symmetric positive bilinear form

$$\mathcal{A}(\varphi,\varphi') = \int_0^1 \left(D_0(u)\frac{d\varphi}{du}\frac{d\varphi'}{du} + \gamma_0(u)\varphi\varphi' \right) du. \tag{5.6}$$

Solutions of (4.25) are stationary points of the energy functional $\mathcal{E}[U_{i,0}] = \mathcal{A}(U_{i,0}, U_{i,0})$ subject to the normalization constraint $\|U_{i,0}\| = \beta_c^{-\frac{1}{2}}$. The set of functions $\{\varphi_k\}_{k=1}^{\infty}$ forms a complete orthogonal system on $L^2([0,1])$ satisfying the Dirichlet boundary condition at the endpoint $u = 1$ (see Appendix B) and can therefore serve as a natural basis for an approximate solution $U_{i,0}^K$; i.e., we consider the truncated generalized Fourier expansion

$$U_{i,0}^K(u) = \sum_{k=1}^{K} c_{i,k}^K \, \varphi_k(u). \tag{5.7}$$

Following standard variational techniques, the stationary points are reached at the approximate eigenvalue $|\lambda_{i,0}^K| = \beta_c \, \mathcal{E}[U_{i,0}^K]$, where the expansion coefficients $c_{i,k}^K$ satisfy the matrix eigenvalue problem,

$$\sum_{k=1}^{K} c_{i,k}^K \, \mathcal{A}(\varphi_k, \varphi_l) = -\lambda_{i,0}^K \, c_{i,l}^K, \qquad l = 1, 2, \dots, K. \tag{5.8}$$

The symmetry of the bilinear form implies that the basis set $(U_{i,0}^K)_{i=1,\dots,K}$ is orthogonal in $L_2([0,1])$. Furthermore, the true solution is recovered by taking the limit

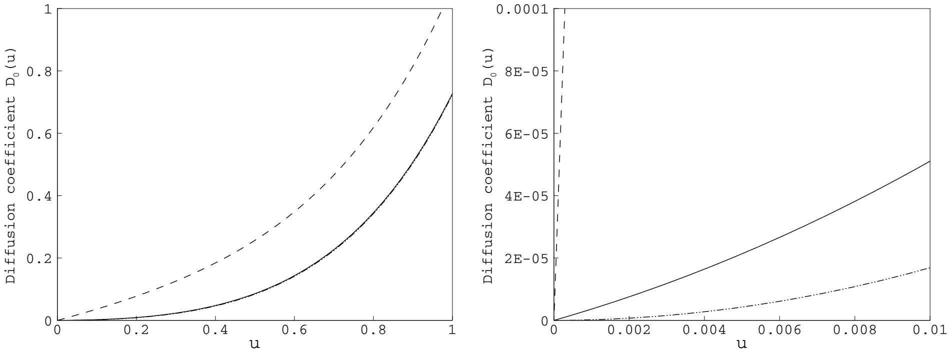


FIG. 5.2. *Leading order diffusion coefficient* $D_0(u)$. *Dashed line:* $\tau = 1$*; continuous line:* $\tau = 0.01$*; dash-dot-dot line:* $\tau = 0.0001$.

$U_{i,0} = \lim_{K\to\infty} U_{i,0}^K$. Similarly, $\lambda_{i,0} = \lim_{K\to\infty} \lambda_{i,0}^K$ and from the Rayleigh–Ritz principle, approximate eigenvalues $|\lambda_{i,0}^K|$ are upper bounds for the true eigenvalues $|\lambda_{i,0}|$ of the infinite-dimensional problem. The convergence of the method depends upon the properties of the perturbation such as its shape and its power spectrum. In most cases of practical interest, lowest order eigenvalues are expected to be obtained at a modest computational price (say $K \le 100$). For the sake of illustration, we consider perturbations given by the general form

$$g(\bar{x}, z) = \tau g_1(z)\bar{x} + \sum_{q=2}^{5} g_q(z)\bar{x}^q. \tag{5.9}$$

In the current analysis, the random functions g_q are assumed to be statistically identical and satisfy the Gaussian distribution: $\langle g_q(0)g_q(z)\rangle = e^{-z^2}$, $q = 1, \dots, 5$, which gives $\Gamma_q(\zeta) = \sqrt{\pi}e^{-\zeta^2/4}$. The corresponding diffusion coefficient is the polynomial of degree 5,

$$D_0(u) = \tau\frac{\sqrt{\pi}e^{-1/4}}{4}u + \sum_{n=2}^{5} d_n u^n. \tag{5.10}$$

The other coefficients are given explicitly in (4.30). In Figure 5.2 are plotted two graphs of D_0 against u for three values of τ. The tiny difference between the two curves $\tau = 0.01$ and $\tau = 0.0001$ can be identified on the magnified figure on the right. When $\tau = 0$, criterion (4.33) is not satisfied and there is no continuous counterpart to the discrete eigenmode. Thus, the coefficient τ can be interpreted as a diffusion parameter and the modal distribution in the waveguide is expected to "lose its regularity" when $\tau \to 0$. This behavior is revealed in Figures 5.3–5.4, where the first and fifth eigenmodes obtained from the discrete system (3.10) and from the continuous model (4.25) are shown. In all cases illustrated we considered $\beta_c = V/2 = 500$ guided modes, and the eigenfunctions $U_{i,0}^K$ are computed with $K = 50$ basis functions. From (4.44), it can be shown that the amplitude of the first order correction terms $v_{i,k}$ will grow as $\tau \to 0$. Thus the leading order solution $U_{i,0}$ will be a good approximation only if the number of guided modes is sufficiently large so that $\varepsilon\|U_{i,1}\| \ll \|U_{i,0}\|$. This explains the discrepancy observed when $\tau = 0.0001$.

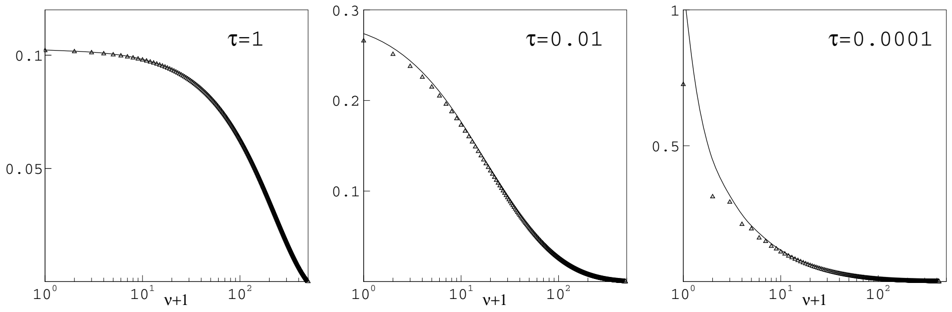


FIG. 5.3. *Influence of the diffusion parameter τ on the "regularity" of the first eigenmode; unbroken line denotes the continuous model $U_{1,0}$, and triangles denote the discrete eigenmode $\mathbf{U}^{(1)}$. Number of guided modes:* 500*.*

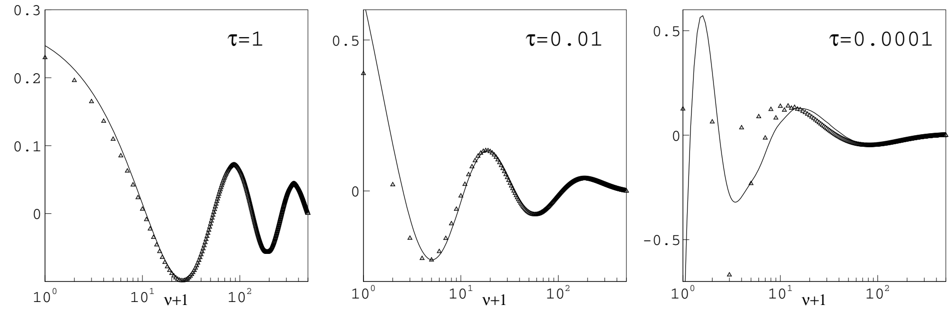


FIG. 5.4. *Influence of the diffusion parameter τ on the "regularity" of the fifth eigenmode; unbroken line denotes the continuous model $U_{5,0}$, and triangles denote the discrete eigenmode $\mathbf{U}^{(5)}$. Number of guided modes:* 500*.*

TABLE 5.3
Long distance power distribution. Fraction of energy carried by the lowest modes ($\beta_c = 500$).

τ	$\nu = 0, 1, \ldots, 10$	$\nu = 0, 1, \ldots, 100$
1	33%	82%
0.01	70%	98%
0.0001	95%	99.8%

The diffusion parameter has noticeable consequences on the modal distribution of $\mathbf{U}^{(1)}$ and therefore the long-distance power distribution

$$\mathbf{A}(z) \approx \mathbf{U}^{(1)} \exp\left(\epsilon^2 \lambda_1 z\right) \left[\mathbf{U}^{(1)}\right]^T \mathbf{A}(0) \quad \text{as} \quad z \to \infty. \tag{5.11}$$

This is clearly illustrated in Figure 5.3, where τ has a significant effect on the modal distribution. Since the fundamental mode $U_{1,0}$ is a positive function in $[0, 1]$ and $\lambda_{1,0} U_{1,0}(0) = \tau(dU_{1,0}/du)_{u=0}$ (see Appendix A), the optical power is likely to be concentrated among the lowest order modes as $\tau \to 0$. This is confirmed in Table 5.3, where the fraction of energy carried by the lowest modes is shown for various values of τ.

6. Conclusion. In this paper we have analyzed the evolution of the modal power distribution of the transverse electric field as it propagates along a multimode slab waveguide with quadratic refractive index profile and with small random deformations. We showed that for waveguides supporting a sufficiently large number of guided modes, the mode coupling mechanism can be ideally described as a diffusion equation. Even for a moderate number of modes, the regular expansion method allows us to obtain very accurate solutions when first order correction terms are taken into account. In practice, the technique described herein provides excellent qualitative predictions for the long-distance modal distribution within fixed computational resources regardless of the number of modes. Furthermore, we were able to identify nearly nondiffusive regimes in which the modal power distribution is not the solution of a diffusion equation and exhibits irregular behavior. In these latter scenarios, we observed strong focusing effects of the wave field in the vicinity of the waveguide axis. Work is almost complete on applying the present technique to three-dimensional fibers of circular cross-section with parabolic index profile [15]. We have good reason to believe that the results demonstrated in this paper could be generalized to other graded-index fibers, and this will be the subject of future work.

Appendix A. Orthogonality of the first order eigenvectors. The purpose of this appendix is to prove that the eigenmodes of the discrete system, taken to first order in ε, are orthogonal to $\mathcal{O}(\varepsilon^2)$. We commence by integrating by parts (4.44). This yields

$$v_{i,j} + v_{j,i} = -\frac{\beta_c D_1(0)}{\lambda_{i,0} - \lambda_{j,0}} \left(\frac{dU_{i,0}}{du} U_{j,0} - \frac{dU_{j,0}}{du} U_{i,0} \right)_{u=0}, \qquad i \neq j. \tag{A.1}$$

Moreover, the eigenvalue equation (4.25), together with $\gamma_0(0) = 0$ and $D_0(0) = 0$, implies

$$\lambda_{i,0} U_{i,0}(0) = \left(\frac{dD_0}{du} \frac{dU_{i,0}}{du} \right) \bigg|_{u=0}, \tag{A.2}$$

and by virtue of (4.20) and (4.21), it can be shown that $W^1_\zeta(0) = (dW^0_\zeta/du)_{u=0}$. Thus, from (4.27),

$$D_1(0) = \frac{dD_0}{du}(0), \tag{A.3}$$

and so we can construct the identity

$$(\lambda_{i,0} - \lambda_{j,0})(U_{i,0}U_{j,0})_{u=0} = D_1(0) \left(\frac{dU_{i,0}}{du} U_{j,0} - \frac{dU_{j,0}}{du} U_{i,0} \right)_{u=0}. \tag{A.4}$$

Therefore, from (A.1) and (A.4) we arrive at the result

$$v_{i,j} + v_{j,i} = -\beta_c (U_{i,0}U_{j,0})_{u=0}. \tag{A.5}$$

Finally, using $\beta_c^{-1} = 2\varepsilon$ and applying the composite trapezoidal rule yields a relationship between the discrete and continuous eigenmode products, namely,

$$\begin{aligned}(\mathbf{V}_0^{(i)} + \varepsilon \mathbf{V}_1^{(i)}) \cdot (\mathbf{V}_0^{(j)} + \varepsilon \mathbf{V}_1^{(j)}) &= \frac{1}{2}(U_{i,0}U_{j,0})_{u=0} + \mathcal{O}(\varepsilon^2) \\ &+ \int_0^{\beta_c} (U_{i,0}(\tilde{\nu}/\beta_c) + \varepsilon U_{i,1}(\tilde{\nu}/\beta_c))(U_{j,0}(\tilde{\nu}/\beta_c) + \varepsilon U_{j,1}(\tilde{\nu}/\beta_c))\, d\tilde{\nu}.\end{aligned} \tag{A.6}$$

This reduces to

$$\begin{aligned}(\mathbf{V}_0^{(i)} + \varepsilon \mathbf{V}_1^{(i)}) \cdot (\mathbf{V}_0^{(j)} + \varepsilon \mathbf{V}_1^{(j)}) &= \varepsilon(v_{i,j} + v_{j,i}) \\ &+ \frac{1}{2}(U_{i,0}U_{j,0})_{u=0} + \mathcal{O}(\varepsilon^2) = \mathcal{O}(\varepsilon^2)\end{aligned} \tag{A.7}$$

by virtue of (A.5) and completes the exercise.

Appendix B. Microbending solution. This appendix proves completeness of the regular orthogonal eigenfunctions satisfying

$$\frac{d}{du}\left(u\frac{d\varphi}{du}\right) = \lambda\varphi, \qquad \varphi(1) = 0, \tag{B.1}$$

together with the normalization condition (4.42), $\|\varphi_i\| = \beta_c^{-\frac{1}{2}}$. We find that (see section 4.3) these functions are

$$\varphi_i(u) = \beta_c^{-\frac{1}{2}} \frac{J_0(j_{0,i}\sqrt{u})}{|J_1(j_{0,i})|}, \tag{B.2}$$

where J_n is the Bessel function of the first kind of order n, $j_{0,i}$ is the location of the ith zero of J_0, and each $\varphi_i(u)$ has the associated eigenvalue $\lambda_{i,0} = -j_{0,i}^2/4$.

B.1. Completeness. By construction, the set of eigenfunctions $\{\varphi_i\}_{i=1}^{\infty}$ defines an orthogonal system on $L^2([0,1])$. The system is complete if the Dalzell-type criterion [24] is satisfied, i.e.,

$$S = 2\beta_c \sum_{i=1}^{\infty} \int_0^1 \left| \int_0^t \varphi_i(u)du \right|^2 dt = 1. \tag{B.3}$$

A straightforward calculation yields

$$S = \sum_{i=1}^{\infty} \frac{8}{3j_{0,i}^2}\left(1 + \frac{4}{j_{0,i}^2}\right). \tag{B.4}$$

Now, let $\alpha \geq 2$ be an integer; then the Cauchy residue theorem gives the identity

$$I_{N,\alpha} = \oint_{C_N} \frac{d\ln J_0(z)}{dz}\frac{dz}{z^\alpha} = 4\pi i \sum_{i=1}^{N} \frac{1}{j_{0,i}^\alpha} + 2\pi i \operatorname{Res}\left\{\frac{d\ln J_0(z)}{z^\alpha dz};\, z = 0\right\}, \tag{B.5}$$

where the closed contour C_N is the circle centered at the origin with radius R_N chosen such that $j_{0,N} < R_N < j_{0,N+1}$. Since $\lim_{N\to\infty} I_{N,\alpha} = 0$, we get, setting respectively $\alpha = 2$ and $\alpha = 4$,

$$\sum_{i=1}^{\infty} \frac{1}{j_{0,i}^2} = \frac{1}{4} \quad \text{and} \quad \sum_{i=1}^{\infty} \frac{1}{j_{0,i}^4} = \frac{1}{32}. \tag{B.6}$$

Substitution of these sums into (B.4) completes the result.

B.2. First order eigenvalues. The first order eigenvalues are specified by (4.45); we can evaluate them as follows. First, multiply (B.1) by $d\varphi/du$ and integrate over 0 to 1 to give

$$\int_0^1 \frac{d\varphi}{du}\frac{d}{du}\left(u\frac{d\varphi}{du}\right)du = \frac{\lambda}{2}\int_0^1 \frac{d\varphi^2}{du}\,du. \tag{B.7}$$

Integrating by parts twice then yields

$$\frac{1}{2}\left(\frac{d\varphi}{du}\right)^2_{u=1} + \frac{1}{2}\int_0^1 \left(\frac{d\varphi}{du}\right)^2 du = -\frac{\lambda}{2}\varphi^2(0). \tag{B.8}$$

Integrating by parts again and using the equality $(d\varphi/du)_{u=0} = \lambda\varphi(0)$ gives finally

$$\int_0^1 \varphi\frac{d^2\varphi}{du^2}\,du = \left(\frac{d\varphi}{du}\right)^2_{u=1} = -\frac{\lambda}{\beta_c}. \tag{B.9}$$

Substituting this into (4.45), using (A.3) and setting $\gamma_1 = 0$ (lossless case), yields

$$\lambda_{i,1} = -\lambda_{i,0}. \tag{B.10}$$

Acknowledgments. The authors are most grateful to Professor John Love (ANU, Australia), Dr. David Allwright and Dr. John Ockendon (OCIAM, University of Oxford, UK), and Dr. Dominic Gallagher and Dr. Tom Felici (Photon Design) for their substantial input to the project.

REFERENCES

[1] R. Olshansky, *Propagation in glass optical waveguides*, Rev. Modern Phys., 51 (1979), pp. 341–367.

[2] A. F. Garito, J. Wang, and R. Gao, *Effects of random perturbations in plastic optical fibers*, Science, 281 (1998), pp. 962–967.

[3] S. E. Golowich, W. White, W. A. Reed, and E. Knudsen, *Quantitative estimates of mode coupling and differential modal attenuation in perfluorinated graded-index plastic optical fiber*, J. Lightwave Tech., 21 (2003), pp. 111–121.

[4] J. Garnier, *Light propagation in square law media with random imperfections*, Wave Motion, 31 (2000), pp. 1–19.

[5] D. Marcuse, *Mode conversion caused by surface imperfections of a dielectric slab waveguide*, in Integrated Optics, IEEE Press, New York, 1972.

[6] D. Marcuse, *Radiation losses of dielectric waveguides in terms of the power spectrum of the wall distortion function*, in Integrated Optics, IEEE Press, New York, 1972.

[7] H. E. Rowe and D. T. Young, *Transmission distortion in random multimode waveguides*, IEEE Trans. Microwave Theory, MTT-20 (1972), pp. 349–365.

[8] D. Marcuse, *Derivation of coupled power equations*, Bell System Tech. J., 51 (1972), pp. 229–237.

[9] D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1991.

[10] D. Gloge, *Optical power flow in multimode fibers*, Bell System Tech. J., 51 (1972), pp. 1767–1783.

[11] D. Marcuse, *Losses and impulse response of a parabolic index fiber with random bends*, Bell System Tech. J., 52 (1973), pp. 1423–1427.

[12] R. Olshansky, *Mode coupling effects in graded-index optical fibers*, Appl. Optics, 14 (1975), pp. 935–945.

[13] J. Zubia, G. Durana, G. Adabaldetreku, J. Arrue, M. A. Losada, and M. Lopez-Higuera, *A new method to calculate mode conversion coefficients in SI multimode optical fibers*, J. Lightwave Tech., 16 (1998), pp. 1195–1202.

[14] A. Djordjevitch and S. Savović, *Numerical solution of the power flow equation in step-index plastic optical fibers*, J. Opt. Soc. Amer. B, 21 (2004), pp. 1437–1442.

[15] E. Perrey-Debain and I. D. Abrahams, *A continuous model for mode mixing in graded-index multimode fibres with random imperfections*, Proc. Roy. Soc. A, submitted.

[16] E. Perrey-Debain and I. D. Abrahams, *A diffusion analysis approach for multimode random optical waveguides*, in Proceedings of the 7th International Conference on Mathematical and Numerical Aspects of Waves (WAVES 2005), Brown University, CT, 2005, pp. 267–269.

[17] A. W. Snyder and J. D. Love, *Optical Waveguide Theory*, Chapman and Hall, London, 1983.

[18] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics (Non-Relativistic Theory)*, 3rd ed., Pergamon Press, Oxford, 1977.

[19] M. Skorobogatiy, S. G. Johnson, S. A. Jacobs, and Y. Fink, *Dielectric profile variations in high-index-contrast waveguides, coupled mode theory, and perturbation expansions*, Phys. Rev. E (3), 67 (2003), article 046613.

[20] J. Garnier, *Energy distribution of the quantum harmonic oscillator under random time-dependent perturbation*, Phys. Rev. E (3), 60 (1999), pp. 3676–3687.

[21] L. B. Dozier and F. D. Tappert, *Statistics of normal mode amplitudes in a random ocean, I. Theory*, J. Acoust. Soc. Amer., 63 (1978), pp. 353–365.

[22] E. Perrey-Debain and I. D. Abrahams, *A band factorization technique for transition matrix element asymptotics*, Comput. Phys. Comm., 175 (2006), pp. 315–322.

[23] J. D. Pryce, *Numerical Solution of Sturm-Liouville Problems*, Clarendon Press, Oxford, 1993.

[24] J. R. Higgins, *Completeness and Basis Properties of Sets of Special Functions*, Cambridge University Press, Cambridge, UK, 2004.

SIAM J. APPL. MATH.
Vol. 68, No. 2, pp. 544–561

THE GENERALIZED RIEMANN PROBLEM FOR A SCALAR NONCONVEX CHAPMAN–JOUGUET COMBUSTION MODEL*

WANCHENG SHENG[†], MEINA SUN[†], AND TONG ZHANG[‡]

Abstract. The generalized Riemann problem for a scalar nonconvex Chapman–Jouget combustion model in a neighborhood of the origin ($t > 0$) on the (x, t) plane is considered. Under the entropy conditions, we exhibit the construction of the solutions. It can be observed that, for some cases, there are essential differences between the structures of the perturbed Riemann solutions and the corresponding Riemann solutions. Especially, a strong detonation in the corresponding Riemann solution may turn into a weak deflagration followed by a shock wave after perturbation, which appears in the numerical simulations of Bao and Jin [*J. Comput. Phys.*, 163 (2000), pp. 216–248] and Zhang and Ying [*J. Comput. Math.*, 23 (2005), pp. 337–350].

Key words. scalar nonconvex CJ combustion model, generalized Riemann problem, detonation, deflagration

AMS subject classifications. 35L45, 35L60, 35L65, 35L67, 58J45, 76L05, 76N10, 80A25

DOI. 10.1137/060672650

1. Introduction.

The Chapman–Jouguet (CJ) combustion theory plays an important role in gas dynamics [5], [16], [2]. In Lagrangian coordinates, one-dimensional adiabatic, inviscid flow of combustible ideal gases with an infinite rate of reaction is described by the hyperbolic system of conservation laws

$$\begin{cases} u_t + p_x = 0, \\ \tau_t - u_x = 0, \\ E_t + (pu)_x = 0, \end{cases} \tag{1.1}$$

with reaction equation

$$q(x,t) = \begin{cases} q(x,0), & \sup\limits_{0<y\le t} T(x,y) \le T_i, \\ 0 & \text{otherwise}, \end{cases} \tag{1.2}$$

where $u, p > 0$, $\tau > 0$, T, and $E > 0$ are the velocity, pressure, specific volume, temperature, and specific energy, respectively, and T_i is the ignition temperature. In addition to the kinetic energy $\frac{1}{2}u^2$ and internal energy e, the total energy E also contains the chemical binding energy q:

$$E = \frac{1}{2}u^2 + e + q. \tag{1.3}$$

The internal energy $e = e(T)$ is a known function of T, which satisfies the law of Boyle and Gay-Lussac for ideal gas [5],

$$p\tau = RT, \tag{1.4}$$

*Received by the editors October 18, 2006; accepted for publication (in revised form) July 10, 2007; published electronically December 5, 2007. This work was supported by NSFC 10671120 and the Key Program from Beijing Educational Commission with KZ 200510028018.

http://www.siam.org/journals/siap/68-2/67265.html

[†] Department of Mathematics, Shanghai University, Shanghai, 200444, People's Republic of China (mawcsheng@staff.shu.edu.cn, smnwhy0350@163.com).

[‡] Institute of Mathematics, AMSS, Chinese Academy of Sciences, Beijing, 100080, People's Republic of China (tzhang@math.ac.cn).

or the van der Waals equation of state for van der Waals gas [11],

$$\left(p+\frac{a}{\tau^2}\right)(\tau-b)=RT, \tag{1.5}$$

where R, a, and b are constants. For ideal gas, the equation of state (1.4) is a convex curve in the (p,τ) plane, while for van der Waals gas, the equation of state (1.4) is a nonconvex curve in the (p,τ) plane in some cases.

Because of the difficulty of the combustion problems in gas dynamics, there are few results except for the Riemann problems in [4] (solution involving only detonation), [15] (solution involving only deflagration), and [21]. In [21], all Riemann solutions are obtained for the Riemann problem (1.1), (1.2) and

$$(u,\tau,p,q)\big|_{t=0}=(u^{\pm},\tau^{\pm},p^{\pm},q^{\pm}),\quad \pm x>0. \tag{1.6}$$

Fickett in 1979 [6] and Majda in 1981 [10] proposed the simplest CJ combustion model in Lagrangian coordinates,

$$\begin{cases} (u+q)_t+f(u)_x=0, \\ q(x,t)=\begin{cases} q(x,0), & \sup\limits_{0\le\tau\le t} u(x,\tau)\le u_i, \\ 0 & \text{otherwise}, \end{cases} \end{cases} \tag{1.7}$$

where u is a lumped quantity with some features of density, velocity, pressure or temperature, q denotes the binding energy of the reactive gas, and $f(u)$ represents the flux function. The model (1.7) describes the combustible gas with an infinite rate of reaction, which implies that a gas particle releases all of its binding energy once its temperature exceeds u_i (ignition temperature).

In this paper, the initial value we are interested in is of the form

$$(u,q)(x,0)=(u_0^{\pm}(x),q_0^{\pm}(x)),\quad \pm x>0, \tag{1.8}$$

where $q_0^{\pm}(x)$ equal a constant q_0 for unburnt gas and zero for burnt gas, and $u_0^{\pm}(x)$ are arbitrary smooth functions with the properties

$$\lim_{x\to 0-} u_0^{-}(x)=u^{-},\quad \lim_{x\to 0+} u_0^{+}(x)=u^{+}. \tag{1.9}$$

The corresponding Riemann problem is an initial problem (1.7) with

$$(u,q)\big|_{t=0}=\begin{cases} (u^{-},q_0), & x<0, \\ (u^{+},0), & x>0. \end{cases} \tag{1.10}$$

The initial value problem (1.8) is a perturbation of (1.10) at the neighborhood of the origin. So, we call (1.8) a generalized Riemann problem. Naturally, we would like to know whether or not, in some neighborhood of the origin, the solution for (1.7) and (1.8) is similar to the corresponding Riemann solution. When $f(u)$ is convex, this problem was studied in [14], the results of which show that they are essentially different for some cases. In the present paper, our attention is focused on the nonconvex case. For simplicity, we assume that $f(u)$ is the simplest nonconvex function; i.e.,

(A) $f(u)$ has only one inflection point $\tilde{u}$, and $f'(\pm\infty)=+\infty$.

The case for $f(u)$ with one inflection point and $f'(\pm\infty) = -\infty$ can be treated similarly without substantial difficulties.

To guarantee the uniqueness of the solution, entropy conditions are needed. Here we mention some works on the entropy conditions for the model (1.7). In 1984, Ying and Teng [17] proved the existence and uniqueness of the Riemann solution for the Zeldovich–von Neumann–Döring (ZND) model

$$\begin{cases} (u+q)_t + f(u)_x = 0, \\ q_t = -k\varphi(u)q, \end{cases} \tag{1.11}$$

where k is the rate of reaction for combustible gas and $\varphi(u)$ is the Heaviside function: $\varphi(u) = 0$ as $u \le u_i$, $\varphi(u) = 1$ as $u > u_i$. Furthermore, they obtained limits of the solution as k tends to infinity and defined the limits as an admissible solution of the Riemann problem for (1.7). Based on Ying and Teng's results, Liu and Zhang [9] summarized a set of entropy conditions, including pointwise and global entropy conditions, with which they obtained the uniqueness of the Riemann problem for CJ model (1.7). These results were all obtained under the assumption that $f(u)$ is strictly convex.

Since a genuine two-dimensional conservation law must be nonconvex in certain directions [3], [20] and the van der Waals equation of state is nonconvex in some cases, it is interesting to investigate a scalar combustion model with a nonconvex flux $f(u)$, which is the indispensable preparation for the study of multidimensional combustion problems. There is another motivation to study the nonconvex model (1.7) [12]. A well-known phenomenon in combustion theory is the transition from deflagration to detonation. However, this phenomenon cannot occur in the convex case because detonation and deflagration waves cannot propagate in the same direction (forward or backward). In the nonconvex case, however, this phenomenon can be observed [12].

For the nonconvex system (1.7), Zhang and Zhang [19] gave the entropy restriction that mimics those in [9] and generalizes the classical Oleinik entropy condition for scalar conservation laws when solving the Riemann problem. In 2003, Li and Zhang [8] proved that the Riemann solutions in [19] are the limit of the Riemann solutions for the nonconvex self-similar ZND combustion model

$$\begin{cases} (u+q)_t + f(u)_x = 0, \\ q_t = -\dfrac{k}{t}\varphi(u)q \end{cases} \tag{1.12}$$

as the rate of reaction goes to infinity. However, through the study of the structure stability of combustion solutions, Sheng and Zhang [12] found that their entropy conditions do not guarantee the uniqueness in some cases, which was not discussed in the two papers [19] and [8]. Sheng and Zhang contributed a set of complete entropy conditions by a different method and uniquely constructed the entropy solutions for the Riemann problem (1.7), (1.8), and (1.9). The ignition problem for (1.7) without convexity was investigated in [13].

With the method of characteristic analysis, we constructively obtain the solutions of (1.7) and (1.8), which include all the possibilities in [14] and have more interesting structures. We find that for most of the cases, the combustion waves in the corresponding Riemann solutions are able to retain their forms after perturbation, in the neighborhood of the origin, while for some other cases, the perturbation brings essential changes to the combustion waves. For instance, the perturbation may transform a strong detonation into a weak deflagration followed by a shock wave; see Case 4 in the present paper. This interesting phenomenon also appears in the numerical solutions in [1] and [18]. The phenomenon is unreadable and has puzzled numerical analysts. Our theoretical results give a reasonable explanation for this phenomenon. In fact, error is unavoidable in computation. The error forms a perturbation of the initial data.

This paper is organized as follows. In section 2, we present some preliminaries containing the pointwise and global entropy conditions and elementary waves. Then the construction of the perturbed Riemann solutions and our main results are exhibited in section 3.

2. Preliminaries. Suppose the piecewise smooth vector function (u, q) satisfies (1.7). Then it is easy to show that $q(x, t)$ is piecewise constant, 0 or q_0, and that the smooth solution $u(x, t)$ is a constant or rarefaction wave (R).

A jump of solution $(u, q)(x, t)$ at $x = x(t)$ should satisfy the Rankine–Hugoniot condition

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \frac{[f]}{[u+q]} =: \sigma, \tag{2.1}$$

where $[f] = f(u_r) - f(u_l)$, $u_l = u(x(t) - 0, t)$, $u_r = u(x(t) + 0, t)$, etc.

The following three kinds of noncombustion discontinuities are admissible.

1. $[q] = 0$, $[u] \neq 0 \Rightarrow \sigma = \frac{[f]}{[u]}$ is a generalized shock, which may be classified into
   (a) shock wave (S): $f'(u_r) < \sigma < f'(u_l)$;
   (b) left–contact discontinuity (LC): $f'(u_r) < \sigma = f'(u_l)$;
   (c) right–contact discontinuity (RC): $f'(u_r) = \sigma < f'(u_l)$;
   (d) double–contact discontinuity (DC): $f'(u_r) = \sigma = f'(u_l)$;
2. $[q] \neq 0$, $[u] = 0 \Rightarrow \sigma = 0$ is a contact jump (J);
3. $[q] \neq 0$, $[u] \neq 0$, $\sigma = 0$ is a combination of S and J (SJ),

where the generalized shock and SJ satisfy the Oleinik entropy condition

$$\frac{f(u) - f(u_l)}{u - u_l} \geq \frac{f(u_r) - f(u_l)}{u_r - u_l} \qquad \text{for} \quad (u - u_l)(u - u_r) \leq 0. \tag{2.2}$$

We next investigate the combustion wave, which has nonzero speed $\sigma \neq 0$, and across which q jumps from q_0 to zero. Let u_l and u_r be the limit values of u in the combustion wave front and wave back, respectively, i.e., $q_l > 0 = q_r$ and $u_l \leq u_i < u_r$, which implies $\sigma = \frac{f(u_r) - f(u_l)}{u_r - (u_l + q_0)} < 0$. Then the following six kinds of combustion waves satisfying the pointwise entropy conditions [12] are admissible.

Pointwise entropy conditions.

a. If there exists a $u_R \in [u_l, u_r)$ such that for all $u \in (u_l, u_r)$,

$$\sigma = \frac{f(u_r) - f(u_l)}{u_r - (u_l + q_0)} = \frac{f(u_R) - f(u_l)}{u_R - u_l} \leq \frac{f(u) - f(u_l)}{u - u_l}, \tag{2.3}$$

the discontinuity σ is called deflagration. Furthermore, it can be divided into three subcases:

1. $f'(u_l) = \sigma < f'(u_r)$: CJ deflagration (CJDF);
2. $f'(u_l) > \sigma < f'(u_r)$: weak deflagration (WDF);
3. $f'(u_l) = \sigma = f'(u_r)$: double–contact combustion (DCC).

b. If there exists a $u_R \in [u_r, +\infty)$ satisfying (2.3) for $u \in (u_l, u_R)$, σ is called detonation. Also, it can be divided into three subcases:

4. $f'(u_l) > \sigma = f'(u_r)$: CJ detonation (CJDT);

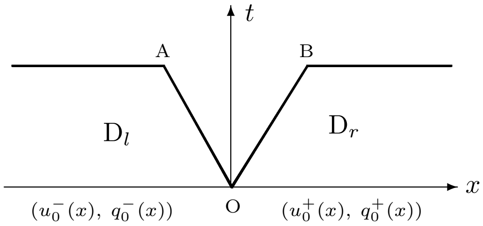


Fig. 1.

5. $f'(u_l) > \sigma > f'(u_r)$: strong detonation (SDT);
6. $f'(u_l) = \sigma > f'(u_r)$: contact detonation (CDT).

For the case that $q_l = 0 < q_r$, $u_l > u_i \geq u_r$, the pointwise entropy conditions for combustion waves can be easily defined by means of transformation $\bar{x} = -x$, $\bar{f} = -f$.

We call R, S, J, SJ, CJDF, WDF, DCC, CJDT, SDT, and CDT elementary waves for (1.7) without convexity.

The aforementioned entropy conditions cannot guarantee the uniqueness and structure stability of the Riemann solutions for (1.7). Hence, global entropy conditions [12] are needed.

Global entropy condition.

If the Riemann problem for (1.7) has several solutions, we choose one which satisfies the following rule:

If $U = \big\{u \;\big|\; \exists u < u_i,$ *such that* $f(u) = f(u + q_0)$ *and* $f'(u + q_0)f'(u_i) > 0\big\} \neq \emptyset$, *take*

$$u_l = \max\big\{u \;\big|\; u \in U\big\};$$

otherwise, take the combustion wave which propagation speed is as low as possible.

Then in the next section, we show that the generalized Riemann problem subject to the above entropy conditions can be solved uniquely.

3. Solutions of the generalized Riemann problem. We will investigate the solutions for the discontinuous initial value problem (1.7), (1.8) in a neighborhood of the origin ($t > 0$) on the (x,t) plane. In fact, in the region where the solution is smooth ($q \equiv$ *const.*), (1.7) reduces to the scalar conservation law $u_t + f(u)_x = 0$, for which the generalized Riemann problem was studied in [3] and the references therein. Hence by [7], the classical solution $(u_l, q_l)(x,t)$ $((u_r, q_r)(x,t))$ can be defined in a strip domain D_l (D_r) for local time. The right boundary of D_l has characteristic OA: $x = \lambda(u^-)t$, and the left boundary of D_r has characteristic OB: $x = \lambda(u^+)t$, where $\lambda(u) = f'(u)$ (see Figure 1).

We will distinguish the different cases according to the different solutions for the corresponding Riemann problem [12]. It is redundant to dwell on all the cases since some of them can be discussed similarly. So we will pick some typical cases and focus our attention on them in the following.

By assumption (A), there are two possibilities: $f'(\tilde{u}) < 0$ or $f'(\tilde{u}) \geq 0$. The latter can be treated as the special case of the former. Therefore, we suppose $f'(\tilde{u}) < 0$ in the following without loss of generality. From $f'(\tilde{u}) < 0$ and $f'(\pm\infty) = +\infty$, we know that there exist u_1 and u_2 such that $f'(u_1) = f'(u_2) = 0$, where $u_1 > \tilde{u} > u_2$. Let u_3, u_4 satisfy $f(u_1) = f(u_3)$, $f(u_2) = f(u_4)$, respectively (Figure 2). For convenience, in the following figures, we denote $(f(u^{\pm}), u^{\pm})$ as $(\pm)$, $(f(u^{\pm}), u^{\pm}{+}q_0)$ as $(\pm')$, $(f(u_i), u_i)$ as (i) in the (f,u) plane, etc.

3.1. Solutions for $u_0^-(x) \leq u_i < u^+$, $q_0^-(x) = q_0 > 0 = q_0^+(x)$. We begin by fixing $u^- \in (u_3, \tilde{u})$. Let u_5 ($\tilde{u} < u_5 < u_1$) satisfy $f'(u_5) = w(u^-, u_5)$, where $w(u,v) = \frac{f(u)-f(v)}{u-v}$. Then our discussion can be divided into the following cases according to the elementary waves in the corresponding Riemann solutions.

Case 1. *The corresponding Riemann problem has noncombustion solution containing no SJ.*

We just consider the case $u_i > u_5$ and $f'(u_i) \geq 0$, for which the Riemann solution consists of a contact jump J for q and RC + R for u. Here "+" means "follows."

When $\dot{u}_0^-(0) > 0$, it is possible to construct a discontinuity without combustion $x = x(t)$ in the domain $\lambda(u_5)t < x < \lambda(u^+)t$ as follows:

$$\left\{ \begin{array}{l} \dfrac{\mathrm{d}x}{\mathrm{d}t} = w(u_l(x,t), u_r(x,t)), \\ x(0) = 0, \end{array} \right. \tag{3.1}$$

where $u_r(x,t)$ is a centered simple wave defined by

$$\frac{x}{t} = \lambda(u_r(x,t)) \quad (u_5 \leq u_r \leq u^+). \tag{3.2}$$

Since $x = x(t)$ satisfies the stability condition at the origin, $\lambda(u_5) = \dot{x}(0) < \lambda(u^-)$, it can be proved by the method used in [3, p. 16] that there exists a solution of (3.1) in the interior of domain $\lambda(u_5)t < x < \lambda(u^+)t$, near the origin.

Conversely, under the condition that there is a discontinuity $x = x(t)$ with $\dot{x}(0) = \lambda(u_5)$ and $\ddot{x}(0) \geq 0$ in the domain $\lambda(u_5)t < x < \lambda(u^+)t$, we now prove that $\dot{u}_0^-(0) \geq 0$. In fact, $\ddot{x}(0)$ can be calculated as follows. Differentiating the first equation in (3.1) with respect to t and letting $t = 0$, one obtains

$$(u^- - u_5)\ddot{x}(0) = (\lambda(u^-) - \dot{x}(0)) \left.\frac{\mathrm{d}u_l}{\mathrm{d}t}\right|_{t=0} - (\lambda(u_5) - \dot{x}(0)) \left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0},$$

where it can be easily checked that

$$\left.\frac{\mathrm{d}u_l}{\mathrm{d}t}\right|_{t=0} = \frac{\partial u_l}{\partial t} + \frac{\partial u_l}{\partial x}\dot{x}(0) = (\lambda(u_5) - \lambda(u^-))\dot{u}_0^-(0).$$

As for $\left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0}$, note that along $x = x(t)$,

$$\lim_{t\to 0} \lambda(u_r(x,t)) = \lim_{t\to 0} \frac{x}{t} = \dot{x}(0) = \lambda(u_5);$$

namely,

$$\lim_{t\to 0} u_r(x(t),t) = u_5.$$

Then from (3.2), we have

$$\lambda'(u_5) \left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0} = \lim_{t\to 0} \frac{t\dot{x}(t) - x(t)}{t^2} = \frac{\ddot{x}(0)}{2}.$$

Thus

$$(u^- - u_5)\ddot{x}(0) = -(\lambda(u^-) - \lambda(u_5))^2 \dot{u}_0^-(0),$$

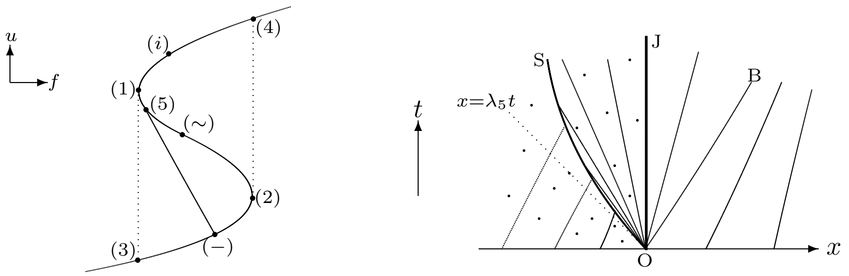


Fig. 2.

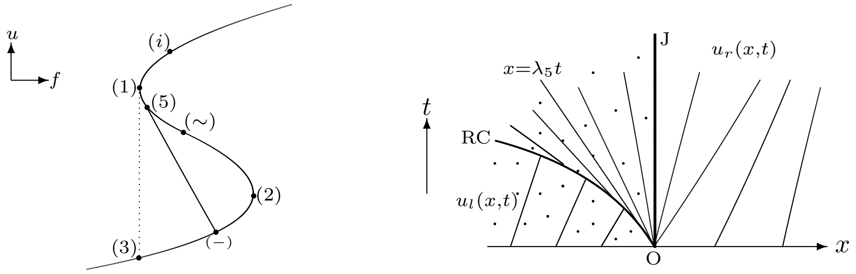


Fig. 3.

which shows that $\ddot{x}(0)$ has the same sign as $\dot{u}_0^-(0)$. It is easily seen that $x = x(t)$ satisfies $\lambda(u_r) < w(u_l, u_r) < \lambda(u_l)$ near the origin, which means that the right–contact discontinuity turns into a shock wave (see Figure 2).

For $\dot{u}_0^-(0) < 0$, it is impossible to construct a solution containing a shock wave, but it is possible to construct an RC $x = x(t)$ such as

$$\left\{\begin{array}{l} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \lambda(\bar{u}) = w(u_l, \bar{u}) \quad \left(-\infty < \dfrac{x}{t} \leq \lambda(u_5)\right), \\ x(0) = 0. \end{array}\right. \tag{3.3}$$

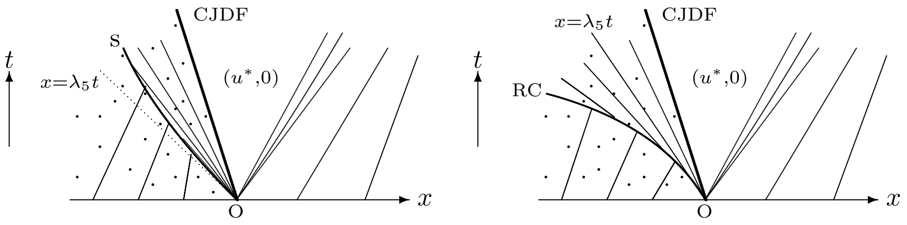
A solution of (1.7) and (1.8) can then be defined, which takes $u_r(x,t)$ on the right-hand side of $x = \lambda(u_5)t$ and $u_l(x,t)$ on the left-hand side of $x = x(t)$ for u, respectively. The characteristic of the solution in the domain $x(t) < x \leq \lambda(u_5)t$ is the tangent of $x = x(t)$ (see Figure 3). Similarly to the case $\dot{u}_0^-(0) > 0$, it can be proved that $\dot{u}_0^-(0) < 0$ is a necessary condition for a solution constructed as above.

Case 2. *A combustion wave CJDF appears in the corresponding Riemann solution:* $u_i > u_5$ *and* $f'(u_i) < 0$.

Since the same discussion as above can be carried out for this case, we omit the details. The solution for $u^+ \geq u^*$ is illustrated in Figure 4, where $u^* > u_i$ satisfies $f'(u_i) = \frac{f(u^*) - f(u_i)}{f(u^*) - (u_i + q_0)}$. The case for $u^+ < u^*$ is the same as for $u^+ \geq u^*$ except that $(u^*, 0)$ and $(u_r, 0)$ are connected by a shock wave instead of a centered simple wave, in which the shock wave satisfies

$$\left\{\begin{array}{l} \dfrac{\mathrm{d}x}{\mathrm{d}t} = w(u^*, u_r(x,t)) \quad \left(\lambda(u^+) \leq \dfrac{x}{t} \leq \lambda(u^*)\right), \\ x(0) = 0. \end{array}\right. \tag{3.4}$$

As we can see, the perturbation has no influence on the CJDF, which goes on to propagate with the speed $f'(u_i)$ in the neighborhood of the origin.



$\dot{u}_0^-(0){>}0$ $\dot{u}_0^-(0){<}0$

FIG. 4.

Case 3. *A combustion wave WDF appears in the corresponding Riemann solution.*

We are concerned with the typical case $u_i = u_5$.

In the case of $\dot{u}_0^-(0) < 0$, it can be proved that the combustion wave WDF turns to CJDF $x = f'(u_i)t$, ahead of which is RC+R lying in the domain $-\infty < \frac{x}{t} \le \lambda(u_i)$. The structure of the solution is similar to that of the case $\dot{u}_0^-(0) < 0$ in Figure 4.

In the case of $\dot{u}_0^-(0) > 0$, the combustion wave WDF remains. Let u^* have the same representation as in Case 2. Then we need to do some analysis for the case $u^+ \ge u^*$, for which the corresponding Riemann solution can be denoted by WDF + $(u^*, 0)$ + R. For any u_l, we may define $\bar{u}(u_l) \in (\tilde{u}, u_5)$, $\hat{u}(u_l) > \bar{u}(u_l)$ such that

$$\begin{cases} w(u_l, \bar{u}) = \lambda(\bar{u}) = \dfrac{f(\hat{u}) - f(u_l)}{\hat{u} - (u_l + q_0)}, \\ \bar{u}(u^- - 0) = u_5, \quad \hat{u}(u^- - 0) = u^*. \end{cases} \tag{3.5}$$

Differentiating the above equations and eliminating $\mathrm{d}\bar{u}$, we finally get, by setting $u_l = u^- - 0$,

$$\left.\frac{\mathrm{d}\hat{u}}{\mathrm{d}u_l}\right|_{u_l = u^- - 0} = \frac{(\lambda(u^-) - \lambda(u_5))(u_5 + q_0 - u^*)}{(\lambda(u^*) - \lambda(u_5))(u_5 - u^-)}. \tag{3.6}$$

Along the WDF $x = x(t)$ with $\dot{x}(0) = \lambda(u_5)$, defined as

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \lambda(\bar{u}) = \dfrac{f(\hat{u}) - f(u_l)}{\hat{u} - (u_l + q_0)}, \\ x(0) = 0, \end{cases} \tag{3.7}$$

it holds that

$$\left.\frac{\mathrm{d}u_l}{\mathrm{d}t}\right|_{t=0} = (\lambda(u_5) - \lambda(u^-))\dot{u}_0^-(0). \tag{3.8}$$

Therefore,

$$\left.\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\right|_{t=0} = \left.\frac{\mathrm{d}\hat{u}}{\mathrm{d}u_l}\right|_{u_l = u^- - 0} \left.\frac{\mathrm{d}u_l}{\mathrm{d}t}\right|_{t=0} = \frac{(\lambda(u_5) - \lambda(u^-))^2(u_5 + q_0 - u^*)}{(\lambda(u_5) - \lambda(u^*))(u_5 - u^-)}\dot{u}_0^-(0). \tag{3.9}$$

Then for $\ddot{x}(0)$ of the WDF, an easy calculation provides

$$(u^- + q_0 - u^*)\ddot{x}(0) = (\lambda(u^-) - \lambda(u_5)) \left.\frac{\mathrm{d}u_l}{\mathrm{d}t}\right|_{t=0} - (\lambda(u^*) - \lambda(u_5)) \left.\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\right|_{t=0},$$

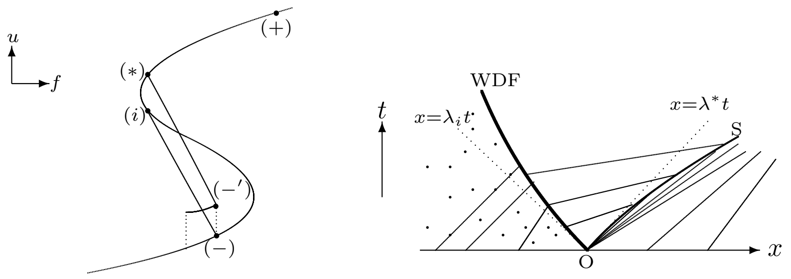


FIG. 5.

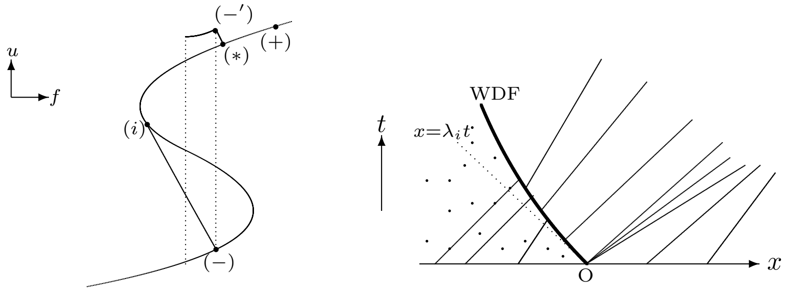


FIG. 6.

which, combined with (3.8) and (3.9), implies

$$\ddot{x}(0) = \frac{(\lambda(u^-) - \lambda(u_5))^2}{u_5 - u^-} \dot{u}_0^-(0) > 0.$$

It is obvious that $\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\Big|_{t=0} > 0$ when $u^* > u_5 + q_0$, which means that the rarefaction wave behind the WDF turns to a shock wave $x = x(t)$ (see Figure 5) determined by

$$(3.10)\qquad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = w(\hat{u}, u_r), & \lambda(u^*) \le \dfrac{x}{t} \le \lambda(u^+), \\ \dfrac{x}{t} = \lambda(u_r), & u^* \le u_r \le u^+, \\ x(0) = 0. \end{cases}$$

For $u^* < u_5 + q_0$, i.e., $\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\Big|_{t=0} < 0$, however, the rarefaction wave remains (see Figure 6). For the case $u^+ < u^*$, we can show by similar discussion that the solution is similar to the corresponding Riemann solution consisting of WDF $+ (u^*, 0) +$ S. We omit the details. In the following, we take $u_i \in (u^-, u_5)$ and $u^+ \in (u_i, u_5)$.

Case 4. *A combustion wave SDT appears in the corresponding Riemann solution.*

We discuss the case $f(u^+) < f(u^-)$, $f'(u^+) < f'(u_5)$, and $q_0 \in (0, q_1]$, where $q_1 > 0$ satisfies $f'(u_5) = \frac{f(u^+) - f(u^-)}{u^+ - (u^- + q_1)}$ (see Figure 7).

It is easy to find that the SDT retains its form when $q_0 \in (0, q_1)$. Namely, a strong detonation $x = x(t)$ can be constructed in the interior of the domain AOB satisfying

$$(3.11)\qquad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \dfrac{f(u_r) - f(u_l)}{u_r - (u_l + q_0)} & \left(\lambda(u^+) \le \dfrac{x}{t} \le \lambda(u^-)\right), \\ x(0) = 0. \end{cases}$$

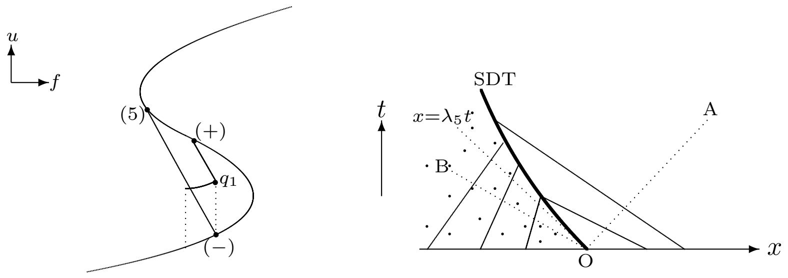


FIG. 7.

Then we consider the complicated case $q_0 = q_1$. According to the Riemann solution, it is possible to define a combustion jump $x = x(t)$ by solving the problem (3.11). Whether an $x = x(t)$ so defined is an SDT depends on whether the stability condition $\lambda(u_r) < \frac{\mathrm{d}x}{\mathrm{d}t} < \lambda(u_l)$ holds for it.

Let $\bar{u}(u_l)$, $\hat{u}(u_l)$ have the same meanings as in Case 3, while $u^* > u_5$ satisfies $f'(u_5) = \frac{f(u^*)-f(u^-)}{f(u^*)-(u^-+q_1)}$, which is different from that in (3.5). In addition, we define $\hat{\hat{u}}(u_l) < \bar{u}(u_l)$ as

$$\begin{cases} \lambda(\bar{u}) = \dfrac{f(\hat{\hat{u}}) - f(u_l)}{\hat{\hat{u}} - (u_l + q_1)}, \\ \hat{\hat{u}}(u^- - 0) = u^+. \end{cases} \tag{3.12}$$

By a calculation similar to that in Case 3, it can be obtained that along $x = x(t)$,

$$\left.\frac{\mathrm{d}\hat{\hat{u}}}{\mathrm{d}t}\right|_{t=0} = -\frac{(\lambda(u^-) - \lambda(u_5))^2(u_5 + q_1 - u^+)}{(\lambda(u^+) - \lambda(u_5))(u_5 - u^-)}\dot{u}_0^-(0).$$

It is easy to see that $x = x(t)$ is an SDT if and only if

$$\left.\frac{\mathrm{d}\hat{\hat{u}}}{\mathrm{d}t}\right|_{t=0} \geq \left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0},$$

which, combined with

$$\left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0} = (\lambda(u_5) - \lambda(u^+))\dot{u}_0^+(0),$$

implies that

$$\frac{(\lambda(u^-) - \lambda(u_5))^2}{u_5 - u^-}\dot{u}_0^-(0) \geq \frac{(\lambda(u^+) - \lambda(u_5))^2}{u_5 + q_1 - u^+}\dot{u}_0^+(0). \tag{3.13}$$

Obviously, (3.13) is satisfied when $\dot{u}_0^-(0) > 0$ and $\dot{u}_0^+(0) < 0$. Moreover, we can derive $\ddot{x}(0) > 0$ for the SDT by the following:

$$(u^- + q_1 - u^+)\ddot{x}(0) = (\lambda(u^+) - \lambda(u_5))^2\dot{u}_0^+(0) - (\lambda(u^-) - \lambda(u_5))^2\dot{u}_0^-(0) > 0.$$

The solution is depicted in Figure 7.

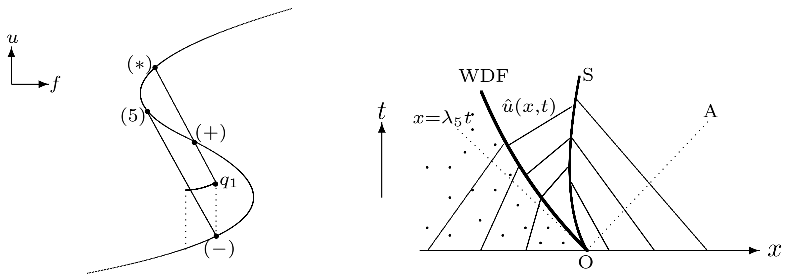


Fig. 8.

Next, let us assume $\dot{u}_0^-(0) > 0$ and $\dot{u}_0^+(0) > 0$. It is certain that the SDT remains when (3.13) is satisfied. Once (3.13) fails to hold, it turns out that the combustion wave SDT turns into a WDF followed by a shock wave (see Figure 8). Now we give the proof for this case.

Due to the entropy conditions, a WDF $x = x^+(t)$ should be constructed in the domain $\lambda(u_5)t \leq x \leq \lambda(u^-)t$, which is determined by (3.7). Then we construct the jump without combustion on the right-hand side of $x = x^+(t)$ as follows:

$$\begin{cases} \dot{x}^-(t) = w(\hat{u}(x,t), u_r(x,t)) \quad \left(x^+(t) \leq \dfrac{x}{t} \leq \lambda(u^-)\right), \\ x^-(0) = 0. \end{cases} \tag{3.14}$$

In order to show the existence of the shock $x = x^-(t)$, we need a certain a priori estimate.

In fact, by (3.7) and (3.14), it can be shown that for any t, t_0 we have

$$\frac{x^-(t) - x^+(t_0)}{t - t_0} = \lambda(\hat{u}(x^-(t),t)) = \lambda(\hat{u}(x^+(t_0),t_0)), \quad t > t_0, \tag{3.15}$$

where

$$\hat{u}(x^-(t),t) = \hat{u}(x^+(t_0),t_0).$$

Differentiating the above equation yields

$$\left.\frac{\mathrm{d}\hat{u}(x^-(t),t)}{\mathrm{d}t}\right|_{t=0} = \left.\frac{\mathrm{d}\hat{u}(x^+(t_0),t_0)}{\mathrm{d}t_0}\right|_{t_0=0} \left.\frac{\mathrm{d}t_0}{\mathrm{d}t}\right|_{t=0}.$$

Noting

$$\left.\frac{\mathrm{d}\hat{u}(x^+(t_0),t_0)}{\mathrm{d}t_0}\right|_{t_0=0} = \frac{(\lambda(u_5) - \lambda(u^-))^2(u_5 + q_1 - u^*)}{(u_5 - u^-)(\lambda(u_5) - \lambda(u^*))}\dot{u}_0^-(0),$$

it follows from (3.15) that $\frac{\mathrm{d}t_0}{\mathrm{d}t}\big|_{t=0} = 1$. Thus we have

$$(u^* - u^+)\ddot{x}^-(0) = (\lambda(u^+) - \lambda(u_5))^2\dot{u}_0^+(0) - \frac{(\lambda(u^-) - \lambda(u_5))^2(u_5 + q_1 - u^*)}{u_5 - u^-}\dot{u}_0^-(0).$$

On the other hand, we get $\ddot{x}^+(0) = \frac{(\lambda(u^-)-\lambda(u_5))^2}{u_5-u^-}\dot{u}_0^-(0)$. Therefore,

$$\ddot{x}^-(0) - \ddot{x}^+(0) = \frac{(\lambda(u^+) - \lambda(u_5))^2}{u^* - u^+}\dot{u}_0^+(0) - \frac{(u_5 + q_1 - u^+)(\lambda(u^-) - \lambda(u_5))^2}{(u_5 - u^-)(u^* - u^+)}\dot{u}_0^-(0) > 0.$$

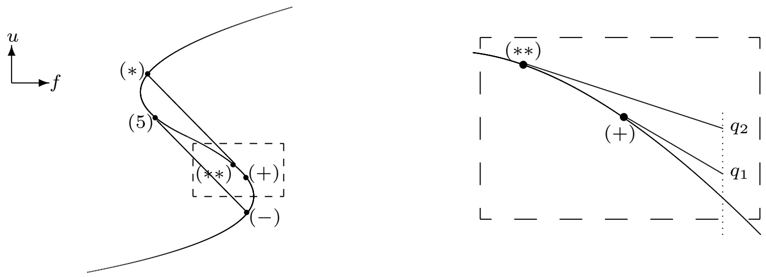


FIG. 9.

Then by virtue of $x^-(0) = x^+(0) = 0$, $\dot{x}^-(0) = \dot{x}^+(0) = \lambda(u_5)$, $\ddot{x}^-(0) > \ddot{x}^+(0) > 0$, it can be proved that there exists a solution $x = x^-(t)$ of (3.14) which takes the value $\hat{u}(x,t)$ on the left-hand side and the value $u_r(x,t)$ on the right-hand side, respectively. Thus we have completed the construction of the solution for this case.

For the remaining cases, $\dot{u}_0^-(0) < 0$ and $\dot{u}_0^+(0) > 0$ or $\dot{u}_0^-(0) < 0$ and $\dot{u}_0^+(0) < 0$, the same discussion as above can be carried out. In brief, the solution of (1.7) and (1.8) may contain an SDT if (3.13) holds; otherwise a WDF followed by a shock may appear.

Remark. In the numerical simulations [18], transition from SDT to WDF followed by a shock is reasonable from the above discussion.

Case 5. *A combustion wave CJDT appears in the corresponding Riemann solution.*

The case $f(u^+) < f(u^-)$, $f'(u^+) > f'(u_5)$, and $q_0 \in [q_1, q_2]$ is considered. Here $q_1, q_2 > 0$ satisfy $f'(u^+) = \frac{f(u^+)-f(u^-)}{u^+-(u^-+q_1)}$, $f'(u_5) = f'(u^{**}) = \frac{f(u^{**})-f(u^-)}{u^{**}-(u^-+q_2)}$, respectively, in which $u^{**} \in (u^+, \tilde{u})$. It is obvious that $q_2 > q_1$ (see Figure 9).

For $q_0 \in [q_1, q_2)$, a discussion similar to that for the convex case shows that the CJDT in the corresponding Riemann solution may either turn into an SDT (if $\dot{u}_0^-(0) > 0$) or retain its form (if $\dot{u}_0^-(0) < 0$).

We now investigate the complex case $q_0 = q_2$. First, we assume $\dot{u}_0^-(0) < 0$. Motivated by the Riemann solution, a CJDT $x = x(t)$ is considered to be constructed in the domain $-\infty < \frac{x}{t} \le \lambda(u_5)$, which is determined by

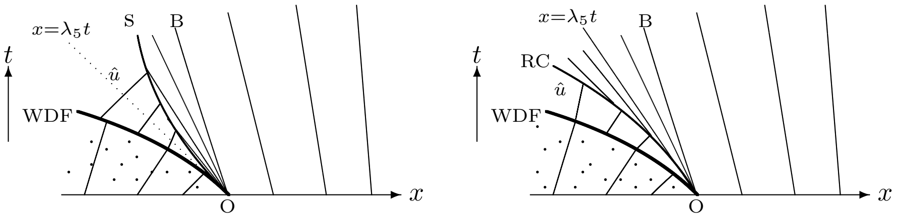
$$\text{(3.16)} \qquad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \lambda(\bar{\bar{u}}) = \dfrac{f(\bar{\bar{u}}) - f(u_l)}{\bar{\bar{u}} - (u_l + q_2)}, \\ x(0) = 0, \end{cases}$$

where $\bar{\bar{u}} \in (u^{**}, \tilde{u})$ and $\bar{\bar{u}}(u^- - 0) = u^{**}$. Obviously, such a CJDT occurs if and only if along $x = x(t)$,

$$\text{(3.17)} \qquad \lambda'(u^{**}) \frac{\mathrm{d}\bar{\bar{u}}}{\mathrm{d}t}\bigg|_{t=0} \ge \lambda'(u_5) \frac{\mathrm{d}\bar{u}}{\mathrm{d}t}\bigg|_{t=0},$$

in which $\bar{u}(u_l)$ satisfies the same definition as in Case 3. From (3.5) and (3.16), we derive

$$\begin{aligned} \lambda'(u_5) \frac{\mathrm{d}\bar{u}}{\mathrm{d}u_l}\bigg|_{u_l = u^- - 0} &= \frac{\lambda(u^-) - \lambda(u_5)}{u^- - u_5}, \\ \lambda'(u^{**}) \frac{\mathrm{d}\bar{\bar{u}}}{\mathrm{d}u_l}\bigg|_{u_l = u^- - 0} &= \frac{\lambda(u^-) - \lambda(u_5)}{u^- + q_2 - u^{**}}. \end{aligned}$$



$u_5 + q_2 > u^*$ $u_5 + q_2 < u^*$

FIG. 10.

Then by noting $\frac{\mathrm{d}u_l}{\mathrm{d}t}\big|_{t=0} = (\lambda(u_5) - \lambda(u^-))\dot{u}_0^-(0)$, (3.17) gives

$$u^{**} - u_5 \geq q_1 > 0,$$

from which we find that the construction of the combustion wave CJDT is impossible. Therefore, under the entropy conditions, a WDF $x = x^+(t)$ satisfying (3.7) is to be constructed in the domain $-\infty < \frac{x}{t} \leq \lambda(u_5)$. Now we do some analysis to determine the noncombustion jump behind the WDF. According to (3.9), it holds that along $x = x^+(t)$, $\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\big|_{t=0} > 0$ if $u_5 + q_2 > u^*$, whereas $\frac{\mathrm{d}\hat{u}}{\mathrm{d}t}\big|_{t=0} < 0$ if $u_5 + q_2 < u^*$. Here $u^* > u_5$ satisfies $f'(u_5) = \frac{f(u^*)-f(u^-)}{u^*-(u^-+q_2)} = \frac{f(u^*)-f(u^{**})}{u^*-u^{**}}$. For the case $u_5 + q_2 > u^*$, a shock wave $x = x^-(t)$ should be constructed (see Figure 10) as follows:

$$\begin{cases} \dot{x}^-(t) = w(\hat{u}, u_r) \quad \left(\lambda(u_5) \leq \dfrac{x}{t} \leq \lambda(u^+)\right), \\ \dfrac{x}{t} = \lambda(u_r) \quad (u^{**} \leq u_r \leq u^+), \\ x^-(0) = 0. \end{cases}$$

Furthermore, a computation similar to that in Case 4 results in

$$\ddot{x}^-(0) = \frac{(\lambda(u_5) - \lambda(u^-))^2(u_5 + q_2 - u^*)}{(u_5 - u^-)(u^{**} - u^*)}\dot{u}_0^-(0) > 0.$$

Instead, a right–contact discontinuity $x = x^-(t)$ such as

$$\begin{cases} \dot{x}^-(t) = \lambda(u_m) = w(\hat{u}, u_m) \quad (x^+(t) \leq x \leq \lambda(u_5)t), \\ x^-(0) = 0 \end{cases} \tag{3.18}$$

is constructed for the case $u_5 + q_2 < u^*$, where $u_m \in (u^{**}, \tilde{u})$. In the same way, we can get

$$\ddot{x}^-(0) = \frac{(\lambda(u_5) - \lambda(u^-))^2(u_5 + q_2 - u^*)}{(u_5 - u^-)(u^{**} - u^*)}\dot{u}_0^-(0) < 0.$$

Then by using the fact that $\ddot{x}^+(0) = \frac{(\lambda(u_5)-\lambda(u^-))^2}{u_5-u^-}\dot{u}_0^-(0) < 0$, we have

$$\ddot{x}^-(0) - \ddot{x}^+(0) = -\frac{(\lambda(u_5) - \lambda(u^+))^2(u_5 + q_2 - u^{**})}{(u_5 - u^-)(u^* - u^{**})}\dot{u}_0^-(0) > 0,$$

which, together with $x^-(0) = x^+(0) = 0$ and $\dot{x}^-(0) = \dot{x}^+(0) = \lambda(u_5)$, guarantees the existence of the solution $x = x^-(t)$ to (3.18) (see Figure 10).

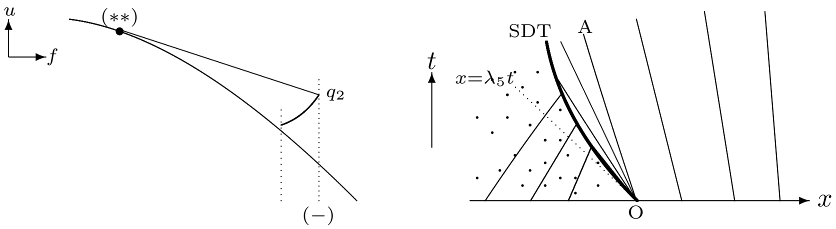


FIG. 11.

Next, we consider the case $\dot{u}_0^-(0) > 0$. It is possible, by the entropy conditions, to define an SDT $x = x(t)$ by solving the following problem:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \dfrac{f(u_r) - f(u_l)}{u_r - (u_l + q_2)} & \left(\lambda(u_5) \leq \dfrac{x}{t} \leq \lambda(u^+)\right), \\ x(0) = 0, \end{cases}$$

where $u_r(x,t)$ is a centered simple wave $x = \lambda(u_r)t$ $(u^{**} \leq u_r \leq u^+)$. It can be verified that the sufficient and necessary condition for the appearance of such an SDT is that along $x = x(t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{f(u_l) - f(u_r)}{u_l + q_2 - u_r}\right)\bigg|_{t=0} \geq \lambda'(u_5)\frac{\mathrm{d}\bar{u}}{\mathrm{d}t}\bigg|_{t=0};$$

namely,

$$\begin{aligned} &[(u^{**} - u^- - q_2)\lambda(u_5) - (f(u^{**}) - f(u^-))]\,\frac{\mathrm{d}u_r}{\mathrm{d}t}\bigg|_{t=0} \\ &\qquad - [(u^{**} - u^- - q_2)\lambda(u^-) - (f(u^{**}) - f(u^-))]\,\frac{\mathrm{d}u_l}{\mathrm{d}t}\bigg|_{t=0} \\ &\geq \frac{(\lambda(u_5) - \lambda(u^-))^2(u^- + q_2 - u^{**})^2}{u_5 - u^-}\,\dot{u}_0^-(0). \end{aligned} \tag{3.19}$$

Note that $f(u^{**}) - f(u^-) = \lambda(u_5)(u^{**} - u^- - q_2)$ and $\frac{\mathrm{d}u_r}{\mathrm{d}t}\big|_{t=0} \neq \infty$; then we find (3.19) is obviously true so that the solution of (1.7) and (1.8) involves an SDT indeed (see Figure 11).

Case 6. *The corresponding Riemann solution is SJ: $x = 0$.*

The case $f(u^+) = f(u^-)$ and $f'(u^+) < f'(u_5)$ is taken into account.

Due to the entropy condition $\lambda(u^+) < w(u^+, u^-) < \lambda(u^-)$, it is possible to construct a shock wave $x = x(t)$ with $\dot{x}(0) = 0$ in the burnt region, which is determined by

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = w(u_l, u_r) & \left(\lambda(u^+) \leq \dfrac{x}{t} \leq \lambda(u^-)\right), \\ x(0) = 0. \end{cases} \tag{3.20}$$

It is obvious that a solution corresponding to no reaction exists or, equivalently, that a shock wave such as (3.20) forms if and only if along $x = x(t)$,

$$\lambda(u^+)\frac{\mathrm{d}u_r}{\mathrm{d}t}\bigg|_{t=0} \geq \lambda(u^-)\frac{\mathrm{d}u_l}{\mathrm{d}t}\bigg|_{t=0};$$

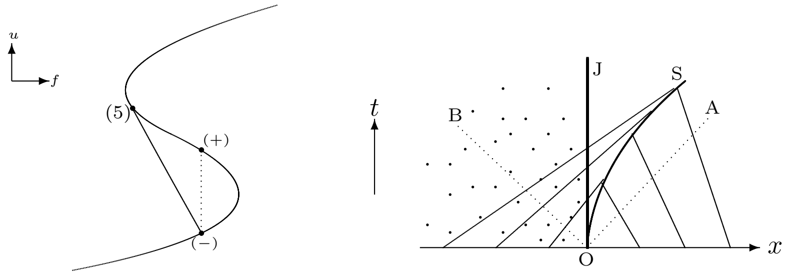


FIG. 12.

namely,

$$(\lambda(u^+))^2\dot{u}_0^+(0) \leq (\lambda(u^-))^2\dot{u}_0^-(0). \tag{3.21}$$

For the case $\dot{u}_0^+(0) < 0$ and $\dot{u}_0^-(0) > 0$, (3.21) obviously holds. So the solution can be denoted by $\mathrm{S}+\mathrm{J}$ (see Figure 12).

For the case $\dot{u}_0^+(0) > 0$ and $\dot{u}_0^-(0) > 0$, we give the discussion in detail. Certainly, burning does not happen, and the solution shown in Figure 12 can be constructed if (3.21) is satisfied. Now we deal with the cases when (3.21) fails, that is, the cases that involve combustion waves.

Suppose the SJ at $t = 0$ becomes an SDT $x = x(t)$ for $t > 0$, in which $x(t)$ with $\dot{x}(0) = 0$, $\ddot{x}(0) < 0$ can be expressed as (3.11). Then we have

$$(u^- + q_0 - u^+)\ddot{x}(0) = (\lambda(u^+))^2\dot{u}_0^+(0) - (\lambda(u^-))^2\dot{u}_0^-(0) > 0,$$

which implies $q_0 < u^+ - u^- =: q_1$. On the other hand, assuming $q_0 \in (0, q_1)$, it can easily be proved that there exists such an SDT as above.

When $q_0 = q_1$, it can be claimed that the SJ turns into either an SDT or a WDF behind which there is a shock wave. In fact, it is easy to show that the combustion wave is an SDT $x = x(t)$ if and only if along $x = x(t)$,

$$\left.\frac{\mathrm{d}\hat{\hat{u}}}{\mathrm{d}t}\right|_{t=0} \geq \left.\frac{\mathrm{d}u_r}{\mathrm{d}t}\right|_{t=0}, \tag{3.22}$$

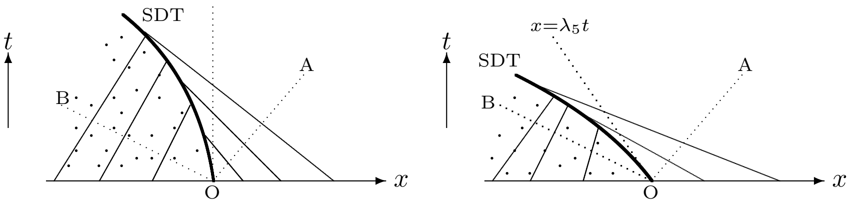
where $\hat{\hat{u}}(u_l)$ satisfies (3.12). With the condition $\dot{x}(0) = 0$, (3.22) is equivalent to

$$\lambda(u^-)(\lambda(u^-) - \lambda(u_5))\dot{u}_0^-(0) \geq \lambda(u^+)(\lambda(u^+) - \lambda(u_5))\dot{u}_0^+(0). \tag{3.23}$$

Thus the SDT with $\dot{x}(0) = 0$ occurs once (3.23) is satisfied. If the SDT has the property $\dot{x}(0+0) = \lambda(u_5)$, (3.22) becomes

$$(\lambda(u^-) - \lambda(u_5))^2\dot{u}_0^-(0) \geq (\lambda(u^+) - \lambda(u_5))^2\dot{u}_0^+(0). \tag{3.24}$$

Therefore we have two possibilities: an SDT with $\dot{x}(0) = 0$ (if (3.23) holds) or an SDT with $\dot{x}(0+0) = \lambda(u_5)$ (if (3.23) fails and (3.24) holds) (see Figure 13). When (3.24) fails to hold, the WDF $x = x^+(t)$ with $\dot{x}^+(0+0) = \lambda(u_5)$ is sure to appear by the entropy conditions. The structure of the solution is similar to that of Figure 8, in which the WDF $x = x^+(t)$ and the shock wave $x = x^-(t)$ also satisfy (3.7) and (3.14), respectively. The discussion on the existence of the shock wave is the same as that in Case 4.



(3.23) holds

(3.24) holds

Fig. 13.

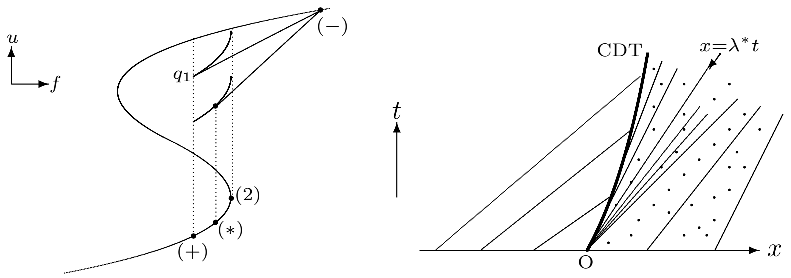


Fig. 14.

When $q_0 \in (q_1, +\infty)$, we can easily find that the perturbation gives rise to WDF + S.

Similarly we deal with the other cases, i.e., $\dot{u}_0^+(0) > 0$ (< 0) and $\dot{u}_0^-(0) < 0$, to obtain that the SJ becomes an SDT ($0 < q_0 < q_1$) or a WDF followed by a shock wave ($q_0 \geq q_1$) if a combustion wave emerges.

Notice that the case $f(u^+) = f(u^-)$, $f'(u^+) > f'(u_5)$ can be treated similarly to the convex case when $q_0 \in (0, q_2)$, where q_2 has the same meaning as in Case 5, while for $q_0 \in [q_2, +\infty)$, the same discussion as above can be carried out. Thus we do not elaborate on it here.

So far we have completed the discussion of all typical cases when the gas is unburnt at the left-hand side of the origin and burnt at the right-hand side at the initial time. Now we turn to the converse case to see what the perturbation will bring.

3.2. Solutions for $u^- > u_i \geq u_0^+(x)$, $q_0^-(x) = 0 < q_0 = q_0^+(x)$. In this subsection, we just mention briefly the cases which differ from those in section 3.1 and the convex ones. In the following, we fix $u^+ \in (-\infty, u_2)$.

Case 7. *A combustion wave CDT appears in the corresponding Riemann solution.*

First, we consider the case $u_i \in [u_2, u^+)$, $f(u^-) > f(u_2)$, and $f'(u^-) > f'(u^+)$. Let $q_1 > 0$ satisfy $f'(u^+) = \frac{f(u^+)-f(u^-)}{u^+ + q_1 - u^-}$ (see Figure 14). Then to ensure the appearance of the CDT, we take $q_0 \in (0, q_1]$. A proof similar to that in Case 1 shows that for $\dot{u}_0^-(0) > 0$, the CDT preserves its form as $x = x(t)$ satisfying

$$
(3.25) \qquad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \lambda(\bar{u}) = \dfrac{f(u_l) - f(\bar{u})}{u_l - (\bar{u} + q_0)}, \\ x(0) = 0, \\ \dot{x}(0) = \lambda(u^*), \end{cases}
$$

where $\bar{u}(u_l) \in (u^+, u_2)$ with $\bar{u}(u^- - 0) = u^*$ (see Figure 14). And for $\dot{u}_0^-(0) < 0$, the

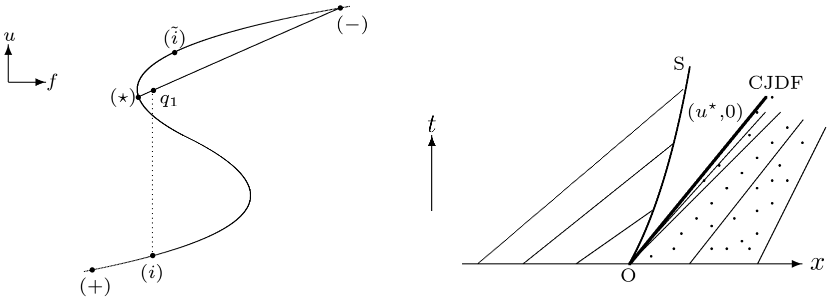


FIG. 15.

CDT changes into an SDT lying in the domain $\lambda(u^*)t \le x \le \lambda(u^-)t$.

Let us consider another case $u_i < u_2$, $f(\tilde{u}_i) \ge f(u_i)$, $u^- > \tilde{u}_i$, and $q_0 = q_2$, in which $q_2 > 0$, $\tilde{u}_i > u_1$ satisfy $f'(u_i) = f'(\tilde{u}_i) = \frac{f(u^-)-f(u_i)}{u^- -(u_i+q_2)}$ (see Figure 15). We also have two possible situations: as above, the CDT becomes an SDT in the domain $\lambda(u_i)t \le x \le \lambda(u^+)t$ provided that $\dot{u}_0^-(0) < 0$, while for $\dot{u}_0^-(0) > 0$, the combustion wave CJDF $x = \lambda(u_i)t$ appears. The state behind the CJDF is $(u^\star, 0)$, which is connected with $(u_r(x,t), 0)$ by a shock wave. Here $u^\star < \tilde{u}_i$, satisfying $f'(u_i) = \frac{f(u^-)-f(u^\star)}{u^- - u^\star} = \frac{f(u_i)-f(u^\star)}{u_i+q_2-u^\star}$ (see Figure 15).

Case 8. *A combustion wave DCC appears in the corresponding Riemann solution.*

From the property of DCC: $\lambda(u_r) = \sigma = \lambda(u_l)$, we easily find that the perturbation cannot affect a combustion wave of this kind, which is able to propagate with the speed at initial time.

We summarize our results in the following.

THEOREM 3.1. *There exists a unique solution to the generalized Riemann problem* (1.7) *and* (1.8). *For most of the cases, the corresponding Riemann solutions are stable, while for some typical cases, a small perturbation of initial data may lead to essential changes to the corresponding Riemann solutions. Especially, a strong detonation in the corresponding Riemann solution may turn into a weak deflagration followed by a shock wave after perturbation, which appears in the numerical simulations.*

REFERENCES

[1] W. BAO AND S. JIN, *The random projection method for hyperbolic conservation laws with stiff reaction terms*, J. Comput. Phys., 163 (2000), pp. 216–248.

[2] J. B. BDZIL AND D. S. STEWART, *The dynamics of detonation in explosive systems*, in Annual Review of Fluid Mechanics, Vol. 39, Annual Reviews, Palo Alto, CA, 2007, pp. 263–292.

[3] T. CHANG AND L. HSIAO, *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Pitman Monogr. Surveys Pure Appl. Math. 41, Longman Scientific and Technical, Harlow, UK, 1989.

[4] A. J. CHORIN, *Random choice methods with application to reacting gas flow*, J. Comput. Phys., 25 (1977), p. 253.

[5] R. COURANT AND K. O. FRIEDRICHS, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.

[6] W. FICKETT, *Detonation in miniature*, Amer. J. Phys., 47 (1979), pp. 1050–1059.

[7] T. T. LI AND W. C. YU, *Boundary Value Problem for Quasilinear Hyperbolic Systems*, Duke Univ. Math. Ser. 5, Duke University Press, Durham, NC, 1985.

[8] J. LI AND P. ZHANG, *The transition from Zeldovich–von Neumann–Döring to Chapman–Jouget theories for a nonconvex scalar combustion model*, SIAM J. Math. Anal., 34 (2003), pp. 675–699.

[9] T. P. LIU AND T. ZHANG, *A scalar combustion model*, Arch. Rational Mech. Anal., 114 (1991), pp. 297–312.

[10] A. Majda, *A qualitative model for dynamic combustion*, SIAM J. Appl. Math., 41 (1981), pp. 70–93.

[11] F. W. Sears and G. L. Salinger, *Thermodynamics, Kinetic Theory, and Statistical Thermodynamics*, Addison–Wesley, Reading, MA, 1975.

[12] W. C. Sheng and T. Zhang, *Structural stability of solutions to the Riemann problem for a scalar nonconvex combustion model*, Discrete Contin. Dyn. Syst., submitted.

[13] M. N. Sun and W. C. Sheng, *The ignition problem for a scalar nonconvex combustion model*, J. Differential Equations, 231 (2006), pp. 673–692.

[14] M. N. Sun and W. C. Sheng, *The generalized Riemann problem for a scalar Chapman-Jouguet combustion model*, Z. Angew. Math. Phys., to appear.

[15] Z.-H. Teng, A. J. Chorin, and T.-P. Liu, *Riemann problems for reacting gas, with applications to transition*, SIAM J. Appl. Math., 42 (1982), pp. 964–981.

[16] C. D. William, *The detonation of explosives*, Sci. Amer., 256 (1987), pp. 98–104.

[17] L. Ying and Z. Teng, *Riemann problem for a reaction and convection hyperbolic system*, Approx. Theory Appl., 1 (1984), pp. 95–122.

[18] X. T. Zhang and L. Ying, *Dependence of qualitative behavior of the numerical solutions on the ignition temperature for a combustion model*, J. Comput. Math., 23 (2005), pp. 337–350.

[19] P. Zhang and T. Zhang, *The Riemann problem for scalar CJ-combustion model without convexity*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 195–206.

[20] T. Zhang and Y. X. Zheng, *Two-dimensional Riemann problem for a single conservation law*, Trans. Amer. Math. Soc., 132 (1989), pp. 589–619.

[21] T. Zhang and Y. X. Zheng, *Riemann problem for gas dynamic combustion*, J. Differential Equations, 77 (1989), pp. 203–230.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 562–573

ADMISSIBILITY OF A WIDE CLUSTER SOLUTION IN "ANISOTROPIC" HIGHER-ORDER TRAFFIC FLOW MODELS[*]

RUI-YUE XU[†], PENG ZHANG[‡], SHI-QIANG DAI[†], AND S. C. WONG[§]

Abstract. We analytically investigate a wide cluster solution and show that it is not admitted in some of the traffic flow models in the literature. For those traffic flow models that admit the wide cluster solution, the relationship between two important control parameters and the critical densities that divide an equilibrium solution into stable and unstable regions is thoroughly discussed in detail. We find that such wide clusters exist with a free traffic density in an unstable region, and with one or three critical densities. These results are different from the cases in the well-known higher-order traffic flow models of Payne and Whitham [H. J. Payne, "Models of freeway traffic and control," in *Mathematical Models of Public Systems*, A. G. Bekey, ed., Simulation Council Proc. Ser. 1, La Jolla, CA, 1971, pp. 51–61], [G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, New York, 1974], Kühne [R. D. Kühne, "Macroscopic freeway model for dense traffic-stop-start waves and incident detection," in *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, J. Volmuller and R. Hamerslag, eds., VNU Science Press, Utrecht, 1984, pp. 21–42], and Kerner and Konhäuser [B. S. Kerner and P. Konhäuser, *Phys. Rev. E* (3), 50 (1994), pp. 54–83].

Key words. Aw and Rascle model, hyperbolic conservation law, wide cluster, shock wave, Rankine–Hugoniot condition

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/06066641X

1. Introduction.

Lighthill and Whitham [14] and Richards [17] independently proposed a hydrodynamic approach to study the traffic flow problems on a homogeneous highway, which is known in the literature as the first-order LWR model. The model has recently been extended to multilane [4] and multiclass models [3, 9, 10, 25, 31, 30, 32]. To embody the important nonlinear phenomena in traffic flow problems, some classical higher-order traffic flow models were developed [13, 12, 15, 23] and are characterized by their capability to reproduce the stop-and-go waves that are frequently observed on the highways. The formation of clusters is related to the instability of congested traffic, of which the development into free traffic and jams is typical of phase transitions and hystereses [12, 29, 28]. However, Daganzo [5] criticized the "isotropic" nature of these models, following which a stream of so-called anisotropic higher-order models [1, 8, 11, 19, 26, 27] was developed.

Recently, an analytical tool that is based on the weak solution theory was proposed to give a full and concise description of a wide cluster solution in higher-order models. Zhang, Wong, and Dai [33] used the model proposed by Jiang, Wu, and

[*] Received by the editors July 31, 2006; accepted for publication (in revised form) August 13, 2007; published electronically December 5, 2007. This work was jointly supported by the National Natural Science Foundation of China (grants 70629001, 10532060), the National Basic Research Program of China (grant 2006CB705503), the Research Grants Council of the Hong Kong Special Administrative Region, China (project HKU7187/05E), the Shanghai Leading Academic Discipline Project (grant Y0103), and the University of Hong Kong (grant 10207394).

http://www.siam.org/journals/siap/68-2/66641.html

[†] Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai 200072, People's Republic of China (monkeyxu1983@163.com, sqdai@126.com).

[‡] Corresponding author. Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai 200072, People's Republic of China (pengzhang@ustc.edu).

[§] Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, People's Republic of China (hhecwsc@hkucc.hku.hk).

Zhu [11] to demonstrate the solution procedure. Zhang and Wong [28] further showed that two conservation forms of the Payne–Whitham (PW) models [15, 23] have different cluster solutions that are asymptotic to the solutions of the Kühne [13] and Kerner and Konhäuser (KK) [12] models.

In this paper, we thoroughly investigate the admissibility of a wide cluster solution for several aforementioned "anisotropic" models, using the technique that was developed in [33, 28]. The model equations are introduced in section 2. In section 3, the procedure for solving a wide cluster is briefly discussed. While a criterion is prescribed as an essential condition, the formulations in [26, 27] are excluded from further consideration because they do not admit a wide cluster solution (see section 3.1). Here, we note that cluster solutions were also derived in Greenberg [6], Siebel and Mauser [20], and Wilson and Berg [24]. With two control parameters in the "pressure" $p(\rho) = \alpha(\rho/\rho_m)^\gamma$ of Aw and Rascle [1], the admissible regions of (γ, α) for the wide cluster solution are displayed. The regions of (γ, α) that are related to the stability of an equilibrium solution are also discussed (see section 3.2). It is novel to find that the free traffic flow (the constant flow with the minimal density) of a wide cluster solution is not stable in numerous regions, whereas the jam (the flow with the maximal density in the cluster) together with the whole solution is always stable. In some of these regions, it is evident that a wide cluster solution is admitted with one or three critical densities, which are used to divide equilibrium flows into stable and unstable regions. The results from numerical simulations (see section 4) are in good agreement with all analytical findings.

2. Model equations.

In macroscopic descriptions, vehicles on a highway are analogous to flows in compressible hydrodynamics [14, 17]. We have, by the mass conservation,

$$\rho_t + q_x = 0, \tag{1}$$

where $\rho(x,t)$ is the density, $q(x,t)$ is the flow, and no off- or on-ramp flows are considered along the highway. By defining an average speed $v = q/\rho$, the acceleration is considered in higher-order models. We write the acceleration of the discussed "anisotropic" models in the following mutual form:

$$v_t + vv_x = \frac{v_e(\rho) - v}{\tau} + \rho p'(\rho) v_x, \tag{2}$$

where $p(\rho)$ is the "pressure" with $p'(\rho) > 0$, and $v_e(\rho)$ is the equilibrium speed-density relationship with $v_e'(\rho) < 0$. For the derivation of a stable wide cluster, we assume a nonconvex fundamental diagram $q_e(\rho) = \rho v_e(\rho)$. Precisely, $q_e''(\rho) < 0$ for $0 \le \rho < \rho_I$, and $q_e''(\rho) > 0$ for $\rho_I < \rho \le \rho_m$, where $q_e''(\rho_I) = 0$, and ρ_m is the jam (maximum) density. Furthermore, we assume that $\tau > \tau_0$ for some fixed $\tau_0 > 0$, and thus the relaxation term is bounded. Nevertheless, this model is asymptotic to the LWR model [14, 17] if we allow $\tau \to 0$, which leads to $v \to v_e(\rho)$; see [16].

It is easy to derive the two characteristic speeds of system (1)–(2): $\lambda_1 = v - \rho p'(\rho)$ and $\lambda_2 = v$, which are no greater than the motion speed. Therefore, a perturbation propagates only upstream with respect to the perturbed moving vehicle. This is why the formulation is classified as "anisotropic." In contrast, the two characteristic speeds in the classical models are, respectively, smaller and greater than the traffic speed v, meaning that the perturbation propagates both upstream and downstream.

Several choices of the function $p(\rho)$ are listed in Table 1, corresponding to different

TABLE 1

"Anisotropic" formulations of higher-order traffic flow models, which are distinguished by the "pressure" $p(\rho)$ or the sound speed $c(\rho) = \rho p'(\rho)$, where $T(\rho) = t_r[1 + E/(1+\rho)]$ and constants $c_0, E, t_r > 0$.

Models	$c(\rho)$	$p(\rho)$
Aw and Rascle [1]	$\rho\, p'(\rho)$	$p(\rho)$
Jiang, Wu, and Zhu [11]	c_0	$c_0 \ln \rho$
Zhang [27]	$-\rho\, v_e'(\rho)$	$-v_e(\rho)$
Xue and Dai [26]	$-\rho\, \frac{t_r}{T(\rho)} v_e'(\rho)$	$-\frac{t_r}{T(\rho)} v_e(\rho)$

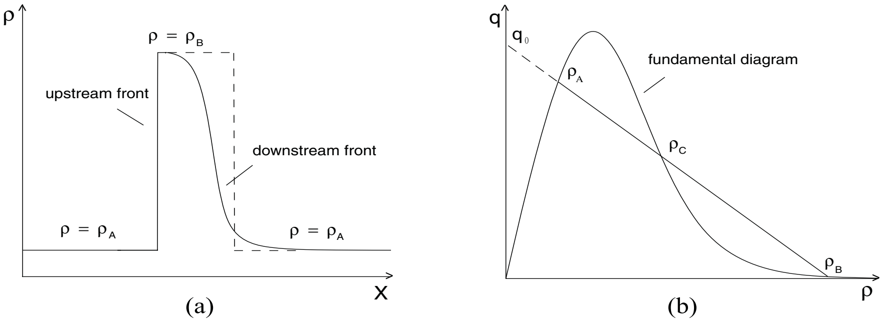


FIG. 1. *Illustration of a wide cluster solution:* (a) *profile of the density distribution;* (b) *phase plot $q = q(\rho)$ in comparison to the fundamental diagram $q = q_e(\rho)$.*

formulations in [1, 11, 26, 27]. We note that Aw and Rascle [1] contained a more general model by assuming an increasing function $p(\rho)$, and that the relaxation term was considered in [16] as an improvement; see also [2, 7] for further development of the Aw and Rascle model. Similarly, we add such a term to the formulation of Zhang [27]; without the term, it is unlikely to derive a wide cluster solution.

3. Admissibility of a wide cluster solution.

The profile of a wide cluster solution is shown in Figure 1(a). The cluster is expected to move backward with a constant velocity $a < 0$, without change in the shape. In other words, we assume a traveling wave solution $\rho(x,t) = \rho(X)$ and $v(x,t) = v(X)$ with $X = x - at$. Following the procedures in [33, 28], we can show that the downstream front is a smooth transition layer and the upstream front is a shock wave.

3.1. The downstream and upstream fronts.

For the downstream front that smoothly links a higher density region $\rho = \rho_B$ to a lower density region $\rho = \rho_A$, $\rho_A < \rho_B$, equations (1) and (2) are applicable. Hence, we replace $\rho(x,t)$ and $v(x,t)$ with $\rho(X)$ and $v(X)$ in the equations, which yields

$$q = a\rho + q_0 \tag{3}$$

and

$$\frac{d\rho}{dX} = g(\rho)\frac{q_e(\rho) - (a\rho + q_0)}{p'(\rho)\rho^2 - q_0}, \tag{4}$$

where the integration constant $q_0 > 0$, and thus the function $g(\rho) = \rho^2(\tau q_0)^{-1} > 0$. Equation (3) suggests a linear relation between the flow q and the density ρ, which is represented by a segment $\overline{AB}$ in Figure 1(b). As we assume that $d\rho/dX|_{\rho=\rho_A} = d\rho/dX|_{\rho=\rho_B} = 0$ (Figure 1(a)), (4) gives

$$(5) \qquad q_e(\rho_A) = a\rho_A + q_0, \quad q_e(\rho_B) = a\rho_B + q_0.$$

This implies that the two constant states ρ_A and ρ_B are in equilibrium (see Figure 1(b)). See [23] for a similar derivation. Let $(\rho_C, q_e(\rho_C))$ be the intersection of the segment $\overline{AB}$ and the fundamental diagram $q = q_e(\rho)$, i.e.,

$$(6) \qquad q_e(\rho_C) = a\rho_C + q_0.$$

It is obvious that the numerator of (4) is positive for $\rho \in (\rho_A, \rho_C)$ and negative for $\rho \in (\rho_C, \rho_B)$ (see Figure 1(b)). Accordingly, a decreasing transition layer ($d\rho/dX < 0$) is guaranteed if and only if

$$(7) \qquad p'(\rho_C)\rho_C^2 - q_0 = 0, \quad (p'(\rho)\rho^2 - q_0)(\rho - \rho_C) > 0 \quad \text{for } \rho \neq \rho_C.$$

Intuitively, we have $q_e'(\rho_C) < a < q_e'(\rho_B)$; see Figure 1(b). This along with (5)–(7) gives $v_e'(\rho_C) + p'(\rho_C) < 0$, and $v_e'(\rho_B) + p'(\rho_B) > 0$. By the linear stability conditions (see section 3.2), the two inequalities imply the following property.

PROPERTY 1. *For the solvability of a transition layer,* (i) ρ_C *must be located in an unstable region, whereas* (ii) ρ_B *must be located in a stable region.*

The two formulations attained by choosing $p(\rho) = -v_e(\rho)$ in [27] and $p(\rho) = -t_r v_e(\rho)/T$ in [26] (see Table 1) are not able to generate a transition layer, because equilibria of [27] and [26] can be easily shown to be uniformly stable and unstable, respectively. This also implies that they are not able to reproduce a wide cluster. For the choice of $p(\rho) = c_0 \ln \rho$ in [11] (see Table 1), the detailed discussion on the solution of a wide cluster was given in [33]. These three models are excluded from the forthcoming discussion.

We also note that Siebel and Mauser [19, 20, 21] adopted a strictly concave flow-density relationship $q_e(\rho) = \rho v_e(\rho)$ with the "pressure" $p(\rho) = -v_e(\rho)$. Although the convective term of their models resembled the model in Zhang [27], Siebel and Mauser introduced a coefficient $\beta(\rho, v)$ in the relaxation term which may change sign to reflect the interaction or difference between the relaxation time and reaction time of drivers. This allowed two critical densities (and thus an unstable regime) which correspond to the roots of the function $\beta(\rho, v)$. Then a traveling wave solution was obtained for this novel formulation [20].

In another development, Greenberg [6] derived the traveling wave solutions (clusters) using Lagrangian coordinates, in which the headway $s = 1/\rho$ and the car index m were taken as solution variables. Conservation across the discontinuous upstream front was also considered in [6], which in essence was similar to our discussion in what follows. Here, we remark that some intrinsic relations were implied in the formulations of Greenberg, Klar, and Rascle [8], Greenberg [6], and Siebel and Mauser [19].

In the forthcoming discussion, we assume that the denominator of (4) is an increasing function of ρ, that is,

$$(8) \qquad (p'(\rho)\rho^2)' \equiv \rho(\rho p(\rho))'' > 0,$$

which is sufficient to ensure the inequality in (7). Essentially, (8) or the convexity of the function $\rho p(\rho)$ was also assumed in [1, 8, 6, 19, 20, 21].

A monotonically increasing and smooth connection from ρ_A to ρ_B is impossible at the upstream front according to (3) and (4) and the detailed discussion in [33, 28]. We consider a shock wave with the following conservation form of (2),

$$\frac{\partial \rho(v+p(\rho))}{\partial t} + \frac{\partial \rho v(v+p(\rho))}{\partial x} = \frac{q_e(\rho)-q}{\tau}, \tag{9}$$

in [1, 16]. It is difficult, if not impossible, to define other conservation forms, except for that defined in [33] in which $c(\rho) = \rho p'(\rho)$ was taken as a constant. We note that different conservation forms result in different values of the characteristic parameters for solving a wide cluster [28].

To deal with the assumed shock that is also a traveling wave with the moving speed $a < 0$, we apply the Rankine–Hugoniot conditions to the conservation system of (1) and (9). This gives two equalities: one is implied in (5) and the other reads

$$a = \frac{q_e(\rho_B)(v_e(\rho_B)+p(\rho_B)) - q_e(\rho_A)(v_e(\rho_A)+p(\rho_A))}{(q_e(\rho_B)+\rho_B p(\rho_B)) - (q_e(\rho_A)+\rho_A p(\rho_A))}, \tag{10}$$

where the constant state $\rho = \rho_A$ (together with $\rho = \rho_B$) can easily be verified to be in equilibrium by (1) and (9). By (5)–(7) and (10) we have five independent algebraic equations to solve for five unknowns: a, ρ_A, ρ_B, ρ_C, and q_0. The solution of this algebraic equation system determines the assumed wide cluster solution.

3.2. Control parameters and solvability of the wide cluster. The speed-density relationship is taken as

$$v_e(\rho) = v_f((1+e^{12.5(\rho/\rho_m - 0.25)})^{-1} - (1+e^{12.5\times 0.75})^{-1}), \tag{11}$$

where v_f is the free-flow speed, ρ_m is the jam density, and the point of inflexion of the fundamental diagram $q_e(\rho) = \rho v_e(\rho)$ is located at $\rho_I \approx 0.333598$. Equation (11) is similar to that in [12] and was also applied in [33]. We take the "pressure" as

$$p(\rho) = \alpha(\rho/\rho_m)^\gamma, \quad \alpha, \gamma > 0. \tag{12}$$

Let $(\rho_0, q_e(\rho_0))$ be an equilibrium point in the fundamental diagram. The linear stability conditions for a constant solution $\rho = \rho_0$ can then be easily determined as $\lambda_1(\rho_0) \le q_e'(\rho_0) \le \lambda_2(\rho_0)$, which implies that the kinematic wave speed lies between the first and second characteristic speeds. See Whitham [23] for the relevant discussion. The inequalities are equivalent to

$$H(\rho_0) \equiv -1 - \frac{v_e'(\rho_0)}{p'(\rho_0)} \le 0, \tag{13}$$

because it is assumed that $p'(\rho_0) > 0$, and the condition of a single point at the boundary with $\rho_0 = 0$ is excluded. Taking the equality of (13) and with two control parameters α and γ, we indicate the solvability of the critical densities and the resultant stable and unstable regions. Hereafter, we denote dimensionless variables by placing a bar over them, such that a density is scaled by ρ_m and a speed (including the parameter α) is scaled by v_f. Using the dimensionless variables, we note that these critical densities and all characteristic parameters of the cluster solution depend on γ and $\bar{\alpha}$ only.

For the case $\gamma > 1$, say $\gamma = 2$ as is applied in the construction of Figure 2(a), monotone changes of the function $H(\rho_0)$ are divided into three intervals. This suggests

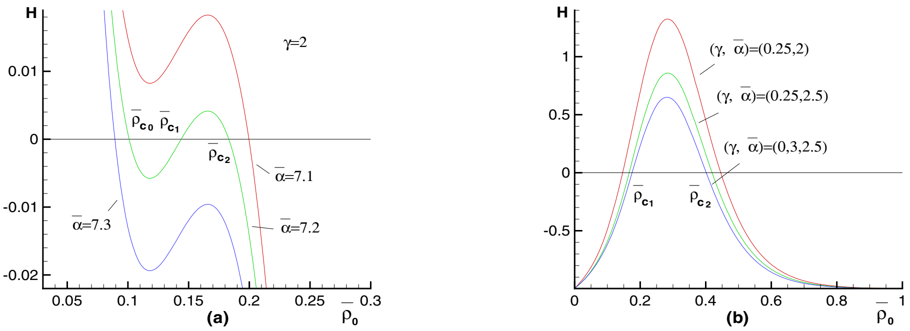


FIG. 2. *Solvability of critical densities and division of stable and unstable regions for an equilibrium constant solution* $\rho = \rho_0$ *for* (a) $\gamma > 1$ *and* (b) $\gamma \leq 1$.

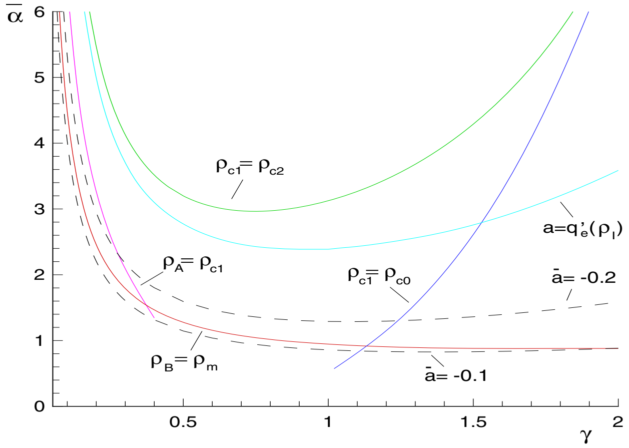


FIG. 3. *Division of the* γ-$\bar{\alpha}$ *plane for the study of a wide cluster solution. The solution exists below the curve* $a = q_e'(\rho_I)$, *and the traveling wave velocity* a *and the height* $\rho_B - \rho_A$ *decrease when* $(\gamma, \bar{\alpha})$ *approaches the curve. The equilibrium density* $\rho = \rho_A$ *of the cluster solution is located in a stable region only for approximately* $\gamma < 0.4$ *and* $(\gamma, \bar{\alpha})$ *in or below the curve* $\rho_A = \rho_{c_1}$.

at most three critical densities, ρ_{c_0}, ρ_{c_1}, and ρ_{c_2}. In this case, say with $\bar{\alpha} = 7.2$ as in the figure, the three critical densities divide the interval $(0, 1]$ into two stable intervals, $[\bar{\rho}_{c_0}, \bar{\rho}_{c_1}]$ and $[\bar{\rho}_{c_2}, 1]$, and two unstable intervals, $(0, \bar{\rho}_{c_0})$ and $(\bar{\rho}_{c_1}, \bar{\rho}_{c_2})$. As $\bar{\alpha}$ decreases (with reference to another curve $H(\bar{\rho}_0)$ for $\bar{\alpha} = 7.1 < 7.2$), $\bar{\rho}_{c_1}$ and $\bar{\rho}_{c_0}$ become identical for some $\bar{\alpha} \in (7.1, 7.2)$. This is to simultaneously have $H(\rho_{c_0}) = 0$ and $H'(\rho_{c_0}) = 0$, which determine a curve that is denoted by $\rho_{c_1} = \rho_{c_0}$ in the γ-$\bar{\alpha}$ coordinate plane (Figure 3). In the region that is below the curve $\rho_{c_1} = \rho_{c_0}$ (Figure 3), the critical densities $\bar{\rho}_{c_0}$ and $\bar{\rho}_{c_1}$ together with the interval $[\bar{\rho}_{c_0}, \bar{\rho}_{c_1}]$ must vanish because $\bar{\alpha}$ becomes smaller; see also the reference curve $H(\bar{\rho}_0)$ for $\bar{\alpha} = 7.1 < 7.2$ in Figure 2(a). In this case, we have an unstable interval $(0, \bar{\rho}_{c_2})$ and a stable interval $[\bar{\rho}_{c_2}, 1]$. Similarly, we can draw a curve $\rho_{c_1} = \rho_{c_2}$ in Figure 3, which is determined by setting $H(\rho_{c_2}) = 0$ and $H'(\rho_{c_2}) = 0$. The critical densities $\bar{\rho}_{c_1}$ and $\bar{\rho}_{c_2}$ along with the interval $(\bar{\rho}_{c_1}, \bar{\rho}_{c_2})$ vanish for $(\gamma, \bar{\alpha})$ in the region that is above this curve (Figure 3). See also the reference curve $H(\bar{\rho}_0)$ for $\bar{\alpha} = 7.3 > 7.2$ in Figure 2(a). In this case, we have an unstable interval $(0, \bar{\rho}_{c_0})$ and a stable interval $[\bar{\rho}_{c_0}, 1]$.

For the case $\bar{\gamma} \leq 1$, monotone changes of the function $H(\rho_0)$ are divided into two intervals. This suggests at most two critical densities $\bar{\rho}_{c_1}$ and $\bar{\rho}_{c_2}$. We define the same curve $\rho_{c_1} = \rho_{c_2}$ as previously discussed, and the curve goes smoothly in the whole region for $\gamma > 1$ and $\gamma \leq 1$ (Figure 3). However, in the region that is above the curve $\rho_{c_1} = \rho_{c_2}$ and for $\gamma \leq 1$, we have an overall stable interval $(0, 1]$.

We now turn our attention to the solvability and the characteristic parameters of the discussed wide cluster, which is related to the critical densities and the division of the stable and unstable intervals through Property 1 and the following discussion. With two control parameters $\bar{\alpha}$ and γ, we note that one equation in addition to (5)–(7) and (10) determines a curve in the γ-$\bar{\alpha}$ coordinate plane implicitly, which is also shown in Figure 3.

Defining the curve $\rho_B = \rho_m$, we can verify that the region below the curve $\rho_B = \rho_m$ suggests that $\rho_B > \rho_m$, by which the wide cluster solution is not collision-free. Therefore, this region is not considered in the forthcoming discussion. Furthermore, we define the curve $a = q_e'(\rho_I)$ ($\approx -0.542579 v_f$). By Figure 1(b), it is evident that the traveling wave speed a reaches its minimum with $\rho_A = \rho_C = \rho_B = \rho_I$ for $(\gamma, \bar{\alpha})$ in this curve. Here, the inequalities

$$\rho_A < \min(\rho_C, \rho_I) < \rho_B \tag{14}$$

are assumed for the wide cluster solution because of the nonconvexity of the function $q_e(\rho)$ (cf. Figure 1(b)). According to Property 1, it is implied that ρ_I is a critical density in the above limiting solution. Actually, we derive the same curve when defining $\rho_{c_2} = \rho_I$. For $(\gamma, \bar{\alpha})$ that is above this curve ($a = q_e'(\rho_I)$ or $\rho_{c_2} = \rho_I$), (5)–(7) and (10) are insolvable. This means that the iteration that is applied to solve these equations is never convergent under the restriction of (14). On the other hand, (5)–(7) and (10) are solvable for $(\gamma, \bar{\alpha})$ that is between the curves $\rho_B = \rho_m$ and $a = q_e'(\rho_I)$. Here, the curve $a = q_e'(\rho_I)$ serves as the other boundary to separate the two regions in which the wide cluster is, respectively, solvable and insolvable.

Let α increase. Then we find that the traveling wave speed a of the wide cluster decreases until a reaches its limiting value in the curve $a = q_e'(\rho_I)$. Two reference curves $\bar{a} = -0.1$ and $\bar{a} = -0.2$ are depicted in Figure 3 to show such a monotonic decreasing property. By the mass conservation at the discontinuous upstream front, a decreasing traveling wave speed usually suggests a drastic drop in the height $\rho_B - \rho_A$ of the wide cluster. The critical density ρ_{c_2} also decreases in this trend, which implies that $\rho_{c_2} > \rho_I$ for $(\gamma, \bar{\alpha})$ below the curve $a = q_e'(\rho_I)$, and $\rho_{c_2} < \rho_I$ for $(\gamma, \bar{\alpha})$ between the curves $a = q_e'(\rho_I)$ and $\rho_{c_1} = \rho_{c_2}$. This seems to suggest that it is essential that there exist a critical density that is greater than the inflexion ρ_I for the solvability of the wide cluster. On the other hand, it is novel that the wide cluster is insolvable even though we do have a critical density $\rho_{c_2} < \rho_I$ that is close to ρ_I.

The foregoing discussion together with Property 1 also suggests the locational relations between the critical densities and the two characteristic densities ρ_C and ρ_B. That is, $\rho = \rho_C$ and $\rho = \rho_B$ (as a constant portion of the wide cluster), respectively, are located in two adjacent unstable and stable intervals that is separated by $\rho_{c_2} > \rho_I$. This fact is also related to the following analytical property.

Property 2. *There is at most one critical density in* $[\rho_I, \rho_m]$.

This property holds simply because a critical density is also the root of the function $\rho^2 p'(\rho) H(\rho)$, and $(\rho^2 p'(\rho) H(\rho))' = -\rho(\rho p(\rho))'' - \rho(\rho v_e(\rho))'' < 0$ in $[\rho_I, \rho_m]$. For the solvability of the wide cluster, we do have such a critical density ρ_{c_2}. Here, it is implied that $\rho_{c_0} \leq \rho_{c_1} < \rho_I$ if the critical density ρ_{c_0} or ρ_{c_1} does exist.

To learn the locational relation between ρ_A and the critical densities, we draw the curve $\rho_A = \rho_{c_1}$ in Figure 3, which exists for $\gamma \leq 0.4$ approximately, where it meets with the curve $\rho_B = \rho_m$. Between the two curves we have $\rho_A \leq \rho_{c_1}$, which suggests that the equilibrium $\rho = \rho_A$ (as a constant portion of the wide cluster) is stable. In other regions where ρ_{c_1} does not vanish, we obviously have $\rho_{c_1} < \rho_A < \rho_I < \rho_{c_2}$, which suggests that $\rho = \rho_A$ is unstable. In the region that is below the curve $\rho_{c_0} = \rho_{c_1}$ for $\gamma > 1$, where ρ_{c_0} and ρ_{c_1} vanish, $\rho = \rho_A < \rho_I < \rho_{c_2}$ still belongs to the unstable region $(0, \rho_{c_2})$.

4. Wide clusters derived from numerical simulation. We write the system of (1) and (9) as the following standard conservation or balance laws:

$$u_t + f(u)_x = s(u), \tag{15}$$

where $u = (\rho, h)^T$, $h = \rho(v + p(\rho))$, $f(u) = (h - \rho p(\rho),\ \rho^{-1}h^2 - hp(\rho))^T$, and $s(u) = (0, \tau^{-1}(q_e(\rho) - h + \rho p(\rho)))^T$. For a numerical simulation, a conservative scheme of system (15) can be written as

$$u_i^{n+1} = u_i^n - \frac{\Delta t^n}{\Delta x}(\hat{f}_{i+1/2}^n - \hat{f}_{i-1/2}^n) + \Delta t^n s(u_i^n),$$

where $\Delta x = L/N$, and N is the grid number of the computational interval $(0, L)$. We apply the Lax–Friedrichs numerical flux:

$$\hat{f}_{i+1/2}^n = \frac{1}{2}(f(u_i^n) + f(u_{i+1}^n) - \mu^n(u_{i+1}^n - u_i^n)), \quad \mu^n = \max_u \max(|\lambda_1|, |\lambda_2|),$$

where $\lambda_{1,2}$ are the two characteristic speeds, and the maximum is taken over u_i^n for all i at time level n. The CFL condition that is necessary for numerical stability turns out to be $\Delta t^n \leq \mu^n \Delta x$; we always take $\Delta t^n = 0.7\mu^n \Delta x$ in numerical simulations. As the grids should be sufficiently refined to obtain the discussed wide cluster solution, we take a large grid number $N = 10000$ in all examples. For illustration purposes, only one state in every 20 grid points is shown in the figures. See [18, 22] for detailed accounts of the Lax–Friedrichs scheme. The reason for the application of this scheme rather than the Godunov or a higher-order scheme was explained in [28].

For an initial constant distribution $\rho(x, 0) = \rho_0$, which is in equilibrium with $v(x, 0) = v_e(\rho(x, 0))$, the small perturbation

$$\rho(x, 0) = \rho_0 + 0.005\rho_0(sgn(0.05 - |x/L - 0.5|) + 1)\sin(20\pi(x/L - 0.5)) \tag{16}$$

induces amplifying oscillations if density ρ_0 is located in an unstable region. Here, we define $sgn(s) = 1$ for $s \geq 0$, and $sgn(s) = -1$ for $s < 0$; the integral average density of $\rho(x, 0)$ over the computational interval $[0, L]$ is not changed by the perturbation. Applying the periodic boundary conditions

$$\rho(0, t) = \rho(L, 0), \quad v(0, t) = v(L, 0), \tag{17}$$

which ensure the conservation of the total vehicles in numerical simulation, the oscillations may evolve into stop-and-go waves or wide clusters in the long run if ρ_0 is sufficiently large. The dynamics of the evolution was well described in the PW [15, 23], Kühne [13], and KK [12] models; see also [28] for more relevant discussion.

As discussed in previous sections, the evolution of the perturbed constant flow $\rho = \rho_0$ is dependent on the control parameters γ and $\bar{\alpha}$, as shown in Figure 3. Our

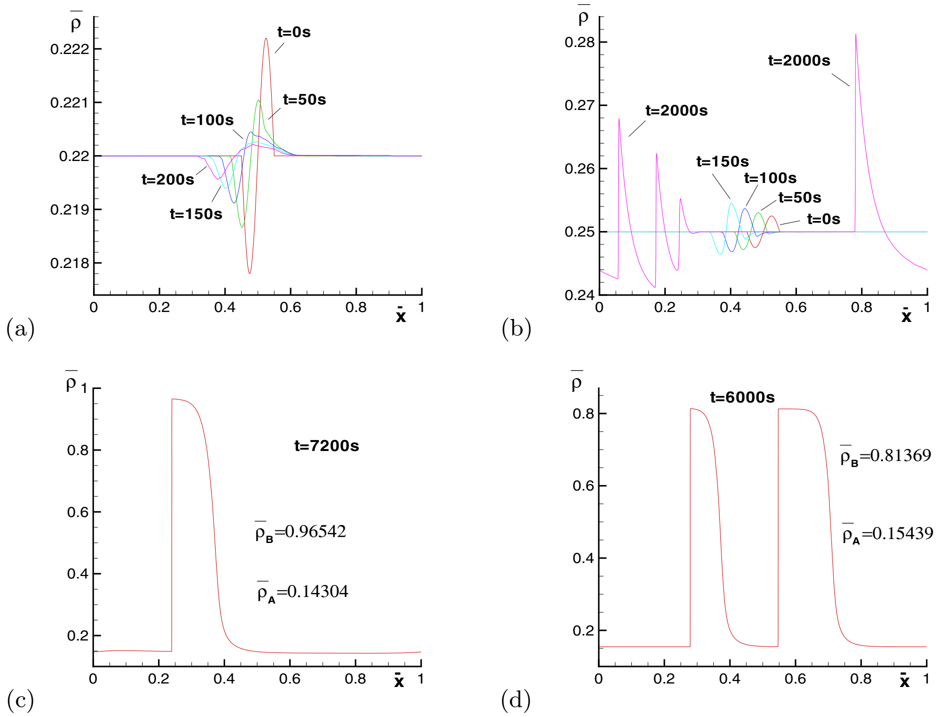


FIG. 4. *Evolution of unstable equilibrium flow $\rho = \rho_0$ with small perturbation for $(\gamma, \bar{\alpha})$ in different domains in Figure* 3. (a) $(\gamma, \bar{\alpha}) = (0.3, 4.2)$*,* $\bar{\rho}_0 = 0.22$*, and* $\tau = 10s$*;* (b) $(\gamma, \bar{\alpha}) = (0.7, 2.8)$*,* $\bar{\rho}_0 = 0.25$*, and* $\tau = 10s$*;* (c) $(\gamma, \bar{\alpha}) = (0.25, 2.1)$*,* $\bar{\rho}_0 = 0.25$*, and* $\tau = 18s$*;* (d) $(\gamma, \bar{\alpha}) = (0.5, 1.5)$*,* $\bar{\rho}_0 = 0.32$*, and* $\tau = 10s$*.*

TABLE 2

Characteristic parameters of a wide cluster and critical densities for comparison to the numerical results in Figures 4 and 5, where "IS" means insolvable and the values in the bracket are obtained from numerical simulation.

γ	$\bar{\alpha}$	$\bar{\rho}_A$	$\bar{\rho}_B$	$\bar{\rho}_C$	$\bar{a}$	$\bar{\rho}_{c0}$	$\bar{\rho}_{c1}$	$\bar{\rho}_{c2}$
0.3	4.2	IS	IS	IS	IS	IS	IS	IS
0.7	2.8	IS	IS	IS	IS	IS	0.226662	0.303168
0.25	2.1	0.142860 (0.14304)	0.968573 (0.96542)	0.332912	−0.137028	IS	0.150555	0.440170
0.5	1.5	0.153584 (0.15439)	0.817781 (0.81369)	0.334882	−0.176989	IS	0.139590	0.423337
1.5	1.5	0.162911 (0.16368)	0.680572 (0.67614)	0.346706	−0.229506	IS	IS	0.401206

numerical tests agree well with all these descriptions. In the domain that is on or above the curve $\rho_{c_1} = \rho_{c_2}$ in Figure 3, the perturbation decays with time because the solution $\rho = \rho_0$ is stable if $\bar{\gamma} \leq 1$ or, otherwise, if ρ_0 is greater than ρ_{c_0}, which is usually very small (see Figure 4(a)). In the domain between the curves $\rho_{c_1} = \rho_{c_2}$ and $a = q_e'(\rho_I)$, the perturbation increases with time but a wide cluster solution can never be developed regardless of the length of simulation (see Figure 4(b)). Here and hereafter, for all figures, the related parameters are shown in Table 2 and the variable x is scaled by $\bar{x} = x/L$. Through numerical simulation, we can always derive one or more wide clusters in other domains where the solution is predicted analytically. Two

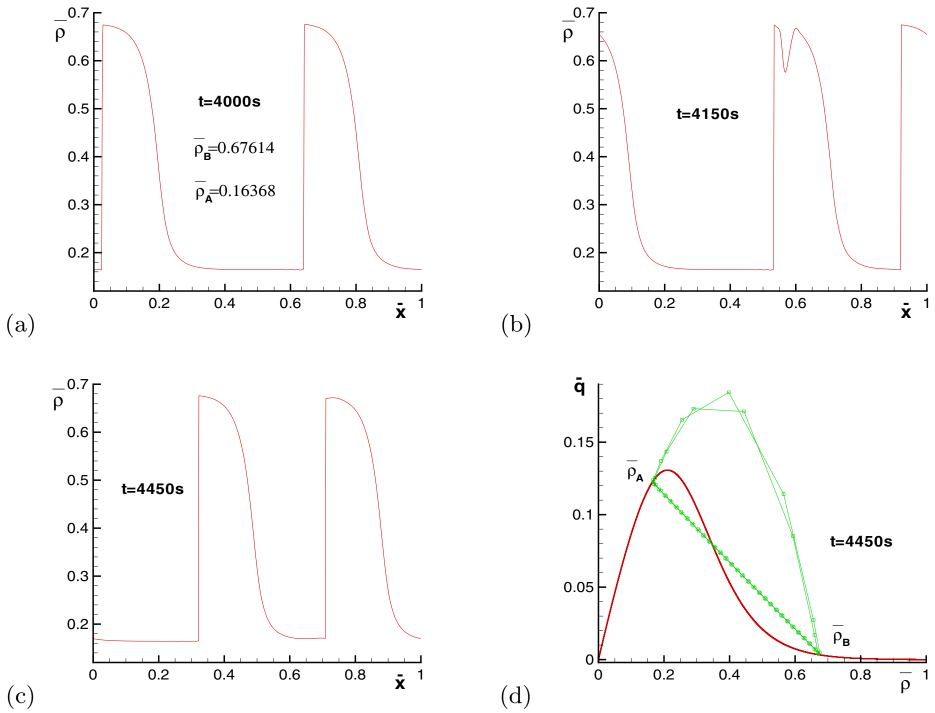


FIG. 5. *Stability test of a wide cluster solution with* $(\gamma, \bar{\alpha}) = (1.5, 1.5)$*,* $\bar{\rho}_0 = 0.33$*, and* $\tau = 10s$*:* (a) *two fully developed clusters;* (b) *density change due to a perturbation to the cluster solution;* (c) *recovery of the cluster solution from the perturbation;* (d) *the fundamental diagram* $(\bar{\rho}, \bar{q}_e(\bar{\rho}))$ *and the phase plot* $(\bar{\rho}, \bar{q})$*.*

such examples are shown in Figures 4(c) and (d).

It is novel that all predicted wide clusters are stable through numerical testing, even though the density $\rho = \rho_A$ of the free traffic is located in an unstable region. It is evident that a stable wide cluster solution exists with one or three critical densities for $\bar{\gamma} > 1$, as predicted analytically in Figure 3. We show a stability test in Figure 5 with $(\bar{\gamma}, \alpha)$ in this region, in which there is only one critical density (see also Table 2). Figure 5(a) shows two wide clusters that are derived at $t = 4000s$. By a perturbation that changes the speed $\bar{v}$ to $\bar{v} \mp 0.1$ for $\bar{x}$ between 0.5 and 0.5 ∓ 0.01, one cluster is found to be slightly distorted at $t = 4150s$. However, it soon recovers, as shown in Figure 5(c). In comparison to the fundamental diagram, we show the phase plot of the solution in Figure 5(d), where the segment $\overline{AB}$ represents the acceleration path of the downstream front. Nevertheless, the deceleration path of the two upstream fronts is now replaced by two curves from A to B. This takes place because a shock profile has to be smoothed by numerical viscosities, which can hardly be avoided in any scheme. The clusters that are shown in all of these figures follow the pattern that is anticipated in Figures 1(a) and (b).

A stable wide cluster solution with a length of unstable equilibrium $\rho = \rho_A$ might be well explained by the stable structure of the upstream front. Actually, the relaxation term of (9) vanishes for solution states on the both sides near the upstream front. Moreover, it is easy to verify that the Lax entropy conditions are satisfied with respect to this discontinuity (see also [6]). When the periodic boundary conditions are applied, a perturbation to $\rho = \rho_A$ (or other solution states) of the wide cluster is expected to be overtaken and thus "absorbed" by the upstream-moving shock. However, an affirmative conclusion could be made only through a rigorous mathematical proof, which is an interesting question for future study.

The parameter values of ρ_A and ρ_B acquired from numerical simulations are also shown in the figures, which are in good agreement with those that are derived from (5)–(7) and (10), and are listed in Table 2. This demonstrates that our numerical solutions converge to the analytical solutions of the described wide cluster.

5. Conclusions and discussions. We have thoroughly investigated the admissibility of a wide cluster solution in "anisotropic" higher-order models, in which the acceleration equation takes the functional form in (2). By this functional form (and also those in [6] and [19]), it appears essential that for the existence of the wide cluster solution the speed-density relationship $v_e(\rho)$ and the "pressure" $p(\rho)$ should be in "conflict" such that there exists an unstable regime in the vicinity of the congestion. Accordingly, it is evident that some formulations that are mentioned in this paper (and probably others) do not admit any wide cluster solution. Even with the functional forms that allow an unstable regime, the "pressure" $p(\rho)$ and the speed-density relationship $v_e(\rho)$ that represents the "force" for relaxation or fluctuation should act "harmoniously" to admit the solution. With reference to Figure 3, this means that a wide cluster is not always ensured or physically sound for all combinations of control parameters $(\gamma, \bar{\alpha})$. The intrinsic relationships between the aforementioned functions or "forces" can be left for future study.

REFERENCES

[1] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[2] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[3] P. Bagnerini and M. Rascle, *A multiclass homogenized hyperbolic model of traffic flow*, SIAM J. Math. Anal., 35 (2003), pp. 949–973.

[4] C. F. Daganzo, *A behavioral theory of multi-lane traffic flow, part* I: *Long homogeneous freeway sections*, Transportation Res. B, 36 (2002), pp. 131–158.

[5] C. F. Daganzo, *Requiem for second-order fluid approximations of traffic flow*, Transportation Res. B, 29 (1995), pp. 277–286.

[6] J. M. Greenberg, *Congestion redux*, SIAM J. Appl. Math., 64 (2004), pp. 1175–1185.

[7] J. M. Greenberg, *Extensions and amplifications of a traffic flow model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[8] J. M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.

[9] A. K. Gupta and V. K. Katiyar, *A new multi-class continuum model for traffic flow*, Transportmetrica, 3 (2007), pp. 73–85.

[10] M. Herty, C. Kirchener, and S. Moutari, *Multi-class traffic models on road networks*, Commun. Math. Sci., 4 (2006), pp. 591–608.

[11] R. Jiang, Q. S. Wu, and Z. J. Zhu, *A new continuum model for traffic flow and numerical tests*, Transportation Res. B, 36 (2002), pp. 405–419.

[12] B. S. Kerner and P. Konhäuser, *Structure and parameters of clusters in traffic flow*, Phys. Rev. E (3), 50 (1994), pp. 54–83.

[13] R. D. Kühne, *Macroscopic freeway model for dense traffic-stop-start waves and incident detection*, in Proceedings of the 9th International Symposium on Transportation and Traffic Theory, J. Volmuller and R. Hamerslag, eds., VNU Science Press, Utrecht, 1984, pp. 21–42.

[14] M. J. Lighthill and G. B. Whitham, *On kinematic waves:* II *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[15] H. J. Payne, *Models of freeway traffic and control*, in Mathematical Models of Public Systems, Vol. 1, A. G. Bekey, ed., Simulation Council Proc. Ser. 1, La Jolla, CA, 1971, pp. 51–61.

[16] M. Rascle, *An improved macroscopic model of traffic flow: Derivation and links with the Lighthill-Whitham model*, Math. Comput. Modelling, 35 (2002), pp. 581–590.

[17] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.

[18] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, Lecture Notes in Math. 1697, A. Quarteroni, ed., Springer, Berlin, 1998, pp. 325–432.

[19] F. Siebel and W. Mauser, *On the fundamental diagram of traffic flow*, SIAM J. Appl. Math., 66 (2006), pp. 1150–1162.

[20] F. Siebel and W. Mauser, *Synchronized flow and wide moving jams from balanced vehicular traffic*, Phys. Rev. E (3), 73 (2006), article 066108.

[21] F. Siebel and W. Mauser, *Balanced vehicular traffic at a bottleneck*, Phys. A, submitted.

[22] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer-Verlag, Berlin, 1999.

[23] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, New York, 1974.

[24] R. E. Wilson and P. Berg, *Existence and classification of traveling wave solutions to second order highway traffic models*, in Traffic and Granular Flow '01, M. Fukui, Y. Sugiyama, M. Schreckenberg, and D. E. Wolf, eds., Springer, Berlin, 2003, pp. 85–90.

[25] G. C. K. Wong and S. C. Wong, *A multi-class traffic flow model—an extension of LWR model with heterogeneous drivers*, Transportation Res. A, 36 (2002), pp. 827–841.

[26] Y. Xue and S. Q. Dai, *Continuum traffic model with the consideration of two delay time scales*, Phys. Rev. E (3), 68 (2003), article 066123.

[27] H. M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transportation Res. B, 36 (2002), pp. 275–290.

[28] P. Zhang and S. C. Wong, *Essence of conservation forms in the traveling wave solutions of higher-order traffic flow models*, Phys. Rev. E (3), 74 (2006), article 026109.

[29] P. Zhang, R. X. Liu, and S. C. Wong, *High-resolution numerical approximation of traffic flow problems with variable lanes and free-flow velocities*, Phys. Rev. E (3), 71 (2005), article 056704.

[30] P. Zhang, R. X. Liu, S. C. Wong, and S. Q. Dai, *Hyperbolicity and kinematic waves of a class of multi-population partial differential equations*, European J. Appl. Math., 17 (2006), pp. 171–200.

[31] M. P. Zhang, C. W. Shu, G. C. K. Wong, and S. C. Wong, *A weighted essentially non-oscillatory numerical scheme for a multi-class Lighthill-Whitham-Richards traffic flow model*, J. Comput. Phys., 191 (2003), pp. 639–659.

[32] P. Zhang, S. C. Wong, and C. W. Shu, *A weighted essentially non-oscillatory numerical scheme for a multi-class traffic flow model on an inhomogeneous highway*, J. Comput. Phys., 212 (2006), pp. 739–756.

[33] P. Zhang, S. C. Wong, and S. Q. Dai, *Characteristic parameters of a wide cluster in a higher-order traffic flow model*, Chinese Phys. Lett., 23 (2006), pp. 516–519.

SIAM J. Appl. Math.
Vol. 68, No. 2, pp. 574–597

FAST TOMOGRAPHIC RECONSTRUCTION VIA ROTATION-BASED HIERARCHICAL BACKPROJECTION*

ASHVIN GEORGE† AND YORAM BRESLER†

Abstract. We introduce a family of fast algorithms for tomographic backprojection in the parallel-beam geometry. The algorithms reduce the computational cost of backprojecting P projections onto an $N \times N$ pixel image from the conventional $O(N^2P)$ to $O(N^2 \log P)$. The new algorithms aggregate the projections in a hierarchical structure, with images in the hierarchy formed by the rotation and addition of other images made up of fewer projections. While these algorithms are related to existing fast algorithms, this work places them within the signal processing framework, providing a systematic means to optimize and adjust the trade-off between computational cost and accuracy. Rotations are performed separably in order that higher-order interpolators may be used with low computational cost. The same ideas are applied to create a tomographic projection algorithm, which computes projections of an $N \times N$ pixel image onto P view-angles at a cost of $O(N^2 \log P)$.

Key words. radon transform, fast algorithms, backprojection, projection, tomography, separable rotation, spline interpolation

AMS subject classifications. 92C55, 44A12, 65R10, 68U10

DOI. 10.1137/060668614

1. Introduction.

The problem of computed tomography in two dimensions is to reconstruct an image from a set of its line-integral projections. In addition to the ubiquitous computed tomography (CT) scanner, other medical applications of tomography are positron emission tomography (PET), single-photon computed tomography (SPECT), and, to a lesser extent, magnetic resonance imaging (MRI). Outside of medical imaging, tomographic imaging is used in security scanning, nondestructive evaluation, transmission electron microscopy (TEM), synthetic aperture radar (SAR), radio astronomy, geophysics, and other areas [14, 5]. The computational cost of the classic method used to estimate the image from its projections—the filtered backprojection (FBP) or convolution backprojection algorithm—is dominated by the step of backprojection. Traditionally, the backprojection of an $N \times N$ pixel image from P projections has a computational complexity of $O(N^2P)$. Fast algorithms exist that achieve $O(N^2 \log P)$ complexity. In practice, the traditional algorithm has been preferred because of the inadequate image quality of the fast algorithms. Our family of algorithms have $O(N^2 \log P)$ complexity and provide image quality comparable to the conventional method.

In medical and security applications, fast backprojection is necessitated by the increasing demand to rapidly process large amounts of data due to (a) the use of larger multirow detectors in CT machines and (b) the use of tomography to image moving objects (such as the beating heart or baggage on a conveyer belt). Another area where fast algorithms are useful is when the data is noisy, sparse, or otherwise degraded and an iterative reconstruction algorithm, involving multiple successive backprojections and reprojections, is used.

*Received by the editors August 30, 2006; accepted for publication (in revised form) September 7, 2007; published electronically December 5, 2007. This work was supported by NSF grants CCR 99-72980 and CCR 02-09203.

http://www.siam.org/journals/siap/68-2/66861.html

†Coordinated Science Laboratory, University of Illinois at Urbana Champaign, 1308 W.Main St., Urbana, IL 61801 (akgeorge@uiuc.edu, ybresler@uiuc.edu).

1.1. Basics of tomography. A *parallel-beam projection* $\mathcal{P}_\theta f$ of a two-dimensional (2D) image $f(x_1, x_2)$ is the set of all parallel straight-line integrals through the image, oriented at the view-angle θ, i.e., $\mathcal{P}_\theta f(t) \triangleq \int_{-\infty}^{\infty} f(t\cos\theta - s\sin\theta,\, t\sin\theta + s\cos\theta)ds$. The Radon transform is the set of all such projections at view-angles $\theta \in [0, \pi)$. Reconstructing a 2D image from a set of *parallel-beam line-integral projections* is equivalent to inverting the Radon transform.

In practice the set of view-angles is a discrete set. The *discrete-angle Radon transform* $\mathcal{R}$ is defined by $\mathcal{R}f(t,p) \triangleq \mathcal{P}_{\theta_p} f(t)$, where $p = 1, \ldots, P$.

The FBP algorithm [23], based on the continuous inversion formula discovered by Radon as early as 1917 (see [14]), involves first filtering the projections with a linear *ramp* filter (with frequency response $H(\omega) = |\omega|$), and then *backprojecting* those filtered projections $q(t,p)$. The discrete-angle *backprojection* operator $\mathcal{B}_{\vec{\theta}}$ (where the set of view-angles $\vec{\theta} \in [0, \pi]^P$) is defined by

$$(\mathcal{B}_{\vec{\theta}} q)(x_1, x_2) \triangleq \frac{\pi}{P} \sum_{p=1}^{P} q(x_1 \cos\theta_p + x_2 \sin\theta_p, p). \tag{1.1}$$

While backprojection can be defined for arbitrarily spaced view-angles $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_P)$, in this definition and the rest of the paper the simplifying assumption is made that the P view-angles are uniformly spaced in some π-length interval, i.e., $\theta_p = \theta_1 + (\pi/P)(p-1)$.

A straightforward implementation of the backprojection equation (1.1) in the discrete domain—reconstructing an $N \times N$ pixel image from P (sampled) projections—has a computational complexity of $O(N^2 P)$ because for each of the $O(N^2)$ points in the output image, $O(P)$ multiplications and additions must be performed to interpolate the sampled projections and evaluate the summation. This is the essence of the traditional, slow algorithm.

1.2. Other fast algorithms. Existing fast algorithms are either FFT-based [28, 17, 24, 1]—exploiting the low complexity of the FFT (fast Fourier transform)—or hierarchical [10, 9, 13, 26, 22, 11, 4]—taking a divide-and-conquer approach directly in the data or image domains..

Most FFT-based algorithms rely on the projection-slice theorem and fast methods to recover the image samples from nonuniformly spaced Fourier samples [16, 21]. Recent versions of these reconstruction algorithms [24] are reported to achieve significant speedup compared to conventional FBP. Andersson's algorithm [1] is conceptually different and interesting (it is based on expressing backprojection, in the log-polar coordinate system, as a convolution that can be implemented in the Fourier domain via an FFT) but appears to be not as computationally competitive in practice. Averbuch et al. [3, 2] define a rapidly computable, invertible Radon transform for *discrete* images, which also makes use of FFTs. Because the projections we consider are approximations to the projections of an underlying *continuous* image, they can be inverted only approximately. While the Averbuch groups' definition of the Radon transform is invertible, their inversion requires an iterative procedure whose complexity, although $O(N^2 \log N)$, has high proportionality constants and is therefore expensive.

Our algorithms fall into the second category. Hierarchical algorithms achieve their low complexity by recursively decomposing the problem of reconstructing an N-by-N image from P projections into smaller problems. This produces an algorithmic tree whose root is the original problem and whose leaves are the smallest problems into which it is decomposed. This hierarchy, when designed correctly, results in a reduction of the complexity of the algorithm from $O(N^2P)$ to $O(N^2 \log P)$, or $O(NP \log N)$, depending on the specific hierarchy chosen.

Basu and Bresler's algorithm [4] performs the dividing-and-conquering in the space domain. It may be classified, therefore, as a decimation-in-space algorithm. It uses the fact, known as the bow-tie result, which allows smaller-sized images to be backprojected from fewer projections. Dividing the image plane into a hierarchy of square subimages leads to a backprojection algorithm of $O(NP \log N)$ complexity.

There are several fast projection and backprojection algorithms [10, 9, 13, 26, 22, 11] that perform the dividing-and-conquering in the projection domain, partitioning and aggregating projections by their view angles. They may, therefore, be classified as decimation-in-angle algorithms. Brady's projection algorithm (see [10, 9]) calculates the discrete Radon transform of an $N \times N$ image (summations along sets of parallel lines through the image) in $O(N^2 \log N)$ time by hierarchically sharing partial sums of projections in similar directions. This hierarchical sharing of partial sums is reversed in Brandt et al.'s [11] and Nilsson's [26] backprojection algorithms. They recognize that images that are backprojected from a limited number of projections at angles close to a common direction may be sampled sparsely along the common direction. Thus the whole backprojected image is recursively constructed, with $O(N^2 \log P)$ complexity, by combining smaller (sparsely sampled) images. Danielsson [13] and Ingerhed [22] apply the hierarchical sharing of similar partial sums to the sinogram domain (similar to the way Brady [9] applies it to the image domain), recognizing that backprojection at a particular pixel in the image is the summation along a sinusoidal trace through the sinogram.

Our algorithm too applies the divide-and-conquer approach in the projection domain. It is most closely related to Brandt's algorithm, but places the idea of projection domain hierarchical backprojection within the framework of signal processing. The signal processing framework allows us to make rigorous the notion of sparsely sampling an image made up of few projections. It also allows for the use of optimizations such as separable image transformations (and consequently recursive infinite impulse response (IIR) filtering), fractional shifting, and integer-factor up-sampling. This improved understanding allows us to make significant improvements in accuracy and computational cost.

Each of the existing fast algorithms processes the data differently and may be therefore more appropriate for particular applications or computer architectures. Memory access, rather than arithmetic cost, can dominate computation on current computing architectures. For practical image sizes our algorithm offers an order of magnitude gain in both arithmetic and memory access over the conventional method for comparable image quality. Our algorithm has some other characteristics that make it particularly attractive compared to existing fast algorithms such as the decimation-in-space [4] and Fourier methods [24, 17]. It allows for the easy sequential processing of projections as they are acquired. The reduction in latency, resulting from the ability to begin computation without waiting for all the projections to be acquired, may be important in real-time imaging. This sequential processing is not as obviously or naturally implementable in the decimation-in-space algorithm, and is particularly difficult in Fourier-based algorithms. The cost of our algorithm also scales with the size of the region on which the reconstruction is performed—even in the case when the region of reconstruction has a nonconvex shape. This is not the case with Fourier-based methods. A comparison of different fast algorithms will require careful and optimized implementation of the algorithms to be compared on the computing architectures of interest. It is left for future work.

1.3. Reprojection. Reprojection, the operation of computing projections at a set of view-angles from a given image, has several applications in imaging including in algorithms for beam-hardening correction, streak suppression, removal of artifacts due to the presence of high-density objects, correction for missing data or partial volume effects, and compensation for attenuation errors (in PET and SPECT). As mentioned before, iterative reconstruction algorithms also involve reprojection (see [8] and the references therein) and make use of the fact that reprojection is the adjoint operation to backprojection [25]. By flow graph transposition [12], fast reprojection algorithms are derived from the fast backprojection algorithms that we introduce.

1.4. Other imaging geometries. The parallel beam tomographic geometry, discussed in this paper, arises in reconstruction problems in MRI and in electron microscopy, and so the algorithm described here can be applied directly to those problems. The algorithm can also be easily extended to the 3D Radon-transform problem (in which integrals are performed on parallel planes rather than lines).

In CT, projection data is usually available not in the parallel-beam, but in the divergent-beam (fan-beam or cone-beam) geometry. Reconstruction algorithms are performed directly on the divergent-beam data, or after it has been rebinned to the parallel-beam configuration [20]. The algorithms presented here can be applied to rebinned data but not (directly) to divergent-beam data.

Fast algorithms for other imaging geometries (most importantly fan-beam and cone-beam) are derived by extending our approach, and the Fourier-domain understanding which leads to the efficient sampling of images made up of few projections, to those geometries. This paper has been restricted to the 2D parallel geometry because this algorithm illustrates most of the important ideas and is a stepping stone for the more general algorithms.

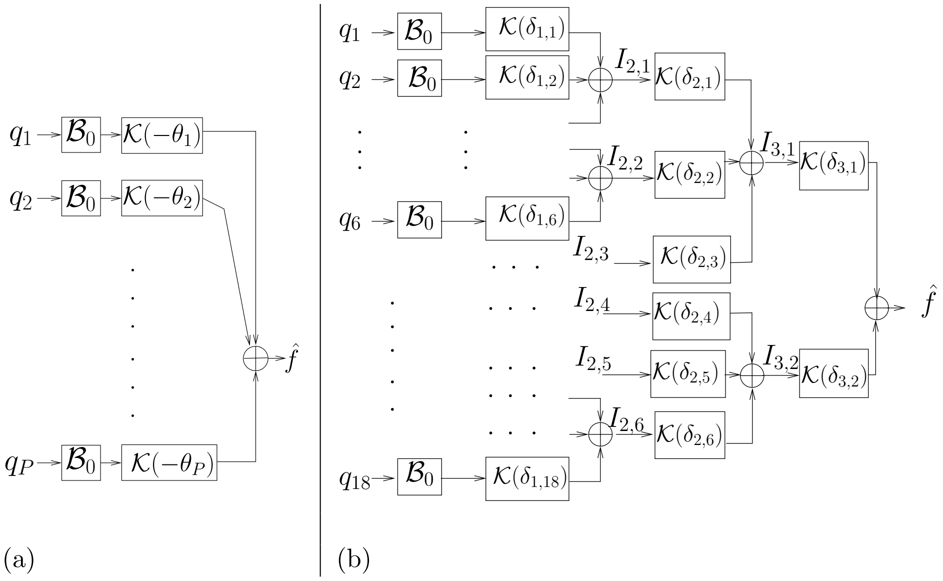
1.5. Brief summary. Following preliminary explanations in section 2, the actual algorithm is presented in section 3. The algorithm involves producing sparsely sampled images from single projections, and then combining them in a hierarchical manner using digital image transformation operations and additions of images. At every stage in the hierarchy (of $\log P$ levels), the number of images decreases by a factor of 3, while the density of samples increases by roughly the same factor. This keeps the cost of each level in the hierarchy $O(N^2)$ and, therefore, the cost of the whole algorithm $O(N^2 \log P)$. This scheme is motivated by Fourier-domain analysis (section 2.2.3), which explains the key idea of the algorithm: how an image backprojected from projections with a small angular range varies slowly in a direction transverse to those angles, and therefore can be sampled sparsely. Additional implementational details of the algorithm are included in section 3.

2. Fast hierarchical backprojection.

2.1. Backprojection as the sum of rotated images. The $O(N^2 \log P)$ hierarchical backprojection algorithms that we introduce are based on the decomposition of backprojection in terms of image rotations.

For convenience denote the P individual filtered projections as $q_p(t) \triangleq q(t,p)$, $p = 1, 2, \dots, P$. Restricting the backprojection operator $\mathcal{B}_{\vec{\theta}}$ (1.1) to a scalar $\theta \in [0, \pi)$, we define the backprojection of a function $q(t)$ at a single-angle θ as $(\mathcal{B}_\theta q)(\vec{x}) \triangleq (\pi/P)q(x_1 \cos\theta + x_2 \sin\theta)$. In particular,

$$(\mathcal{B}_0 q)(\vec{x}) = (\pi/P)q(x_1) = (\pi/P)q([\,1\ 0\,]\vec{x}). \tag{2.1}$$



q_1
q_2
q_P
$\mathcal{B}_0$
$\mathcal{K}(-\theta_1)$
$\mathcal{K}(-\theta_2)$
$\mathcal{K}(-\theta_P)$
$\hat{f}$
(a)
q_1
q_2
q_6
q_{18}
$\mathcal{K}(\delta_{1,1})$
$\mathcal{K}(\delta_{1,2})$
$\mathcal{K}(\delta_{1,6})$
$\mathcal{K}(\delta_{1,18})$
$I_{2,1}$
$I_{2,2}$
$I_{2,3}$
$I_{2,4}$
$I_{2,5}$
$I_{2,6}$
$\mathcal{K}(\delta_{2,1})$
$\mathcal{K}(\delta_{2,2})$
$\mathcal{K}(\delta_{2,3})$
$\mathcal{K}(\delta_{2,4})$
$\mathcal{K}(\delta_{2,5})$
$\mathcal{K}(\delta_{2,6})$
$I_{3,1}$
$I_{3,2}$
$\mathcal{K}(\delta_{3,1})$
$\mathcal{K}(\delta_{3,2})$
$\hat{f}$
(b)

FIG. 2.1. *Rotation-based backprojection.* (a) *Backprojection as the sum of rotated zero-backprojected images, as in* (2.4). (b) *The hierarchical equivalent of* (a).

It is easy to see that the full backprojection can be written in terms of the single-angle backprojection as

$$(\mathcal{B}_{\vec{\theta}}\{q_p\}_{p=1}^{P})(\vec{x}) = \sum_{p=1}^{P} (\mathcal{B}_{\theta_p} q_p)(\vec{x}). \tag{2.2}$$

Denoting the matrix of rotation by angle θ in the plane by $K_\theta = \left[\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right]$ and defining the rotation-by-θ operator $\mathcal{K}(\theta)$ by $(\mathcal{K}(\theta)f)(\vec{x}) \triangleq f(K_\theta \vec{x})$, we obtain immediately from (2.1) that

$$\mathcal{B}_\theta = \mathcal{K}(-\theta)\mathcal{B}_0. \tag{2.3}$$

Combining (2.2) with (2.3) leads to the starting point of the rotation-based hierarchical algorithm: the backprojection operator is equivalent to the summation of rotated zero-backprojected images, i.e.,

$$\hat{f} = B_{\vec{\theta}}\{q_p\}_{p=1}^{P} = \sum_{p=1}^{P} \mathcal{K}(-\theta_p)\mathcal{B}_0 q_p. \tag{2.4}$$

This is illustrated in Figure 2.1(a).

2.2. Hierarchical backprojection.

2.2.1. From parallel to a tree structure. The rotation of an $M \times N$ discrete-index image can be performed in $O(NM)$ operations [30]. Therefore, the complexity of backprojecting P projections onto an $N \times N$ image, according to Figure 2.1(a), is $P \times O(N^2) = O(N^2 P)$, which is no improvement over the conventional algorithm. The first key to improving the complexity to $O(N^2 \log P)$ is conversion to a hierarchical structure. Because the composition of successive rotations is still a rotation,

$$\mathcal{K}(\theta_1)\mathcal{K}(\theta_2)\ldots\mathcal{K}(\theta_N) = \mathcal{K}(\theta_1 + \theta_2 + \cdots + \theta_N), \tag{2.5}$$

we can rearrange the block diagram in Figure 2.1(a) into a hierarchical tree structure as shown in Figure 2.1(b). The intermediate image in the mth branch of the lth level is denoted as $I_{l,m}$. In the initial level

$$I_{1,m} \triangleq \mathcal{B}_0 q_m, \qquad m = 1, 2, \ldots, 2 \cdot 3^L. \tag{2.6}$$

In levels $l = 2, \ldots, L+1$,

$$\begin{aligned} I_{l,m} = {} & \mathcal{K}(\delta_{l-1,3m-2}) I_{l-1,3m-2} + \mathcal{K}(\delta_{l-1,3m-1}) I_{l-1,3m-1} \\ & + \mathcal{K}(\delta_{l-1,3m}) I_{l-1,3m}, \qquad m = 1, 2, \ldots, 2 \cdot 3^{L-l+1}, \end{aligned} \tag{2.7}$$

and in the final level

$$\hat{f} = I_{L+2,1} = \mathcal{K}(\delta_{L+1,1}) I_{L+1,1} + \mathcal{K}(\delta_{L+1,2}) I_{L+1,2}. \tag{2.8}$$

Even though the above explanation (2.6)–(2.8) describes the hierarchical algorithm for a set of exactly $P = 2 \cdot 3^L$ view-angles (or equivalently $L = \log_3(P/2)$), the algorithm can be generalized to arbitrary numbers and configurations of view-angles. The *ternary* branch (i.e., involving the combination of three images) is particularly efficient since, as is explained later, it eliminates the need for one third of the rotations. The efficient grouping of projections relies on the fact that a set of projections (or images) of any number can always be divided into groups such that all but one of the groups has exactly three members. Consequently a set of an arbitrary number of projections can always be hierarchically combined such that in every level all but one of the branches is ternary. The remaining branch involves the rotation and addition of two images (in the case of a binary branch) or no addition (in the case of a unary branch).

We will say that the hierarchical algorithm of (2.6)–(2.8) displayed in Figure 2.1(b) is correct if it is equivalent to that in Figure 2.1(a), i.e., if $\hat{f}$ in (2.4) and (2.8) coincide for every set of filtered projections $\{q_p\}_{p=1}^{P}$.

Now, for any set of projection angles θ_i, $i = 1, \ldots, P$, the intermediate rotation angles $\delta_{l,m}$ can be chosen so that the hierarchical algorithm is correct. Such a collection $\{\delta_{l,m}\}$ will be called *admissible* (for this set of projection angles). A trivial admissible set is $\delta_{1,i} = -\theta_i$, $i = 1, \ldots, P$, with the remaining $\delta_{l,m} = 0$. However, as we will show, certain other choices are preferable.

This hierarchical structure alone does not guarantee the $O(N^2 \log P)$ computational complexity. In fact, it involves up to $1.5 \times P - 1$ rotation operations—about 50% more than the single-level algorithm of (2.4) and Figure 2.1(a).

The second and essential key to reducing the complexity is to use the extra degrees of freedom provided by the additional rotations to enable sparse sampling of the underlying continuous images. This is possible by selecting the intermediate rotation angles $\{\delta_{l,m}\}$ such that the intermediate images have *low bandwidth.* Understanding and optimizing this procedure involves the interplay between three aspects: (i) the composition of the intermediate images, and the relationships between the angles of their constituent projections and the intermediate rotation angles; (ii) the spectral supports of intermediate images; and (iii) the sampling requirements of the intermediate images, and their effect on the computational requirements. We explore these aspects in the following subsections.

2.2.2. Intermediate images and rotation angles. To help identify the composition of intermediate images, we introduce a notational tool to pick out the angle of the pth projection in an image generated by a weighted backprojection of the form

$$I = \sum_{s \in \mathcal{N}'} \nu_s \mathcal{K}(\phi_s) \mathcal{B}_0 q_s, \tag{2.9}$$

where $\mathcal{N}' \subset \{1, 2, \ldots, P\}$ is an index set and $\nu_s \in \mathbb{R}$ are real weights.

Definition 2.1. *For an image I as in* (2.9) *and $p \in \mathcal{N}'$, the pth-angle extraction operator Φ_p is defined as $\Phi_p(I) \triangleq -\phi_p$.*

The following two properties follow in a straightforward way from Definition 2.1.

Lemma 2.2. *For the pth-angle extraction operator Φ_p:*

(a) $\Phi_p(\mathcal{K}(\theta)I) = \Phi_p(I) - \theta$.

(b) *Suppose that I_1 and I_2 are given by* (2.9) *with $\mathcal{N}' = \mathcal{N}_1$ and $\mathcal{N}' = \mathcal{N}_2$ respectively. If $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, then*

$$\Phi_p(I_1 + I_2) = \begin{cases} \Phi_p(I_1) & \textit{if } p \in \mathcal{N}_1, \\ \Phi_p(I_2) & \textit{if } p \in \mathcal{N}_2, \\ \textit{undefined} & \textit{if } p \notin (\mathcal{N}_1 \cup \mathcal{N}_2). \end{cases} \tag{2.10}$$

By iterating (2.7) it is easy to see that any intermediate image $I_{l,m}$ is given by

$$I_{l,m} = \sum_{p \in \mathcal{N}_{l,m}} \mathcal{K}(\phi_p) \mathcal{B}_0 q_p \tag{2.11}$$

for some set $\mathcal{N}_{l,m}$ of integer indices and angles $\{\phi_p : p \in \mathcal{N}_{l,m}\}$, which depend on the intermediate rotation angles $\delta_{l,m}$. Furthermore, it follows from (2.7) and (2.8) that for $l = 1, 2, \ldots, L+1$

$$\begin{gathered} \mathcal{N}_{l,m} = \{b, b+1, b+2, \ldots, e\}, \\ \text{where } b = 3^{l-1}(m-1) + 1 \quad \text{and} \quad e = 3^{(l-1)}(m-1) + 3^{l-1}. \end{gathered} \tag{2.12}$$

Combining Lemma 2.2 with (2.11) and (2.12), and denoting by $\lceil x \rceil$ the smallest integer larger than or equal to x, yields the following characterization of the backprojection angles of projections composing intermediate images in Figure 2.1(b).

Lemma 2.3. *Let $I_{l,m}$ be an intermediate image defined recursively by* (2.6)–(2.8). *Then for $l = 1, 2, \ldots, L+1$, $m = 1, 2, \ldots, 2 \cdot 3^{L-l+1}$, and any $p \in \mathcal{N}_{l,m}$, $\Phi_p(I_{l,m}) = -\sum_{i=1}^{l-1} \delta_{i,\mu(p,i)}$, where $\mu(p,l) \triangleq \lceil p/3^{l-1} \rceil$.*

Proof. For the proof, see Appendix A.

In particular, for $l = L+2$, Lemma 2.3 yields a characterization of the admissible set of intermediate rotation angles: $\theta_p = \Phi_p(I_{L+2,1}) = -\sum_{i=1}^{L+1} \delta_{i,\mu(p,i)}$, $p = 1, \ldots, P$. As expected, these conditions impose only P constraints on the up to $1.5P - 1$ free intermediate rotation angles $\{\delta_{l,m}\}$.

The following additional characterization of the intermediate images is useful in optimizing the intermediate rotation angles.

Consider the *virtual* image (i.e., one that is not actually formed) made up of projections indexed by the set $\mathcal{N}_{l,m}$, with each projection backprojected at its nominal view angle θ_p. Denoted $\tilde{I}_{l,m}$, this virtual image is given by

$$\tilde{I}_{l,m} = \sum_{p \in \mathcal{N}_{l,m}} \mathcal{B}_{\theta_p} q_p. \tag{2.13}$$

In general $\tilde{I}_{l,m} \neq I_{l,m}$. Now, by (2.3), we know that

$$\tilde{I}_{l,m} = \sum_{p \in \mathcal{N}_{l,m}} \mathcal{K}(-\theta_p)\mathcal{B}_0 q_p. \tag{2.14}$$

By Definition 2.1,

$$\Phi_p(\tilde{I}_{l,m}) = \theta_p \qquad \text{if } p \in \mathcal{N}_{l,m}. \tag{2.15}$$

Inspection of the block diagram in Figure 2.1(b) suggests that the relative angles between projections in an intermediate image are preserved in the final reconstructed image $\hat{f}$. This fact is captured by the following result.

Proposition 2.4. *If the algorithm of Figure* 2.1(b) *is correct, then for all* l, m *in the algorithm,*

$$I_{l,m} = \mathcal{K}(\alpha_{l,m})\tilde{I}_{l,m} \quad \textit{for some } \alpha_{l,m} \in (-\pi, \pi]. \tag{2.16}$$

Proof. For the proof, see Appendix B.

Because $\tilde{I}_{l,m}$ are virtual images, we call the $\{\alpha_{l,m}\}$ *virtual rotation angles.* They are related to the (actual) intermediate rotation angles of the hierarchical algorithm as follows.

The definition of $\mathcal{N}_{l,m}$ and the definition of the hierarchical algorithm together imply that (for $l = 2, 3, \ldots, L+1$) $\mathcal{N}_{l,m} = \mathcal{N}_{l-1,3m-2} \cup \mathcal{N}_{l-1,3m-1} \cup \mathcal{N}_{l-1,3m}$, and consequently by (2.13),

$$\tilde{I}_{l,m} = \tilde{I}_{l-1,3m-2} + \tilde{I}_{l-1,3m-1} + \tilde{I}_{l-1,3m}. \tag{2.17}$$

By Proposition 2.4,

$$\begin{aligned}\mathcal{K}(-\alpha_{l,m})I_{l,m} = \mathcal{K}(-\alpha_{l-1,3m-2})I_{l-1,3m-2} &+ \mathcal{K}(-\alpha_{l-1,3m-1})I_{l-1,3m-1} \\ &+ \mathcal{K}(-\alpha_{l-1,3m})I_{l-1,3m},\end{aligned}$$

and therefore

$$\begin{aligned}I_{l,m} = \mathcal{K}(\alpha_{l,m} - \alpha_{l-1,3m-2})I_{l-1,3m-2} &+ \mathcal{K}(\alpha_{l,m} - \alpha_{l-1,3m-1})I_{l-1,3m-1} \\ &+ \mathcal{K}(\alpha_{l,m} - \alpha_{l-1,3m})I_{l-1,3m}.\end{aligned}$$

Comparing this to (2.7), it follows that $\delta_{l-1,3m-i} = \alpha_{l,m} - \alpha_{l-1,3m-i}$, i.e.,

$$\delta_{l,3m-i} = \alpha_{l+1,m} - \alpha_{l,3m-i}. \tag{2.18}$$

2.2.3. Optimal rotation angles. The optimal angles $\alpha^*_{l,m}$ are chosen so that the intermediate images can be sparsely sampled. We use a rectangular sampling lattice with sampling intervals Δ_f and Δ_s; i.e., the discrete image $I^d(n_1, n_2) = I(\Delta_f n_1, \Delta_s n_2)$ is a sampled version of the continuous image. We maintain $\Delta_f \leq \Delta_s$ and, in keeping with the terminology of Brandt et al. [11], we call the first (horizontal) coordinate the *fast* direction and the second (vertical) coordinate the *slow* direction. These sampling intervals are chosen to satisfy the *sampling theorem* [15, p. 37] which, as explained in sections 2.2.4 and 3.3.1, dictates how small the sampling intervals need to be to represent an image with a given spectral support.

Equation (2.14) leads us to the spectral support of $\tilde{I}_{l,m}$. It is easy to see that the spectral support of $\mathcal{B}_0 q$ is restricted to the ω_1-axis (because $(\mathcal{B}_0 q)(x_1, x_2) = q(x_1)$

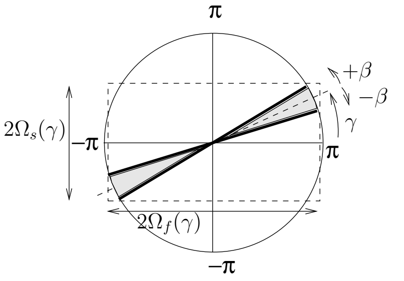


Fig. 2.2. *If the spectral support is the shaded wedge-shaped region within dark lines at angles* $\gamma \pm \beta$*, then the bandwidth in the fast and slow directions,* Ω_f *and* Ω_s*, is as indicated.*

and $\int (\mathcal{B}_0 q)(x_1, x_2) e^{j\omega_2 x_2} dx_2 = 2\pi q(x_1)\delta(\omega_2))$. Furthermore, if q has a bandwidth of π (i.e., its Fourier transform $Q(\omega) = 0$ for $\omega > \pi$), then the 2D spectral support of $\mathcal{B}_0 q$ is exactly $\{(\omega_1, 0) : \omega_1 \leq \pi\}$. It follows from the behavior of the Fourier transform under a spatial linear coordinate transformation ($g(\vec{x}) = f(K_\theta \vec{x}) \implies G(\vec{\omega}) = \frac{1}{|K_\theta|} F(K_\theta^{-T} \vec{\omega}) = F(K_\theta \vec{\omega})$) and (2.14) that the spectral support of $\tilde{I}_{l,m}$ is a wedge-shaped region $\tilde{W}_{l,m}$, as shown in Figure 2.2. In particular, if $\mathcal{N}_{l,m} = \{b, b+1, \ldots, e\}$, the wedge lies between θ_b and θ_e (i.e., $\theta_b = \gamma - \beta$ and $\theta_e = \gamma + \beta$ in Figure 2.2).

By (2.16), $W_{l,m}$, the spectral support of $I_{l,m}$, is also a wedge: in fact, it is just $\tilde{W}_{l,m}$ rotated by $\alpha_{l,m}$. The optimal angle $\alpha^*_{l,m}$ is one that minimizes the bandwidth $\Omega_s \Omega_f$ of $I_{l,m}$, where Ω_s and Ω_f are the bandwidths in slow and fast directions as shown in Figure 2.2. It is easily shown that the optimum virtual rotation angles and corresponding highest frequencies are

$$\alpha^*_{l,m} = \frac{\theta_b + \theta_e}{2}, \qquad \Omega_s(I_{l,m}) = \pi \sin \frac{\theta_e - \theta_b}{2}, \qquad \Omega_f(I_{l,m}) = \pi. \tag{2.19}$$

In view of (2.18), the optimum intermediate rotation angles $\delta^*_{l,m}$ of the hierarchical algorithm are completely determined by $\alpha^*_{l,m}$ as $\delta^*_{l,3m-i} = \alpha^*_{l+1,m} - \alpha^*_{l,3m-i}$.

Though the algorithm can be tailored to arbitrary sets and numbers of projections, we will simplify the description and analysis of the algorithm by assuming exactly $P = 2 \cdot 3^L$ projections uniformly distributed in angle as follows:

$$\theta_i = -\frac{\pi}{4}(1 - 3^{-L}) + \Delta_\theta (i - 1), \qquad i = 1, 2, \ldots, 2 \cdot 3^L,\ \Delta_\theta = \frac{\pi}{2 \cdot 3^L}. \tag{2.20}$$

This choice yields explicit expressions for the optimum rotation angles of the intermediate images. For $l = 2, 3, \ldots, L$,

$$\delta^*_{l,m} = \begin{cases} \Delta_\theta 3^{l-1} & \text{if } m = 1, 4, 7, \ldots, \\ 0 & \text{if } m = 2, 5, 8, \ldots, \\ -\Delta_\theta 3^{l-1} & \text{if } m = 3, 6, 9, \ldots. \end{cases} \tag{2.21}$$

Note that the center image of each triplet in the hierarchy is rotated by 0 radians—a free operation. This is the motivation for the choice of the ternary hierarchy in the common case of uniformly spaced projection angles as in (2.20). Turning now to the last, $l = L + 1$, level in the hierarchy, we again use the 0 radian rotation and another free rotation by $\pi/2$, which merely involves rearrangement (transposition) of pixels.

The algorithm of (2.6)–(2.8) accordingly simplifies to

$$
\begin{aligned}
&I_{l,m} = \mathcal{K}(\Delta_\theta 3^{l-2}) I_{l-1,3m-2} + I_{l-1,3m-1} + \mathcal{K}(-\Delta_\theta 3^{l-2}) I_{l-1,3m} \\
&\qquad \text{for } l = 2, 3, \ldots, L+1 \quad \text{and} \quad m = 1, 2, \ldots, 2 \cdot 3^{L-l+1}, \\
&I_{L+2,1} = I_{L+1,1} + \mathcal{K}(-\pi/2) I_{L+1,2} \qquad \text{and} \qquad \hat{f} = I_{L+2,1}.
\end{aligned} \tag{2.22}
$$

For this algorithm, with the rotation angles chosen per (2.21), substitution of (2.12) into (2.19) yields that the optimum bandwidths of $I_{l,m}$ are

$$
\Omega_s(I_{l,m}) = \pi \sin(\pi(3^{l-1} - 1)/(2P)), \qquad \Omega_f(I_{l,m}) = \pi. \tag{2.23}
$$

These bandwidths will determine the computational requirements and the discretization scheme of the algorithm.

2.2.4. Computational cost. The *sampling theorem* [15, p. 37] dictates how small the sampling intervals need to be in order for the intermediate image $I_{l,m}$ to be recovered from its sampled version $I^d_{l,m}$ (where $I^d_{l,m}(n_1, n_2) = I_{l,m}(n_1 \Delta_f^{l,m}, n_2 \Delta_s^{l,m})$). We combine the criterion of the *sampling theorem* ($\Delta_s < \pi/\Omega_s$ and $\Delta_f < \pi/\Omega_f$) with (2.23) to find the size of the discrete image $I^d_{l,m}$ in the lth level:

$$
\begin{aligned}
\text{size}(I^d_{l,m}) \approx (N/\Delta_s)(N/\Delta_f) &= (N\Omega_s(I_{l,m})/\pi)(N\Omega_f(I_{l,m})/\pi) \\
&= N^2 \sin(\pi(3^{l-1} - 1)/(2P)) \\
&< N^2(\pi 3^{l-1}/(2P)) = O(N^2 3^l/P).
\end{aligned}
$$

In the next section we will show that the cost of rotating a discrete-domain image containing S samples, to a given accuracy, is $O(S)$ arithmetic operations (adds and multiplies). Hence, because $\text{size}(I^d_{l,m}) = O(N^2 3^l/P)$, and because there are $P/3^{l-1}$ images in level l, the arithmetic complexity of the algorithm is

$$
\sum_{l=1}^{L+1} \frac{P}{3^{l-1}} O\left(\frac{N^2 3^l}{P}\right) = O(N^2 \log P). \tag{2.24}
$$

Rather than the number of arithmetic operations, the number of memory accesses or memory bandwidth is often the bottleneck in current computer architectures. Fortunately, the hierarchical algorithm provides a significant improvement in this respect too. In fact, counting the number of arithmetic operations provides a count of the number of memory accesses also, as memory accesses are performed only when arithmetic operations are performed. The number of memory accesses is therefore $O(N^2 \log P)$, which represents an improvement over the $O(N^2 P)$ memory accesses required in the conventional BP algorithm. Finally, a simple analysis shows that, if executed in place, the hierarchical algorithm requires $O(NP)$ memory—the same as the conventional algorithm.

3. Hierarchical backprojection in the discrete domain. We have established that the hierarchical algorithm discussed thus far has the favorable $O(N^2 \log P)$, scaling of the computational cost. However, for fixed N and P, the constants in the cost expression become important and will determine the actual speedup offered by the algorithm. Furthermore, various structural and parametric choices in the discrete index design of the algorithm will determine the trade-off between computation and

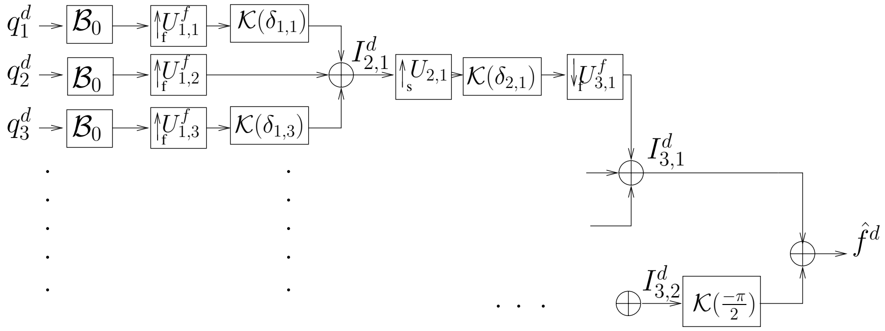

q_1^d
$\mathcal{B}_0$
$\uparrow_{\mathrm{f}} U_{1,1}^f$
$\mathcal{K}(\delta_{1,1})$
q_2^d
$\mathcal{B}_0$
$\uparrow_{\mathrm{f}} U_{1,2}^f$
$I_{2,1}^d$
$\uparrow_{\mathrm{s}} U_{2,1}$
$\mathcal{K}(\delta_{2,1})$
$\downarrow_{\mathrm{f}} U_{3,1}^f$
q_3^d
$\mathcal{B}_0$
$\uparrow_{\mathrm{f}} U_{1,3}^f$
$\mathcal{K}(\delta_{1,3})$
$I_{3,1}^d$
$\hat{f}^d$
$I_{3,2}^d$
$\mathcal{K}(\frac{-\pi}{2})$

FIG. 3.1. *Hierarchical backprojection in the discrete-domain with oversampling. This is a modification of Figure* 2.1(b)*. The modifications include very sparse sampling of the initial zero-backprojected* ($\mathcal{B}_0$) *images, upsampling in the slow direction* ($\uparrow_s$) *at every stage in the hierarchy to increase the density of sampling, and initial up-sampling* $\uparrow_f$ *and final down-sampling* $\downarrow_f$ *in the fast direction to tune the quality of the reconstruction.*

accuracy. We address these issues in this section, describing what we believe is a particularly favorable design.

The block diagram of this digital algorithm is shown in Figure 3.1 for an example with $L = 2$, i.e., for a set of $P = 2 \cdot 3^2 = 18$ projections. The projection view-angles are equally spaced with $\Delta_\theta = \pi/18$. The digital projections $q_p^d[m] = q_p(m\Delta_q)$ are sampled versions of the continuous projections, with uniform sampling interval Δ_q. The digital images $I_{l,m}^d$ are the sampled versions of the underlying continuous images $I_{l,m}$. The operation of sampling a continuous-domain image $I_{l,m}(\vec{x})$ on a rectangular lattice, with sampling periods Δ_1 and Δ_2, is denoted by $\mathcal{D}_{\vec{\Delta}}$, where $\vec{\Delta} = [\Delta_1, \Delta_2]$, and produces the 2D sequence or discrete-domain image $I_{l,m}^d[\vec{m}] = (\mathcal{D}_{\vec{\Delta}} I_{l,m})[\vec{m}] \triangleq I(\Delta_1 m_1, \Delta_2 m_2)$.

Blocks marked $\mathcal{B}_0$ represent the digital zero-angle backprojection operator (see (2.1)) $(B_0 q^d)[\vec{m}] = (\pi/P) q^d[m_1]$. Blocks marked $\mathcal{K}_\delta$ represent the discrete-domain rotation operator. The application of affine coordinate transformations ($(\mathcal{A}f)(\vec{x}) = f(A\vec{x})$, where A is a 2×2 matrix) to digital images is addressed in section 3.1, and a particularly efficient, separable implementation of the digital rotation operator using two shears is described in section 3.2.

The blocks labeled $\uparrow_s U_{l,m}$ represent up-sampling in the slow coordinate by factor $U_{l,m}$ (a digital affine transformation with $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1/U_{l,m} \end{smallmatrix}\right]$). Their role is to adjust the sampling periods $\vec{\Delta}_{l,m}$ of the intermediate images to the increasing slow bandwidth (viz. (2.23)). Section 3.3 explains how $\{U_{l,m}\}$ are determined. Finally, the blocks labeled $\uparrow_f U_{l,m}^f$ and $\downarrow_f U_{l,m}^f$ represent similar up- and down-resampling, respectively, in the fast direction, whose roles are discussed in sections 3.3.1 and 3.4.2.

3.1. Digital affine transformations. Each digital image $f^d[\vec{m}]$ in the algorithm will be considered to be a representation of an underlying continuous image $f_{\vec{\Delta}}^{\psi}(\vec{x})$ with respect to a particular basis function $\psi(\vec{x}) : \mathbb{R}^2 \to \mathbb{R}$ and sampling period $\vec{\Delta} \in \mathbb{R}^2$, the two images being related by

$$f_{\vec{\Delta}}^{\psi}(\vec{x}) \triangleq (E_{\vec{\Delta}}^{\psi} f^d)(\vec{x}) \triangleq \sum_{\vec{n} \in \mathbb{Z}^2} f^d[\vec{n}] \psi\left(\frac{x_1}{\Delta_1} - n_1, \frac{x_2}{\Delta_2} - n_2\right). \tag{3.1}$$

We will say that $f^{\psi}_{\vec{\Delta}}$ is the $\Delta - \psi$ continuous domain extension of f^d, and $E^{\psi}_{\vec{\Delta}}$ is the corresponding extension operator. We assume that $\Psi(\vec{x})$ has a bounded Fourier transform and $f^d \in l_2(\mathbb{Z}^2)$ so that the sum in (3.1) converges in $L_2(\mathbb{R}^2)$. The function ψ is chosen as an interpolant (vanishing on $\mathbb{Z}^2$, except at the origin), so that the digital image coincides with the samples of its $\Delta - \psi$ extension on the rectangular sampling lattice with period $\vec{\Delta}$, i.e.,

$$\mathcal{D}_{\vec{\Delta}} E^{\psi}_{\vec{\Delta}} f^d = f^d \qquad \forall f^d \in l_2. \tag{3.2}$$

With the relationship between digital and continuous-domain images established, we can define an affine transformation of a digital image.

The digital $\Delta - \psi$ affine transformation $\mathcal{A}^{\psi,\vec{\Delta}} : \ell_2(\mathbb{Z}^2) \rightarrow \ell_2(\mathbb{Z}^2)$ corresponding to the continuous affine transformation $\mathcal{A} : L_2(\mathbb{R}^2) \rightarrow L_2(\mathbb{R}^2)$ is defined by

$$\mathcal{A}^{\psi,\vec{\Delta}} f^d \triangleq \mathcal{D}_{\vec{\Delta}} \mathcal{A} E^{\psi}_{\vec{\Delta}} f^d = \mathcal{D}_{\vec{\Delta}} \mathcal{A} f^{\vec{\psi}}_{\vec{\Delta}}. \tag{3.3}$$

In other words, the transformation $\mathcal{A}^{\psi,\vec{\Delta}}$ is defined by applying the continuous affine transformation $\mathcal{A}$ to the $\Delta - \psi$ extension of the digital image, and then resampling. (Recall that $(\mathcal{A}f)(\vec{x}) \triangleq f(A\vec{x})$.) However, because $\mathcal{A}^{\psi,\vec{\Delta}}$ is a mapping between digital images, it is performed purely in the discrete-index domain. Under some additional conditions, we can show that this definition is *consistent*; that is, affine transformation $A^{\psi,\vec{\Delta}}$ yields the same result whether applied in the continuous or digital domain.

Let

$$F^d(\vec{\lambda}) \triangleq \sum_{\vec{n} \in \mathbb{Z}^2} f^d[\vec{n}] e^{-j\vec{\lambda}' \cdot \vec{n}}, \qquad \vec{\lambda} \in \mathbb{R}^2, \tag{3.4}$$

be the discrete-time Fourier transform (DTFT) of f^d. The square $[-\pi, \pi]^2$ is the principal period of $F^d(\lambda)$. If, on $[-\pi, \pi]^2$, $F^d(\lambda)$ vanishes outside a region W, f^d will be called band-limited to W, or $f^d \in B(W)$.

LEMMA 3.1. *Suppose* $f^d \in B(A^{-T}[-\pi, \pi]^2)$ *and* $\Psi(\vec{x}) = \mathrm{Sinc}(\pi x_1)\,\mathrm{Sinc}(\pi x_2)$*; then*

$$E^{\psi}_{\vec{\Delta}} \mathcal{A}^{\psi,\vec{\Delta}} f^d = \mathcal{A} E^{\psi}_{\vec{\Delta}} f^d. \tag{3.5}$$

Proof. For the proof, see Appendix C.

Recall now that the composition property $\mathcal{K}(\theta_1 + \theta_2) = \mathcal{K}(\theta_1)\mathcal{K}(\theta_2)$ of rotation is key to the hierarchical decomposition. The corresponding property for digital affine transformations is established by the following result.

PROPOSITION 3.2. *Let* $\mathcal{A}_1$ *and* $\mathcal{A}_2$ *be affine coordination transformations,* $\mathcal{A}_{21} = \mathcal{A}_2\mathcal{A}_1$*, and* $\mathcal{A}_1^{\psi,\vec{\Delta}}$*,* $\mathcal{A}_2^{\psi,\vec{\Delta}}$*, and* $\mathcal{A}_{2,1}^{\psi,\vec{\Delta}}$ *their corresponding digital versions. Suppose* $f^d \in B(A_1[-\pi, \pi]^2)$ *and* $\Psi(\vec{x}) = \mathrm{Sinc}(\pi x_1)\,\mathrm{Sinc}(\pi x_2)$*. Then*

$$\mathcal{A}_2^{\psi,\vec{\Delta}} \mathcal{A}_1^{\psi,\vec{\Delta}} f^d = \mathcal{A}_{21}^{\psi,\vec{\Delta}} f^d. \tag{3.6}$$

Proof. We have

$$\begin{aligned} \mathcal{A}_2^{\psi,\vec{\Delta}} \mathcal{A}_1^{\psi,\vec{\Delta}} f^d &= \mathcal{D}_{\vec{\Delta}} \mathcal{A}_2 E^{\psi}_{\vec{\Delta}} \mathcal{A}_1^{\psi,\vec{\Delta}} f^d = \mathcal{D}_{\vec{\Delta}} \mathcal{A}_2 \mathcal{A}_1 E^{\psi}_{\vec{\Delta}} f^d \\ &= \mathcal{D}_{\vec{\Delta}} \mathcal{A}_{21} E^{\psi}_{\vec{\Delta}} f^d = \mathcal{A}_{21}^{\psi,\vec{\Delta}} f^d, \end{aligned}$$

where the second equality follows from Lemma 3.1. □

Proposition 3.2 is the basis for converting the continuous-domain hierarchical algorithm of Figure 2.1(b) to the digital algorithm illustrated in Figure 3.1.

In practice, in order to limit computational cost, certain approximations must be made. Instead of the ideal band-limited interpolant $\Psi(\vec{x})$ we use, in (3.1), a $\Psi(\vec{x})$ of finite (usually small) support. This yields, via (3.3), a digital affine transformation with a kernel of small support on $\mathbb{Z}^2$, which is cheap to apply. Because $\Psi(\vec{x})$ is only approximately band-limited, the results in Lemma 3.1 and Proposition 3.2 hold only approximately. However, by using a certain degree of oversampling, the error can be made to fall off exponentially fast with the size of the support of the interpolant. More specifically, assume oversampling by factor $\gamma > 1$, i.e., $\Delta_i = \frac{1}{\gamma}\frac{\pi}{\Omega_i}$, and let $M \times M$ be the size of the support of the interpolant Ψ. Then the interpolation error introduced by (3.1) behaves as $ce^{-\pi M(\gamma-1)/\gamma}$ for some constant c (see [8]).

3.2. Two-shear rotations. The computation is further reduced by implementing each rotation in the algorithm as a cascade of one-dimensional (1D) transformations. We use the well-known decomposition of the rotation matrix

$$K_\theta = S_2^{\tan\theta} S_1^{-\sin\theta\cos\theta} S_c^{\cos\theta}$$

into shears along the fast and slow coordinates, $S_1^\alpha = \left[\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right]$, $S_2^\alpha = \left[\begin{smallmatrix} 1 & 0 \\ \alpha & 1 \end{smallmatrix}\right]$, and $S_c^\alpha = \left[\begin{smallmatrix} \alpha & 0 \\ 0 & 1/\alpha \end{smallmatrix}\right]$. The validity of a corresponding decomposition of digital rotation into a cascade of one-dimensional digital shears follows from Proposition 3.2.

Corollary 3.3. *Let* $A_1 = \left[\begin{smallmatrix} 1 & 0 \\ \tan\delta & 1 \end{smallmatrix}\right]$, $A_2 = \left[\begin{smallmatrix} 1 & -\sin\delta\cos\delta \\ 0 & 1 \end{smallmatrix}\right]$, *and* $\Psi(\vec{x}) = \operatorname{Sinc}(\pi x_1)\operatorname{Sinc}(\pi x_2)$, *and suppose that* $f^d \in l_2(\mathbb{Z}^2) \cap B(W)$, *where* $W = \{\vec{\omega} \in [-\pi, \pi]^2 : |\omega_1 + \omega_2 \tan\delta \Delta_1/\Delta_2| \leq \pi\}$. *Then*

$$\mathcal{A}_2^{\psi,\vec{\Delta}} \mathcal{A}_1^{\psi,\vec{\Delta}} f^d = \mathcal{D}_{\vec{\Delta}'} \mathcal{K}(\delta) E_{\vec{\Delta}}^{\psi} f^d, \tag{3.7}$$

where $\vec{\Delta}' = \left[\begin{smallmatrix} \cos\delta & 0 \\ 0 & 1/\cos\delta \end{smallmatrix}\right] \vec{\Delta}$.

Proof. Use Proposition 3.2 to obtain the result. □

The digital shear operations $\mathcal{A}_i^{\psi,\vec{\Delta}}$ are one-dimensional, involving digital filtering of the rows or columns of the digital image individually. This reduces the computational cost (per pixel) of a 2D image rotation using an $M \times M$ kernel $\Psi(\vec{x})$ from $O(M^2)$ to $O(M)$ for the two-shear version.

To further save in computation, we omit the coordinate scaling transformation $S_c^{\cos\theta}$ in the intermediate digital rotations, producing what we call a two-shear rotation. By Corollary 3.3 this implements combined digital rotation and resampling, rather than pure rotation. The δ-rotated image is effectively down-sampled in the fast coordinate and up-sampled in the slow coordinate by a common factor $1/\cos\theta$.

A fractional (noninteger) resampling of the initially backprojected images $\mathcal{B}_0 q_{\theta_p}$ is, therefore, required. Because these initial images are constant in the slow-direction, changing $\Delta_s^{l,m}$ will leave the image unchanged, and, consequently, the resampling in the slow-coordinate is a free operation. But, as displayed in Figure 3.1, in the initial level a fractional up-sampling in the fast direction is required (as $\Delta_f^{l,m} \leq 1.0$).

3.3. Optimum intermediate up-sampling factors.

3.3.1. Necessary sampling intervals. By (2.23), for the optimum rotation-angles the intermediate image bandwidths are $\Omega_s(I_{l,m}) = \pi\sin(\Delta_\theta(3^{l-1}-1)/2)$ and $\Omega_f(I_{l,m}) = \pi$. Therefore, in successive levels of the continuous-domain algorithm, as l increases, the *slow bandwidth* increases. Consequently the slow-sampling period required by the *sampling theorem* [15, p. 37], $\Delta_s^{l,m} < \pi/\Omega_s(I_{l,m})$, decreases. The required slow sampling rate adjustment from level to level is provided by the slow up-sampling operations $\uparrow_s U_{l,m}$ in Figure 3.1. Recall now from section 3.2 that the digital two-shear rotation introduces down-sampling in the slow direction and up-sampling in the fast direction by a factor $1/\cos(\delta_{l,m})$ at each level.

Consequently the sampling periods $\vec{\Delta}_{l,m}$ in adjacent levels of the algorithm, if the algorithm is correct, are related as follows:

$$\Delta_s^{l+1,\lceil m/3\rceil} = \Delta_s^{l,m}/(\kappa_{l,m}U_{l,m}) \quad \text{and} \quad \Delta_f^{l+1,\lceil m/3\rceil} = \kappa_{l,m}\Delta_f^{l,m}, \tag{3.8}$$

where

$$\kappa_{l,m} = \begin{cases} 1 & \text{if } m = 2,5,8,\dots,2\cdot 3^{L-l+1}-1 \text{ (no rotation)},\\ 1/\cos(\Delta_\theta 3^{l-1}) & \text{if } m = 1,3,4,6,\dots,2\cdot 3^{L-l+1} \text{ (rotation by } \pm\Delta_\theta 3^{l-1}). \end{cases} \tag{3.9}$$

Working backwards from $\Delta_s^{L+1,m} = 1$, we find that the slow-sampling period $\Delta_s^{l,m}$ is related to the up-sampling factors $U_{l',m}$ (for $l' > l$) as follows:

$$\Delta_s^{l,m} = \prod_{l'=l}^{L}(U_{l',\mu(l'-l,m)}\kappa_{l',\mu(l'-l,m)}), \tag{3.10}$$

where $\mu(l,m) = \lceil m/3^l\rceil$.

Next, combining equation (3.10) with the sampling condition $\Delta_s^{l,m} \le \pi/\Omega(I_{l,m}) = 1/\sin(\Delta_\theta(3^{l-1}-1)/2)$ provides necessary conditions on the slow direction up-sampling factors $U_{l,m}$.

Consider now the fast sampling interval Δ_f; because $\Omega_f(I_{l,m}) = \pi$ does not change between levels, no change in Δ_f is required to satisfy sampling requirements. However, the change in $\Delta_f^{l,m}$ produced by the two-shear digital rotation has to be accounted for. To avoid the cost of resampling operations in the fast direction of each level, it suffices to up-sample in level $L = 1$ by a factor $U_{1,l}^f$. Working backwards from $\Delta_f^{L+1,m} = 1$ and using (3.8) and (3.9) yields

$$U_{1,l}^f = \prod_{l'=l}^{L}\frac{1}{\kappa_{l',\mu(l'-l,m)}}. \tag{3.11}$$

3.3.2. Optimum slow up-sampling. Next, we formulate the choice of up-sampling factors as an optimization problem. The total computational cost is the sum of the costs of the image transformations and image additions in all branches of the algorithm which are, in turn, proportional to the sizes of the intermediate image. By (3.10), the dimensions of $I_{l,m}^d$ are $N/\Delta_s^{l,m}\times N/\Delta_f^{l,m} = (N/\prod_{l'=l}^{L}(U_{l',\mu(l'-l,m)}\kappa_{l',\mu(l'-l,m)}))$ $\times\, N/\prod_{l'=l}^{L}\kappa_{l',\mu(l'-l,m)}$. So, using 1D digital filters of fixed length M to implement the 1D digital coordinate transformations, the total cost can be shown to be equal to

$$\sum_{l=2}^{L}\sum_{m=1}^{6\cdot 3^{L-l}}\left(\frac{c_{l,m}}{\prod_{l'=l}^{L}U_{l',\mu(l'-l,m)}}\right) + \text{constant}, \tag{3.12}$$

where the constants $c_{l,m}$ depend on the computational costs of the filters used. This expression makes a mildly simplifying assumption. As written, it implies that the number of samples used to represent the (l, m)th image (i.e., the size of the sampled (l, m)th image) is inversely proportional to the slow-sampling interval. The size of the sampled image is equal to the above expression except for small additive constants that are due to the rounding effects that stem from the representation of an image by a whole number of samples and the retaining of a few samples beyond the strict edge of the image. The above expressions are therefore deemed acceptable for the calculation of optimal upsampling factors.

This cost (3.12) is to be minimized by choice of the factors $U_{l,m}$, subject to the sampling constraint derived from (2.23) in section 3.3.1 (i.e., $\Delta_s^{l,m} \leq \pi/\Omega_s(I_{l,m})$).

Practical considerations suggest yet another constraint on the $U_{l,m}$. The coefficients of digital filters implementing digital resampling operations can be either precomputed and stored (at the cost of memory and access time), or computed "on the fly." In the latter case, resampling by integer factors can be implemented more efficiently than that by fractional factors. Consequently, taking advantage of the freedom in selecting the slow up-sampling factors, we restrict them to integers. (No such freedom exists in the choice of fast direction resampling factors, which are fixed by (3.11)—but these fast direction resampling operations contribute a very small fraction of the total cost of the algorithm.)

The problem of minimizing the computational cost is thus equivalent to the following constrained integer optimization problem:

$$\mathcal{U}^* = \arg\min_{\mathcal{U}} \mathcal{J}(\mathcal{U}), \qquad \mathcal{U} = \{U_{l,m} \in \mathbb{N} : l = 2, 3, \ldots, L;\ m = 1, 2, \ldots, 6 \cdot 3^{L-l}\},$$

$$\mathcal{J}(\mathcal{U}) = \sum_{l=2}^{L} \sum_{m=1}^{6\cdot 3^{L-l}} \frac{c_{l,m}}{\prod_{l'=l}^{L} U_{l',\mu(l'-l,m)}}, \qquad \text{where } c_{l,m} \in \mathbb{R},\ \mu(l'-l, m) = \left\lceil m/3^{l'-l} \right\rceil,$$

subject to the constraint that

$$\Delta_s^{l,m} = \prod_{l'=l}^{L} (U_{l',\mu(l'-l,m)} \kappa_{l',\mu(l'-l,m)}) \leq \pi/\Omega_s(I_{l,m}) = 1/\sin(\Delta_\theta(3^{l-1} - 1)/2). \tag{3.13}$$

This optimization problem can be solved cheaply using dynamic programming [19].

3.4. Oversampling. As explained in section 3.1, oversampling the intermediate images will increase accuracy. Applying the oversampling condition uniformly, we require that each image undergoing a digital coordinate transformation be oversampled by at least some specified constant $\gamma > 1$.

3.4.1. Oversampling in the slow direction. The sampling frequency in the slow direction is controlled by the upsampling factors $U_{l,m}$. Upsampling by a factor γ is achieved by simply replacing the constraint (3.13) in the integer optimization problem by $\Delta_s^{l,m} \leq \frac{1}{\gamma}\pi/\Omega_s(I_{l,m})$ for $l = 2, 3, \ldots, L$.

3.4.2. Oversampling in the fast direction. In the fast direction, we simply increase the upsampling factors $U_{1,m}^f$ to incorporate oversampling, and then downsample the image in the fast direction at level L after the last transformation has been performed, to return to the desired sampling scheme (where $\Delta_f = \Delta_s = 1.0$). This modification to the algorithm therefore involves only one additional level of (fractional) x-resampling in level L, as shown in Figure 3.1. Note that though the block diagram shows these last down-sampling operations as separate, they are combined with rotations in four out of the six images in the Lth level for improved computational efficiency. This is achieved by combining the fast-direction shear operation in the decomposed rotation and the fast-direction down-sampling operation.

The exact values of these fast direction resampling fractions is determined both by the parameter γ and the spectral structure of the intermediate images. The γ oversampling condition is that $\Delta_f^{l,m} \leq \frac{1}{\gamma}\pi/\Omega_f(I_{l,m})$ for $l = 2, 3, \ldots, L$. We know that $\Omega_f(I_{l,m}) = \pi$, and $\{\Delta_f^{l,m}\}$ are related to each other strictly according to (3.8), so the largest of these values is $\Delta_f^{L+1,m}$ (as $\kappa_{l,m} \geq 1$). Consequently, the up-sampling factors $U_{l,m}^f$ are modified to $\tilde{U}_{1,m}^f = \gamma U_{1,m}^f$, and new down-sampling is introduced in level L, with down-sampling factors $\tilde{U}_{L+1,m} = \gamma$.

Note that the oversampling condition in the fast direction will not be satisfied for the first-level images, $\mathcal{B}_0 q_{\theta_p}\,|_{p=1,\ldots,P}$ if the input projections are sampled too sparsely (i.e., if $\Delta_q > \frac{1}{\gamma}$).

4. Fast hierarchical reprojection. The reprojection operator is known to be the adjoint of backprojection. Therefore, the fast hierarchical algorithm for reprojection can be derived from the backprojection algorithm by determining the adjoint of the sequence of operations that defines it. Accordingly, as shown in Figure 4.1, the reprojection algorithm may be formally derived from the block-diagrams of the backprojection algorithm (Figure 3.1) by a flow-graph transposition operation, reversing the flow of data [12]. Summation junctions in the backprojection block-diagram are replaced by simple branchings in the reprojection case, affine transforms (rotations by θ) are replaced by their adjoints (rotations by $-\theta$), up-sampling is replaced by down-sampling, and down-sampling by up-sampling, by the same factor. Finally, the zero-angle backprojection is replaced by zero-angle reprojection $\mathcal{P}_0$, where $(\mathcal{P}_0 f)(t) = \int_{-\infty}^{\infty} f(t,s)ds$. In the discrete implementation, it reduces to summation of pixel values along the columns of the image. As in the discrete backprojection algorithm, the rotations are implemented by two-shear rotations. The choice of parameters in the hierarchical reprojection algorithm is guided by the same principles as for the backprojection algorithm.

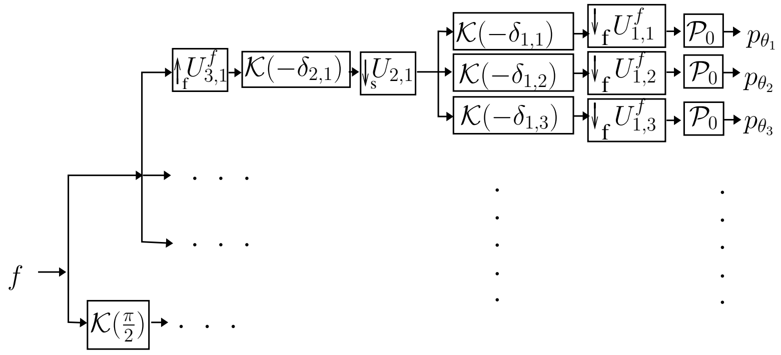


FIG. 4.1. *Ternary hierarchical reprojection. It is formed by a flow-graph transposition (i.e., reversal of the flow of data) of the block-diagram in Figure* 3.1.

5. Numerical experiments. The backprojection algorithms were tested in reconstructing the well-known Shepp–Logan phantom of size 512×512 pixels ($N = 512$).

The sampling-theoretic analysis of tomography [25] dictates that at least $N\pi/2 \approx 805$ projections are needed for reconstruction.

While the description of the algorithm is for $P = 2 \cdot 3^L$ view angles, the hierarchy is easily modified to accommodate $P = 2 \cdot 3^{L-1} \cdot T$ (for some small integers T and L). This is done by modifying the initial level so that it combines projections in groups of T instead of groups of 3.

We analytically generate $P = 1134 = 2 \cdot 3^4 \cdot 7$ projections at equally spaced angles in $(-\pi/4, 3\pi/4)$. (Because of the particular way we implement the digital image transformations, wherein we retain some samples beyond the strict edge of the image, it is cheaper to use a single 7-ary initial level than a ternary level preceded by an initial level with branches of varying branching-factors.) The use of 1134 projections means that there is some oversampling in the angular direction (which is not uncommon in practical systems). With fewer views, the relative acceleration provided by the hierarchical algorithm will be somewhat reduced (and the exact values will be specified later in this section). The projections are sampled with a sample-spacing of one pixel unit (and to simulate real tomographic systems which have detector elements of nonzero width, each projection is analytically convolved with the indicator function of the interval $(-0.5, 0.5)$). The projections are filtered with the well-known Shepp–Logan ramp filter.

The key trade-off in this algorithm is between accuracy, or the quality of the reconstructed image, and computational efficiency. Though visual quality is not easily quantifiable, we choose to measure the quality of the image reconstructed by the fast hierarchical algorithm in a few different ways. As reference images we use both the actual phantom and a reconstruction of the phantom using the conventional method. The conventional method that we use as a reference is pixel-driven backprojection, using linear interpolation between adjacent samples of a projection; cf. [23]. To quantify the difference between our $N \times N$ pixel reconstruction ($\hat{f}[\vec{m}]$) and the reference image ($f_{ref}[\vec{m}]$), we compute both the peak error ($\max_{\vec{m}} |\hat{f}[\vec{m}] - f_{ref}[\vec{m}]|$—the maximum difference between the pixel values of the reconstruction and the reference image) and the RMS (root mean square) error $= \sqrt{\frac{1}{N^2} \sum_{\vec{m}} |\hat{f}[\vec{m}] - f_{ref}[\vec{m}]|^2}$.

The computational cost of the algorithm is measured by counting the number of arithmetic operations—additions and multiplications—involved. By using this measure we can avoid the need to account for differences in processors and programming optimizations that affect the run-time of the particular code implementation. The *acceleration factor* relative to conventional backprojection is defined as the ratio of total operation count (adds + multiplies) for the conventional and the fast algorithm.

The trade-off between accuracy and cost in the fast algorithm is adjusted in two main ways—changing the oversampling parameter and changing the kind of interpolator used. The list of interpolators used is as follows: two FIR (finite impulse response) interpolators—the two-point linear interpolator and the three-point quadratic Schaum [29] interpolator, and three IIR (infinite impulse response) interpolators—Blu's shifted-linear interpolator [6], and Unsers's MOMS (spline) interpolators of quadratic and cubic degree [7].

Figure 5.1 displays the reconstruction of the phantom using the conventional and fast backprojection algorithms. In Figure 5.1(a) is shown the conventional reconstruction, and in Figure 5.1(b) is shown the reconstruction produced by the fast algorithm using the quadratic Schaum interpolator and an oversampling factor of $\gamma = 1.22$ (at an acceleration factor over the conventional algorithm of 9.7). While the pixel values of the phantom lie in the range $[0, 2.0]$, the displayed images in Figure 5.1(a) and (b) use

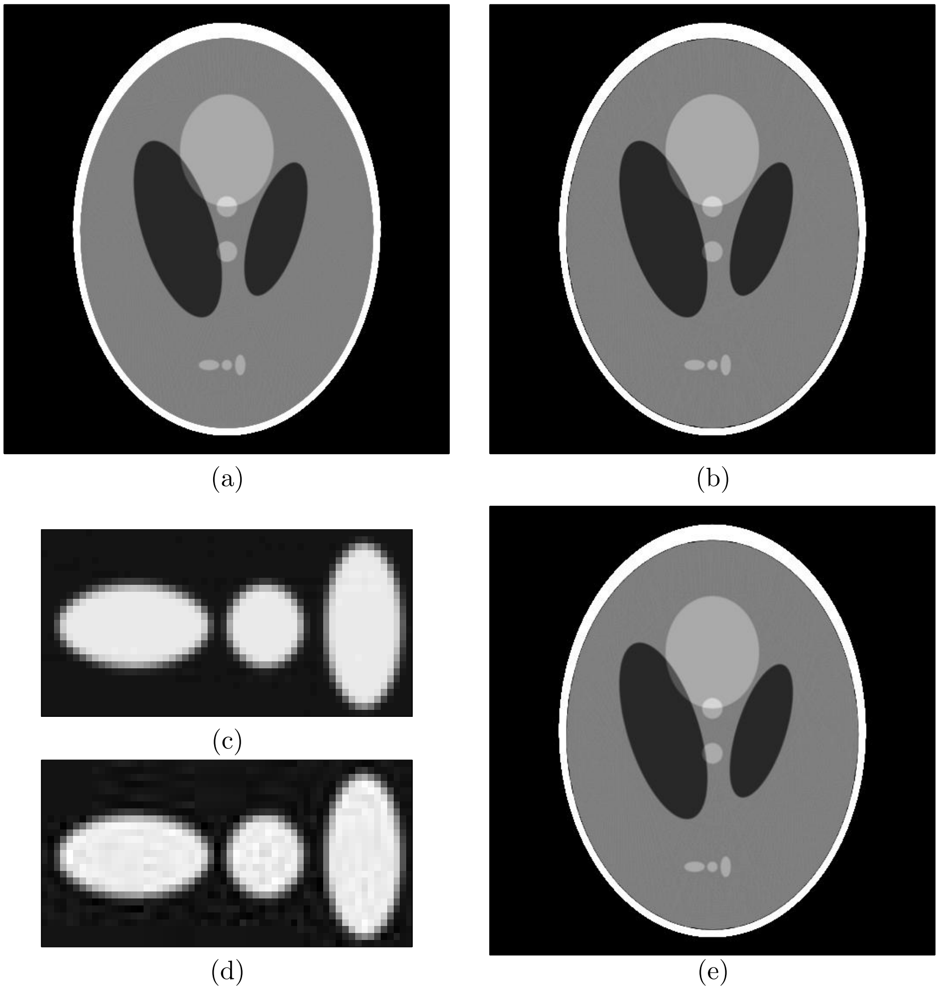


FIG. 5.1. *Sample reconstructions:* (a) *conventional backprojection,* (b) *fast backprojection (using the quadratic Schaum interpolator and oversampling $\gamma = 1.22$). The grayscale window in* (a) *and* (b) *is* $[0.99, 1.05]$*, which is* 3% *of the total range of pixel values.* (c) *Detail of* (a)*, and* (d) *detail of* (b)*. The grayscale window in* (c) *and* (d) *is* $[1.019, 1.031]$*, which is* 0.6% *of the total range of pixel values.* (e) *Fast backprojection (using the quadratic Schaum interpolator and oversampling $\gamma = 1.22$) with additional Kaiser windowing of projections with Kaiser parameter $\beta = 2.0$ (and grayscale window* $[0.99, 1.05]$*).*

a grayscale window of $[0.99, 1.05]$ (3% of the total range of pixel values) to emphasize small differences and assist the comparison of the reconstructions.

As mentioned previously, the relative acceleration is reduced when fewer view angles are used. In comparison to the acceleration of 9.7 achieved when using $P = 1134$ ($= 2 \cdot 3^4 \cdot 7$) projections, under the same conditions of quadratic Schaum interpolator and oversampling of $\gamma = 1.22$, the relative acceleration is 8.7 in the case of $P = 972$ ($= 2 \cdot 3^4 \cdot 6$), and 7.7 in the case of $P = 810$ ($= 2 \cdot 3^4 \cdot 5$).

The close-ups of Figure 5.1(a) and (b) are shown in Figure 5.1(c) and (d) (respectively). The closeups are visually comparable, even with a narrow grayscale window of $[1.019, 1.031]$ (0.6% of the total range of pixel values). The difference between the fast and slow reconstructions is most visible along the inner edge of the skull. The fast reconstruction displays an overshoot along this inner edge that is visible as a dark curve. A profile through the field of view is shown in Figure 5.2. A zoomed-in closeup of the inner edge of the skull is shown in Figure 5.2(b). Here, in addition to the fast algorithm (solid line), conventional algorithm (dashed line), and phantom (circles), is shown a fast reconstruction that uses a Kaiser window [27] (dotted line). The initial ramp filter is modified by a Kaiser window (with Kaiser parameter $\beta = 2.0$) to reduce the overshooting of the fast algorithm. The whole image (from the Kaiser-windowed fast reconstruction) is displayed in Figure 5.1(e). The reconstructions in the rest of this section do not use this Kaiser windowing.

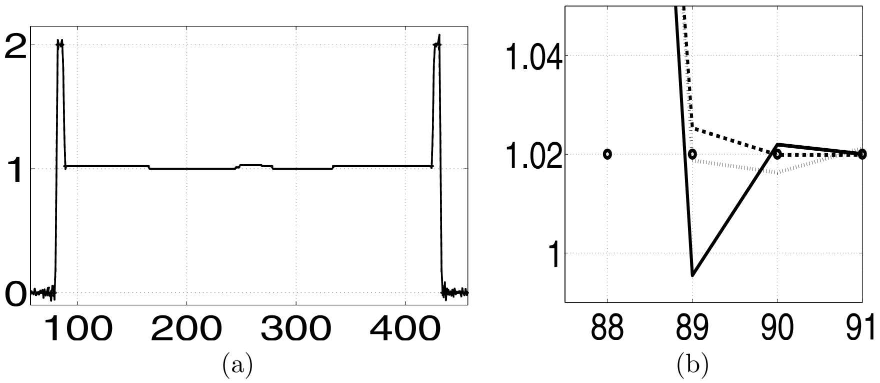


FIG. 5.2. *Profiles through reconstructions.* (a) *A profile of a row (number* 281*) through the conventional reconstruction and the fast reconstruction* without *Kaiser windowing. At this level of detail the differences are hard to see.* (b) *A zoomed in detail of the profile in* (a). *The circles represent the actual values of the phantom, the dashed line is the conventional reconstruction, the solid line is the fast reconstruction, and the dotted line is the fast reconstruction with Kaiser filtering. Notice that that Kaiser filtering reduces overshooting near edges.*

The RMS error of the reconstructions using the fast backprojection versus the acceleration factor over the conventional algorithm is plotted in Figure 5.3(a). The reference image is the original phantom. All five versions of the fast algorithm, each with a different interpolator, are run at different values of the oversampling parameter γ, ranging from 1.0 to 1.82. Data points corresponding to a particular choice of interpolator are connected by a line, and as the oversampling is increased, the cost of the algorithm increases (i.e., the acceleration factor decreases).

As expected, for a given interpolator, as the oversampling (and γ) is increased, the error decreases. For a given oversampling parameter γ, the error of the reconstruction decreases as expected as the complexity of the interpolator is increased—from linear, through shifted-linear, quadratic Schaum, quadratic MOMS, to cubic MOMS. In comparison the RMS error of the conventional algorithm is 0.059 (which is 3% of the maximum pixel value of the image) and is indicated by the horizontal dashed line in Figure 5.3(a).

The peak error of the reconstruction versus the acceleration over the conventional algorithm is plotted in Figure 5.3(b) and (c). Figure 5.3(b) shows the peak error (calculated with respect to the phantom) near the skull, and Figure 5.3(c) shows the peak error (calculated with respect to the conventional reconstruction) far from the skull. The reason for the observed reductions in peak error when oversampling is decreased or when a less complex interpolator is used is the suppression of high-frequency content associated with these choices, which reduces ringing and thereby peak error. The

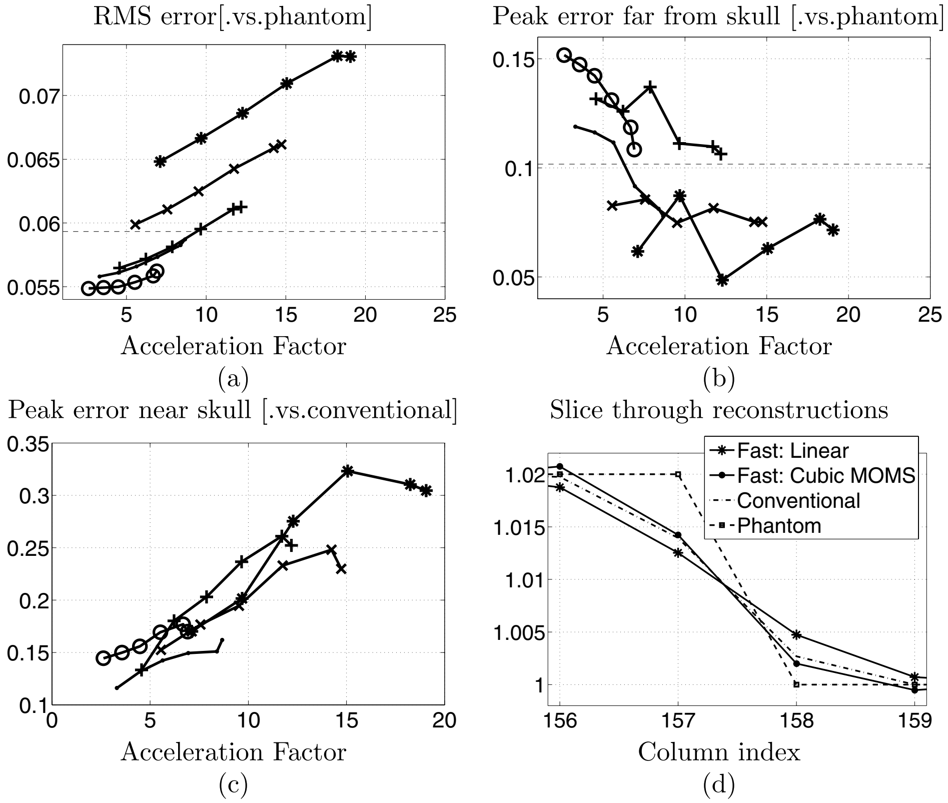


FIG. 5.3. *Comparing reconstructions—Error versus acceleration (*(a), (b), *and* (c)*) and slice profiles* (d). *Error is measured between the reconstruction and a reference image (conventional reconstruction or actual phantom). Acceleration is the cost of the conventional reconstruction divided by the cost of the particular fast reconstruction.* (a) *The RMS error (referenced against phantom),* (b) *the peak error away from the skull (referenced against phantom), and* (c) *the peak error near the skull (referenced against conventional). Interpolation types in* (a), (b), *and* (c): *cubic MOMS* ($\circ$), *quadratic MOMS* ($\cdot$), *quadratic Schaum* ($+$), *shifted linear* ($\times$), *and linear* ($*$). *The oversampling parameter* γ *increases from* 1.0 *to* 1.82*—specifically* $\gamma = 1.00, 1.10, 1.22, 1.37, 1.56$, *and* 1.82. *In* (a) *and* (b) *the horizontal dashed line indicates the error of the conventional reconstruction.* (d) *Profile of row* 256 *of the reconstruction: original phantom (dashed line), conventional reconstruction (dot-dashed line), and fast reconstructions using linear* ($*$) *and cubic-MOMS* ($\bullet$) *interpolation.*

most desirable image quality might therefore not be the set of parameters with the lowest RMS error, but instead one that also has an acceptably low peak error. These curves can be used to select an interpolator and operating point of acceptable quality.

Apart from the error measures described above, the accuracy of the algorithms was adjudged by examining 1D profiles of the reconstructions near the edges of regions in the image. The width of the edge—the number of pixel units taken for the reconstruction to cross from 5% to 95% of the edge-transition (assuming underlying bilinear interpolation)—is calculated by choosing a set of points near the edge of each ellipse making up the phantom, calculating the edge width at each such edge pixel, and averaging the edge width over all the pixels. The edge widths are also calculated using different transition bounds—namely, 10% to 90%, and 20% to 80%. As displayed in Table 5.1, in all cases the edge widths decrease as the complexity of the interpolators is increased from linear (with $\gamma = 1.22$) to shifted-linear, quadratic

Table 5.1

The average edge widths of the reconstructions for different types of interpolators, using three different transition bounds to calculate edges.

Transition Bounds	Conv-enti-onal	Hierarchical (varying interpolators)					
		Linear (γ=1.22)	Shifted-linear (γ=1.0)	Quad Schaum (γ=1.22 with Kaiser)	Quad Schaum (γ=1.22 without Kaiser)	Quad MOMS (γ=1.0)	Cubic MOMS (γ=1.0)
5 to 95 %	2.48	3.36	2.83	2.59	2.27	2.21	2.06
10 to 90 %	1.96	2.62	2.25	2.08	1.85	1.82	1.70
20 to 80 %	1.31	1.73	1.51	1.41	1.26	1.25	1.17
Accel.	1	15.1	14.8	9.7	9.7	8.7	6.9

Schaum (with $\gamma = 1.22$), quadratic MOMS, and cubic MOMS. The Kaiser window increases the edge width (as seen in the case of the quadratic Schaum interpolator). The bottom row displays the acceleration factors of the hierarchical methods. Notice that the fast algorithm, when implemented with interpolators more complex than the quadratic Schaum, achieves a smaller edge width than the conventional algorithm.

The profile of a particular edge for a few representative reconstructions is displayed in Figure 5.3(d). In order of decreasing edge width they are the cubic MOMS, the conventional algorithm, and the linear interpolator. Visually, the edge-transitions are clearly comparable.

6. Conclusions.

Though the rotation-based hierarchical backprojection algorithms presented here are similar to Brandt's multilevel inversion (MI) algorithm [11], there are key differences and improvements.

These algorithms, like MI, involve sampling the intermediate images on rectangular grids, rotated with respect to each other. The sampling criterion in the slow direction explained heuristically in [11, equation 4.1] is equivalent to the sampling criterion derived here (the constraint in (3.13)). While in MI this slow-sampling interval is exactly achieved, we restrict ourselves to computationally efficient integer up-sampling factors in the slow direction and, consequently, satisfy or exceed the sampling criterion in [11]. Brandt et al. also mention “doubling”: they interpolate the projection data at double the original sampling rate to get better accuracy. In our algorithms we introduce variable oversampling in a more structured manner which results in greater flexibility to improve accuracy.

Our choice to perform digital image rotations leads to more performance gains. The decomposition of the rotation operator into shears allows for the intermediate rotations in the algorithm to be implemented as a sequence of 1D digital fractional delays. The separable implementation of the rotation operator allows for the use of digital image transformations and filtering of a higher quality than the bilinear interpolation used by Brandt in the MI algorithm. These higher-quality transformations are possible first because longer FIR filters can be used at low cost when used separably. Furthermore, operations involving IIR filters, such as the shifted-linear and higher-order spline-based filters, can be used (while they could not be used in the non-separable case). Fractional shifts can be efficiently implemented using shift-invariant filters. This also allows for the use of FFT-based shifting.

By analyzing the algorithm within the signal processing framework, we are able to elucidate conditions on the sampling and interpolation of the intermediate images, and optimize various parameters in the algorithm. We are also able to derive a reprojection algorithm which may be designed with the existing machinery of signal processing—optimal low-pass filtering and image transformations. Furthermore, this understanding may be used to derive fast hierarchical algorithms for other tomographic methods whose Fourier-domain behavior is well understood.

Both the backprojection and reprojection algorithms can be adapted to arbitrary numbers of view-angles, not just sets of size $2 \cdot 3^L$. While the ternary branching factor (or radix) is a particularly computationally efficient choice, arbitrary, possibly mixed, radixes may be used. Any number of projections may be processed through a hierarchy consisting of mostly ternary nodes. The algorithm is easily adapted to process projections at nonuniform view-angles. The Fourier-domain analysis dictates the optimal intermediate rotation-angles and the up-sampling factors.

The new family of fast backprojection and reprojection algorithms developed in this paper complements and extends the existing fast algorithms. These algorithms have different tradeoffs between computation and accuracy, and different architectures. A decision as to which of these algorithms is the most effective in any given practical application will require careful comparison of optimized implementations on the computing architectures of interest.

This algorithm can be extended to other projection geometries—the fan-beam [18] and cone-beam, for example—in order to find wider practical application.

Appendix. Proofs of propositions and lemmas.

A. Proof of Lemma 2.3.

Proof (by induction on l). The result is trivially true for $l = 1 : \Phi_p(I_{1,m}) = 0$ for all $p \in N_{1,m} = \{m\}$. Assume that the result is true for $l-1$.

Now consider $p \in \mathcal{N}_{l-1,3m-i}$. By (2.7) and Lemma 2.2, we know that $\Phi_p(I_{l,m}) = \Phi_p(I_{l-1,3m-i}) - \delta_{l-1,3m-i}$. By the inductive hypothesis, we know that $\Phi_p(I_{l-1,3m-i}) = -\sum_{i=1}^{l-2} \delta_{i,\mu(p,i)}$. Equation (2.12) implies that $N_{l-1,3m-i} = \{3^{l-2}(3m-i-1) + k : k = 1, 2, \dots, 3^{l-2}\}$. So $p/3^{l-2} \in \{(3m-i-1) + k/3^{l-2} : k = 1, 2, \dots, 3^{l-2}\}$, and consequently $\mu(p, l-1) \triangleq \lceil p/3^{l-2} \rceil = 3m - i$. Thus $\Phi_p(I_{l,m}) = -\sum_{i=1}^{l-2} \delta_{i,\mu(p,i)} - \delta_{l-1,\mu(p,l-1)} = -\sum_{i=1}^{l-1} \delta_{i,\mu(p,i)}$. $\square$

B. Proof of Proposition 2.4. We will need the following property of the ceiling operator $\lceil . \rceil$.

LEMMA B.1. *Let $x_1, x_2 \in (3^{-k}m, 3^{-k}(m+1)) \subset \mathbb{R}$ for some fixed $k, m \in \mathbb{N}$. Then $\lceil x_1 \rceil = \lceil x_2 \rceil$.*

Proof. Clearly, if $k \geq 0$, the set $\mathbb{Z} = \{3^{-k}3^k p : p \in \mathbb{Z}\} \subseteq \{3^{-k}n : n \in \mathbb{Z}\}$. If there exists $m \in \mathbb{Z}$ such that $x_i \in (3^{-k}m, 3^{-k}(m+1))$ for $i = 1, 2$, then $\lceil x_i \rceil \triangleq \arg\min_{n \in \mathbb{Z}} (n \geq x_i) \geq \arg\min_{y \in \{3^{-k}n : n \in \mathbb{Z}\}} (y \geq x_i) = 3^{-k}(m+1) \implies \lceil x_i \rceil = \lceil (3^{-k}(m+1)) \rceil$; i.e., $\lceil x_1 \rceil = \lceil x_2 \rceil$. $\square$

Proof of Proposition 2.4. Consider a particular l, m in the algorithm described by the block diagram in Figure 2.1(b). By (2.11), $I_{l,m} = \sum_{p \in N_{l,m}} \mathcal{K}(-\Phi_p(I_{l,m}))\mathcal{B}_0 q_p$. Since $\tilde{I}_{l,m} = \sum_{p \in N_{l,m}} \mathcal{K}(-\theta_p)\mathcal{B}_0 q_p$, the statement "there exists a $\alpha_{l,m}$ such that $I_{l,m} = \mathcal{K}(\alpha_{l,m})\tilde{I}_{l,m}$" is equivalent to

$$\sum_{p \in N_{l,m}} \mathcal{K}(-\Phi_p(I_{l,m}))\mathcal{B}_0 q_p = \sum_{p \in N_{l,m}} \mathcal{K}(\alpha_{l,m})\mathcal{K}(-\theta_p)\mathcal{B}_0 q_p = \sum_{p \in N_{l,m}} \mathcal{K}(\alpha_{l,m} - \theta_p)\mathcal{B}_0 q_p.$$

However, because this equality is assumed to hold for any $\{q_p\}_{p=1}^{P}$, it must hold term by term, and is equivalent to the statement that for all $p \in N_{l,m}$, there exists a $\alpha_{l,m}$

such that $-\Phi_p(I_{l,m}) = \alpha_{l,m} - \theta_p$, which is equivalent to

$$\Phi_{p1}(I_{l,m}) - \Phi_{p2}(I_{l,m}) = \theta_{p1} - \theta_{p2} \qquad \forall p1, p2 \in N_{l,m}. \tag{B.1}$$

Hence, we need to prove that the assumption of the proposition implies (B.1). Now, from (2.6)–(2.8) and the definition of Φ_p (Definition 2.1), $\hat{f} = \sum_{p=1}^{P} \mathcal{K}(-\Phi_p(I_{L+2,1}))\mathcal{B}_0 q_p$. But, by an assumption of this proposition—that the final image is correct and equal to that of the nonhierarchical algorithm—$\hat{f} = \sum_{p=1}^{P} \mathcal{K}(-\theta_p)\mathcal{B}_0 q_p$. Hence

$$\Phi_p(I_{L+2,1}) = \theta_p. \tag{B.2}$$

The angle $(\Phi_p(I_{l,m}))$ of a projection in any intermediate image is related to the angle $(\Phi_p(I_{L+2,1}))$ of the same projection in the final image as follows. Using Lemma 2.3, we get

$$\begin{aligned}\Phi_p(I_{L+2,1}) &= -\sum_{i=1}^{L+1} \delta_{i,\mu(p,i)} = -\sum_{i=1}^{l-1} \delta_{i,\mu(p,i)} - \sum_{i=l}^{L+1} \delta_{i,\mu(p,i)} \\ &= \Phi_p(I_{l,m}) - \sum_{i=l}^{L+1} \delta_{i,\mu(p,i)} \qquad \forall p \in N_{l,m},\, \forall l \le (L+2),\end{aligned} \tag{B.3}$$

where $\mu(p,i) = \lceil p/3^{i-1} \rceil$. We also know that if $p \in \mathcal{N}_{l,m} \triangleq \{3^{l-1}(m-1) + k : k = 1, 2, \ldots, 3^{l-1}\}$, as defined in (2.12), then $\frac{p}{3^{i-1}} \subset (3^{l-i}(m-1), 3^{l-i}m]$. So by Lemma B.1, we conclude that for any $p1, p2 \in N_{l,m}$ and $i \ge l$ we have $\mu(p1,i) = \lceil p1/3^{i-1} \rceil = \lceil p2/3^{i-1} \rceil = \mu(p2,i)$. Consequently, using (B.3) and then (B.2),

$$\begin{aligned}\Phi_{p1}(I_{l,m}) - \Phi_{p2}(I_{l,m}) &= \Phi_{p1}(I_{L+2,1}) + \sum_{i=l}^{L+1} \delta_{i,\mu(p1,i)} - \left(\Phi_{p2}(I_{L+2,1}) + \sum_{i=l}^{L+1} \delta_{i,\mu(p2,i)}\right) \\ &= \Phi_{p1}(I_{L+2,1}) - \Phi_{p2}(I_{L+2,1}) = \theta_{p1} - \theta_{p2}.\end{aligned}$$

But this is (B.1), which we have already shown to imply the lemma. □

C. Proof of Lemma 3.1.

Proof. Let $B_2(W)$ denote the space of $L_2(R^2)$ functions bandlimited to $W \subset R^2$, and let $H = \{\vec{\omega} \in \mathbb{R}^2 : |\omega_i| \le \frac{\pi}{\Delta_i}\}$. Then, $E^{\psi}_{\vec{\Delta}} f^d \in B_2(AH) \cap B_2(H)$. To see this note that the Sinc-interpolator Ψ zeroes out the DTFT of f^d outside the principal period $[-\pi, \pi]^2$. Hence, upon coordinate transformation by $\mathcal{A}$, the spectral support is transformed by A^T so that $\mathcal{A}E^{\psi}_{\vec{\Delta}} f^d \in B_2(A^T A^{-T} H) = B_2(H)$. Finally, $E^{\psi}_{\vec{\Delta}} \mathcal{A}^{\psi,\vec{\Delta}} f^d = E^{\psi}_{\vec{\Delta}} \mathcal{D}_{\vec{\Delta}} \mathcal{A} E^{\psi}_{\vec{\Delta}} f^d = \mathcal{A}E^{\psi}_{\vec{\Delta}} f^d$, where the first equality follows from (3.3) and the second from the sampling theorem: $E^{\psi}_{\vec{\Delta}} D_{\vec{\Delta}}$ is an identity on $B_2(H)$. □

REFERENCES

[1] F. Andersson, *Fast inversion of the Radon transform using log-polar coordinates and partial back-projections*, SIAM J. Appl. Math., 65 (2005), pp. 818–837.

[2] A. Averbuch, R. Coifman, D. Donoho, M. Elad, and M. Israeli, *Fast and accurate polar Fourier transform*, Appl. Comput. Harmon. Anal, 21 (2006), pp. 145–167.

[3] A. Averbuch, R. R. Coifman, D. L. Donoho, M. Israeli, and J. Walden, *Fast Slant Stack: A Notion of Radon Transform for Data in a Cartesian Grid Which is Rapidly Computable, Algebraically Exact, Geometrically Faithful and Invertible*, technical report, Department of Statistics, Stanford University, Palo Alto, CA, 2001.

[4] S. Basu and Y. Bresler, *An $O(N^2 \log N)$ filtered backprojection reconstruction algorithm for tomography*, IEEE Trans. Image Process., 9 (2000), pp. 1760–1773.

[5] S. K. Basu, *Fast Algorithms For Tomography*, Ph.D. thesis, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2000.

[6] T. Blu, P. Thevanaz, and M. Unser, *Linear interpolation revitalized*, IEEE Trans. Image Process., 13 (2004), pp. 710–719.

[7] T. Blu, P. Thevenaz, and M. Unser, *Moms: Maximal-order interpolation of minimal support*, IEEE Trans. Image Process., 10 (2001), pp. 1069–1080.

[8] A. Boag, Y. Bresler, and E. Michielssen, *A multilevel domain decomposition algorithm for fast $O(N^2 \log N)$ reprojection of tomographic images*, IEEE Trans. Image Process., 9 (2000), pp. 1573–1582.

[9] M. L. Brady, *A fast discrete approximation algorithm for the Radon transform*, SIAM J. Comput., 27 (1998), pp. 107–119.

[10] M. L. Brady and W. Yong, *Fast parallel discrete approximation algorithms for the Radon transform*, in Proceedings of the 4th Annual ACM Symposium on Parallel Algorithms and Architectures, San Diego, CA, 1992, ACM Press, New York, 1992, pp. 91–99.

[11] A. Brandt, J. Mann, M. Brodski, and M. Galun, *A fast and accurate multilevel inversion of the Radon transform*, SIAM J. Appl. Math., 60 (1999), pp. 437–462.

[12] R. E. Crochiere and A. V. Oppenheim, *Analysis of linear digital networks*, Proc. IEEE, 63 (1975), pp. 581–595.

[13] P.-E. Danielsson, *Iterative Techniques for Projection and Back-Projection*, Technical Report LiTH-ISY-R-1997, Department of Electrical Engineering, Linkoping University, Linkoping, Sweden, 1997.

[14] S. R. Deans, *The Radon Transform and Some of Its Applications*, Wiley, New York, 1983.

[15] D. Dudgeon and R. Mersereau, *Multidimensional Digital Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[16] A. Dutt and V. Rokhlin, *Fast Fourier transforms for nonequispaced data*, SIAM J. Sci. Comput., 14 (1993), pp. 1368–1393.

[17] K. K. Fourmont, *Non-equispaced fast Fourier transforms with applications to tomography*, J. Fourier Anal. Appl., 9 (2003), pp. 431–450.

[18] A. K. George and Y. Bresler, *A Fast and Accurate Decimation-in-Angle Hierarchical Fan-Beam Backprojection Algorithm*, in Proceedings of the 6th Annual IEEE International Symposium on Biomedical Imaging, IEEE Press, Piscataway, NJ, 2006, pp. 1188–1191.

[19] A. K. George and Y. Bresler, *Shear-based fast hierarchical backprojection for parallel-beam tomography*, IEEE Trans. Medical Imaging, 26 (2007), pp. 317–334.

[20] M. Grass, T. Kohler, and R. Proksa, *3D cone-beam CT reconstruction for circular trajectories*, Phys. Med. Biol., 45 (2000), pp. 329–347.

[21] L. Greengard and J.-Y. Lee, *Accelerating the nonuniform fast Fourier Transform*, SIAM Rev., 46 (2004), pp. 443–454.

[22] M. Ingerhed, *Fast Backprojection in Computed Tomography*, Ph.D. thesis, Department of Electrical Engineering, Linkoping University, Linkoping, Sweden, 1999.

[23] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1988.

[24] S. Matej, J. A. Fessler, and I. G. Kazantsev, *Iterative tomographic image reconstruction using Fourier-based forward and back-projectors*, IEEE Trans. Image Process., 23 (2004), pp. 401–412.

[25] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM Monogr. Math. Model. Comput. 5, SIAM, Philadelphia, 2001.

[26] S. Nilsson, *Application of Fast Backprojection Techniques for Some Inverse Problems of Integral Geometry*, Ph.D. thesis, Department of Mathematics, Linkoping University, Linkoping, Sweden, 1997.

[27] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1998.

[28] D. I. Potts and G. I. Steidl, *Fourier reconstruction of functions from their nonstandard sampled Radon transform*, J. Fourier Anal. Appl., 8 (2002), pp. 513–534.

[29] P. Thevanaz, T. Blu, and M. Unser, *Interpolation revisited*, IEEE Trans. Medical Imaging, 19 (2000), pp. 739–758.

[30] M. Unser, P. Thevenaz, and L. Yaroslavsky, *Convolution-based interpolation for fast, high-quality rotation of images*, IEEE Trans. Image Processing, 4 (1995), pp. 1371–1381.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 599–618

A PHASE FIELD METHOD FOR JOINT DENOISING, EDGE DETECTION, AND MOTION ESTIMATION IN IMAGE SEQUENCE PROCESSING*

T. PREUSSER†, M. DROSKE‡, C. S. GARBE§, A. TELEA¶, AND M. RUMPF‡

Abstract. The estimation of optical flow fields from image sequences is incorporated in a Mumford–Shah approach for image denoising and edge detection. Possibly noisy image sequences are considered as input and a piecewise smooth image intensity, a piecewise smooth motion field, and a joint discontinuity set are obtained as minimizers of the functional. The method simultaneously detects image edges and motion field discontinuities in a rigorous and robust way. It is able to handle information on motion that is concentrated on edges. Inherent to it is a natural multiscale approximation that is closely related to the phase field approximation for edge detection by Ambrosio and Tortorelli. We present an implementation for two-dimensional image sequences with finite elements in space and time. This leads to three linear systems of equations, which have to be solved in a suitable iterative minimization procedure. Numerical results and different applications underline the robustness of the approach presented.

Key words. image processing, phase field method, Mumford–Shah, optical flow, denoising, edge detection, segmentation, finite element method

AMS subject classifications. 62H20, 62H35, 65U10, 65N30

DOI. 10.1137/060677409

1. Introduction.

The task of motion estimation from image sequences, or computing the visual representation as optical flow, is a fundamental problem in computer vision. For a number of applications, a dense motion or optical flow field is desirable, yielding a representation of the motion of observed objects for each pixel of the image sequence. In low-level image processing, the accurate computation of object motion in scenes is a long-standing problem which has been addressed extensively. In particular, global variational approaches initiated by the work of Horn and Schunck [19] are increasingly popular. Initial problems such as the smoothing of discontinuities or high computational cost have been solved successfully [25, 7, 8]. Motion estimation also yields important indicators for the detection and recognition of the observed objects. While a number of techniques first estimate the optical flow field and segment objects later in a second phase [37], an approach of computing motion as well as segmenting objects at the same time is much more appealing. First advances in this direction were investigated in [33, 27, 28, 9, 23, 30]. In particular, Kornprobst et al. [22, 3, 4] have considered piecewise smooth motion patterns on image sequences characterized by piecewise smooth objects. Their results are phrased rigorously on the space of

*Received by the editors December 11, 2006; accepted for publication (in revised form) August 24, 2007; published electronically December 7, 2007. This work has been funded by the Priority Program 1114 "Mathematical methods for time series analysis and digital image processing" of the German Research Foundation (DFG).

http://www.siam.org/journals/siap/68-3/67740.html

†Center of Complex Systems and Visualization (CeVis), University of Bremen, Germany (tp@mevis.de).

‡Institute for Numerical Simulation (INS), University of Bonn, Germany (droske@ins.uni-bonn.de, martin.rumpf@ins.uni-bonn.de).

§Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany (Christoph.Garbe@iwr.uni-heidelberg.de).

¶Institute for Mathematics and Computer Science, University of Groningen, The Netherlands (a.c.telea@rug.nl).

functions of bounded variation (BV), and they propose suitable approximations for the numerical implementation. Already, in [22] a joint approach for the segmentation of moving objects in front of a still background and the computation of motion velocities has been proposed. For a given intensity function on an image sequence, a total variation (TV) type functional for the motion field—which allows for jumps in the optical flow velocity—is analyzed in [4, 3]. Recently, Papenberg et al. [29] considered another TV regularization of the motion field and optical flow constraints involving higher order gradients.

The idea of combining different image processing tasks into a single model in order to cope with interdependencies has drawn attention in several different fields. In image registration, for instance, a joint discontinuity approach for simultaneous registration, segmentation, and image restoration has been proposed by Droske and Ring [15] and extended in [16] to incorporate phase field approximations. In these approaches, the phase field is used to describe object boundaries, and sharp interfaces of zero width are replaced by diffuse interfaces of finite width ϵ in which the phase field variable continuously changes its value from 0 to 1. This description of object boundaries draws its name from physics, where it is used for modeling solidification of fluids and associated phase boundaries [31, 38]. Kapur, Yezzi, and Zöllei [20] and Unal et al. [35] have combined segmentation and registration by applying geodesic active contours described by level sets in both images. Vemuri et al. have also used a level set technique to exploit a reference segmentation in an atlas [36]. We refer to [14] for further references.

Recently, Keeling and Ring [21] investigated the relation between optimization and optical flow extraction. A first approach which relates optical flow estimation to Mumford–Shah image segmentation was presented by Nesi [26]. Recently, Rathi et al. investigated active contours for joint segmentation and optical flow extraction [32]. Cremers and Soatto [13, 12] presented an approach for joint motion estimation and motion segmentation with one functional. Incorporating results from Bayesian inference, they derived an energy functional, which can be seen as an extension of the well-known Mumford–Shah [24] approach. Their functional involves the length of boundaries separating regions of different motion as well as a “fidelity term” for the optical flow assumption. Brox, Bruhn, and Weickert [7] present a Chan–Vese-type model for piecewise smooth motion extraction. For given fixed image data the decomposition of image sequences into regions of homogeneous motion is encoded in a set of level set functions, and the regularity of the motion fields in these distinct regions is controlled by a TV functional. Our approach is in particular inspired by these investigations.

We combine denoising and edge detection with the estimation of motion. This results in an energy functional, which incorporates fidelity and smoothness terms for both the image sequence and the flow field. Our focus lies in particular on motion information that is concentrated on edges such as in the case of a moving object with sharp edge contours but without shading and texture. To cope with this, we formulate the optical flow equations appropriately in regions away from edges and on the edge set. Moreover, we incorporate an anisotropic enhancement of the flow along the edges of the image in the sense of Nagel and Enkelmann [25]. This effectively allows us to spread motion information from the edge set onto the whole domain of a moving object. The model is implemented using the phase field approximation in the spirit of Ambrosio and Tortorelli's approach [2] for the original Mumford–Shah functional. The identification of edges is phrased in terms of a phase field function; no a priori knowledge of objects is required, as opposed to formulations of explicit contours. *Particular focus is on optical flow constraints which are not only continuously distributed over shaded or textured regions, but also might be concentrated on edges, e.g., in case of moving objects without texture and shading.* In contrast to a level set approach, the built-in multiscale of the phase field model enables a natural cascadic energy relaxation approach and thus an efficient computation. Indeed, no initial guess for the edge set and the motion field will be required. We present here a truly $(d+1)$-dimensional algorithm, considering time as an additional dimension to the d-dimensional image data. This fully demonstrates the conceptual advantages of the joint approach. Nevertheless, a transfer of the method for only two consecutive time frames is possible but not investigated here. The characteristics of our approach are as follows:

- The distinction of smooth motion fields and optical flow discontinuities is directly linked to edge detection, improving the reliability of the motion estimation.
- The denoising and segmentation task will profit from the explicit coupling of the sequence via the assumption of brightness constancy.
- The phase field approximation is expected to converge to a limit problem for vanishing scale parameter, with a strict notion of edges and motion field discontinuities not involving any additional filtering parameter.
- The algorithm is based on an iteration. In each step a set of three relatively simple linear systems have to be solved for the image intensity, the edge description via the phase field, and the motion field, respectively. Only a small number of iterations is required.

This paper is organized as follows: In section 2 Mumford–Shah-type image denoising and edge detection are reviewed, in section 3 we discuss a generalized optical flow equation, and in section 4 the minimization problem is presented. Section 5 shows how to approximate the segmentation in terms of a variational phase field model. Furthermore, we prove existence of solutions of this model and discuss the limit behavior. Section 6 propounds the corresponding Euler–Lagrange equations, which are discretized applying the usual finite element method in section 7. We conclude with the results in section 8. Finally, in the appendix we provide explicit formulas of all matrices and vectors appearing in the implementation to enable readers to reproduce the algorithm.

2. Recalling the Mumford–Shah functional. In their pioneering paper, Mumford and Shah [24] proposed the minimization of the following energy functional:

$$E_{MS}[u,S] = \lambda \int_\Omega (u-u_0)^2 \,\mathrm{d}\mathcal{L} + \frac{\mu}{2}\int_{\Omega\setminus S} |\nabla u|^2 \,\mathrm{d}\mathcal{L} + \nu\mathcal{H}^{d-1}(S), \tag{2.1}$$

where u_0 is the initial image defined on an image domain $\Omega \subset \mathbb{R}^d$ and λ, μ, ν are positive weights. Here, one asks for a piecewise smooth representation u of u_0 and an edge set S, such that u approximates u_0 in the least squares sense, u ought to be smooth apart from the free discontinuity set S, and in addition S should be smooth and thus small with respect to the $(d-1)$-dimensional Hausdorff measure $\mathcal{H}^{d-1}$. Mathematically, this problem has been treated in the space of functions of bounded variation BV, more precisely in the specific subset SBV [1]. In this paper, we will pick up a phase field approximation for the Mumford–Shah functional (2.1) proposed by Ambrosio and Tortorelli [2]. They describe the edge set S by a phase field ζ which is supposed to be small on S and close to 1 apart from edges, i.e., one asks for minimizers of the energy functional

$$E_\epsilon[u,\zeta] = \int_\Omega \lambda(u-u_0)^2 + \frac{\mu}{2}(\zeta^2+k_\epsilon)\,|\nabla u|^2 + \nu\epsilon\,|\nabla\zeta|^2 + \frac{\nu}{4\epsilon}(1-\zeta)^2\,\mathrm{d}\mathcal{L}, \tag{2.2}$$

where ϵ is a scale parameter and $k_\epsilon = o(\epsilon) \ll 1$ a small positive regularizing parameter, which mathematically ensures strict coercivity with respect to u. On edges the weight ζ^2 is expected to vanish. Hence, the second term measures smoothness of u but only away from edges. The last two terms in the integral encode the approximation of the $(d-1)$-dimensional area of the edge set and the strong preference for a phase field value $\zeta \approx 1$ far from edges, respectively. For larger ϵ one obtains coarse, blurred representations of the edge sets and corresponding smoother images u. With decreasing ϵ we successively refine the representation of the edges and include more image details.

3. Generalized optical flow equation. In image sequences we observe different types of motion fields: locally smooth motion visible via variations of object shading and texture in time, or jumps in the motion velocity apparent at edges of objects moving in front of a background. We aim for an identification of corresponding piecewise smooth optical flow fields in piecewise smooth image sequences

$$u : [0,T]\times\Omega \mapsto \mathbb{R}; \quad (t,x) \to u(t,x)$$

for a finite time interval $[0,T]$ and a spatial domain $\Omega \subset \mathbb{R}^d$ with $d = 1,2,3$. In what follows, we assume $\partial\Omega$ to be Lipschitz. The flow fields are allowed to jump on edges in the image sequence. On edges, the derivative Du splits into a singular and a regular part. The regular part is a classical gradient $\nabla_{(t,x)}u$ in space and time, whereas the singular part lives on the singularity set S—the set of edge surfaces in space-time. Time slices of S are the actual image edges at the specific time. We denote by $n_S \in \mathbb{R}^{d+1}$ the normal on S with respect to space-time. The singular part represents the jump of the image intensity on S, i.e., one observes that $D^s u = (u^+ - u^-)n_S$. Here, u^+ and u^- are the upper and lower intensity values on both sides of S, respectively. Now, we suppose that the image sequence u reflects an underlying motion with a piecewise smooth motion velocity v, which is allowed to jump only on S. Thus, S represents object boundaries moving in front of a background, which might as well be in motion. In strict mathematical terms, we suppose that $u, v \in SBV$ (the set of functions of bounded variation and vanishing Cantor part in the gradient) [17, 1]. In this general setting without any smoothness assumption on u and v, we have to ask for a generalized optical flow equation. In fact, away from moving object edges we derive, as usual, from the brightness constancy constraint equation (BCCE) $u(t+s,\, x+s\,v) = $ const on motion trajectories $\{(t+s,\, x+s\,v) \mid s \in [0,T]\}$, that

$$\nabla_{(t,x)}u \cdot w = 0, \tag{3.1}$$

where $w = (1,v)$ is the space-time motion velocity. On edges, the situation is more complex and in general requires prior knowledge. For instance, a white circular disk moving in front of a black background is visually identical to a black mask with a circular hole moving with the same speed on a white background (the aperture problem). Hence, it is ambiguous on which side of the edge $w^\pm$ vanishes and on which side a nontrivial optical flow equation $n_S \cdot w^\pm = 0$ holds. We will not resolve this ambiguity via semantic assumptions. In what follows, we assume instead that locally only one object—in our example either the circle or the mask—is moving on a stationary background. Hence, we rule out that foreground and background are in motion. In other words, our background is that part of the image which is not moving. Then one of the two values of w on both sides of the edge vanishes by assumption, and we can rewrite the optical flow constraint on the edge without identifying foreground or background by

$$n_S \cdot (w^+ + w^-) = 0. \tag{3.2}$$

This in particular includes the case of a sliding motion without any modification of the object overlap, where $n_S \cdot w^+ = n_S \cdot w^- = 0$.

4. Mumford–Shah approach to optical flow. Now, we ask for a simultaneous denoising, segmentation, and flow extraction on image sequences. Hence, we will incorporate the motion field generating an image sequence into a variational method. Let us formulate a corresponding minimization problem in the spirit of the Mumford–Shah model.

Definition 4.1 (Mumford–Shah-type optical flow approach). *Given a noisy initial image sequence $u_0 : D \to \mathbb{R}$ on the space-time domain $D = [0,T] \times \Omega$, we define the energy*

$$\begin{aligned} E_{\mathrm{MSopt}}[u,w,S] = {} & \int_D \frac{\lambda_u}{2}(u-u_0)^2 \,\mathrm{d}\mathcal{L} + \int_{D\setminus S} \frac{\lambda_w}{2}\left(w \cdot \nabla_{(t,x)} u\right)^2 \mathrm{d}\mathcal{L} \\ & + \int_{D\setminus S} \frac{\mu_u}{2}\left|\nabla_{(t,x)} u\right|^2 \mathrm{d}\mathcal{L} + \int_{D\setminus S} \frac{\mu_w}{q}\left|\nabla_{(t,x)} w\right|^q \mathrm{d}L + \nu \mathcal{H}^d(S) \end{aligned} \tag{4.1}$$

for a piecewise smooth image sequence u, and a piecewise smooth motion field $w = (1, v)$ with a joint jump set S. Furthermore, we require the optical flow constraint $n_S \cdot (w^+ + w^-) = 0$ on S from (3.2). *Now, one asks for a minimizer (u, w, S) of the corresponding constraint minimization problem.*

The first and second terms of the energy are fidelity terms with respect to the image intensity and the regular part of the optical flow constraint, respectively. The third and fourth terms encode the smoothness requirement of u and w. Finally, the last term represents the area of the edge surfaces S. The fidelity weights λ_u, λ_w, the regularity weights μ_u, μ_w, and the weight ν controlling the phase field are supposed to be positive and $q \geq 2$. Let us emphasize that, without any guidance from the local time modulation of shading or texture on both sides of an edge, there is still an undecidable ambiguity with respect to foreground and background.

5. Phase field approximation. Similar to the original model for denoising and edge detection (2.1), the above Mumford–Shah approach (4.1) with its explicit dependence on the geometry of the edge set is difficult to implement without any additional strong assumptions either on the image sequence or on the motion field. For a corresponding parametric approach we refer to the recent results by Cremers and Soatto [13, 12]. The level set approach recently presented by Brox, Bruhn, and Weickert [7] does not explicitly encode motion concentrated on edges. We do not aim to impose any additional assumption on the image sequence u and the motion field v and ask for a suitable approximation of the above model. To gain more flexibility and, in addition, to incorporate a simple multiscale into the model, we propose here a phase field formulation (2.2) in the spirit of Ambrosio and Tortorelli [2]. Let us note that in [3] Aubert, Deriche, and Kornprobst already proposed considering this type of phase field approximation for the regularization of the motion field. We introduce an auxiliary variable ζ—the phase field—describing the edge set S. Away from S we aim for $\zeta \approx 1$, and on S the phase field ζ should vanish. As in the original Ambrosio–Tortorelli model, a scale parameter ϵ controls the thickness of the region with small phase field values. We consider the following energy functionals in the Mumford–Shah optical flow model (4.1):

$$E^{\epsilon}_{\mathrm{fid},u}[u] = \int_D \frac{\lambda_u}{2}(u-u_0)^2 \,\mathrm{d}\mathcal{L}, \tag{5.1}$$

$$E^{\epsilon}_{\mathrm{fid},w}[u,w] = \int_D \frac{\lambda_w}{2}\left(w\cdot\nabla_{(t,x)}u\right)^2 \mathrm{d}\mathcal{L}, \tag{5.2}$$

$$E^{\epsilon}_{\mathrm{reg},u}[u,\zeta] = \int_D \frac{\mu_u}{2}(\zeta^2+k_\epsilon)\left|\nabla_{(t,x)}u\right|^2 \mathrm{d}\mathcal{L}, \tag{5.3}$$

$$E^{\epsilon}_{\mathrm{phase}}[\zeta] = \int_D \left(\nu\epsilon\left|\nabla_{(t,x)}\zeta\right|^2 + \frac{\nu}{4\epsilon}(1-\zeta)^2\right)\mathrm{d}\mathcal{L}. \tag{5.4}$$

These energy contributions control the approximation of the initial image u_0 (5.1) and the optical flow constraints (5.2), the regularity of u (5.3), and the shape of the phase field ζ (5.4). Here, as in the original model, $k_\epsilon = o(\epsilon) > 0$ is a "safety" coefficient, which is needed later to establish existence of solutions of our approximate problem. Still missing is a regularity term for the motion field corresponding to the fourth energy term in the Mumford–Shah model (4.1). If we would consider in a straightforward way the integral

$$\tilde{E}^{\epsilon}_{\mathrm{reg},w}[w,\zeta] = \int_D \frac{\mu_w}{2}(\zeta^2+k_\epsilon)\left|\nabla_{(t,x)}w\right|^2 \mathrm{d}\mathcal{L}, \tag{5.5}$$

the motion field will form approximate jumps on S but without any coupling of a concentrated motion constraint on S and the motion field in homogeneous regions on the image sequence. Figure 8.1 clearly outlines this drawback in the case of a circle with constant white image intensity inside moving on a textured background. As an alternative one might want to decouple the scales for image edges and motion edges introducing a second phase field with a much finer scale parameter $\tilde{\epsilon} \ll \epsilon$ for the representation of motion singularities. But this is not very practical, taking into account a suitable discretization on digital images with limited pixel resolution. Here the parameter ϵ is already in the range of the pixel size. Furthermore, in case of finite energy we would obtain motion fields w bounded in $W^{1,2}$, which is not sufficient to ensure compactness of the optical flow integrand in (5.2). Thus, to allow for piecewise smooth motion fields and to enable an extension of motion velocities first concentrated on edges via the variational approach, we consider

$$E^{\epsilon}_{\mathrm{reg},w}[w,\zeta] = \int_D \frac{\mu_w}{q}\left|P_\delta[\zeta]\nabla_{(t,x)}w\right|^q \mathrm{d}\mathcal{L}. \tag{5.6}$$

Here, the following properties are encoded in the operator $P_\delta[\zeta]$:

- Close to the edges, where $\zeta \le \theta^-$ for some θ^- with $0 < \theta^- < 1$, $P_\delta[\zeta]$ should behave like the original edge indicator ζ^2 proposed by Ambrosio and Tortorelli [2].
- Away from the edges, where $\zeta \ge \theta^+$ for $\theta^- < \theta^+ < 1$, $P_\delta[\zeta]$ is expected to be the identity matrix, which enforces an isotropic smoothness modulus for the motion field w.

- In the spirit of the classical approach by Nagel and Enkelmann [25], $P_\delta[\zeta]$ will be an (approximate) projection onto level sets of the phase field function in the intermediate region. These level sets are surfaces approximately parallel to the edge set in space-time. Thus, information on the optical flow is mediated along the edge set, without a coupling across edge surfaces.

An explicit definition for $P_\delta[\zeta]$ fulfilling these properties is

$$P_\delta[\zeta] = \alpha(\zeta^2)\left(\mathbb{1} + k_\epsilon - \beta(\zeta^2)\frac{\nabla_{(t,x)}\zeta}{\left|\nabla_{(t,x)}\zeta\right|_\delta} \otimes \frac{\nabla_{(t,x)}\zeta}{\left|\nabla_{(t,x)}\zeta\right|_\delta}\right),$$

where $|z|_\delta = (|z|^2 + \delta^2)^{\frac{1}{2}}$ represents a regularized normal. Furthermore, $\alpha : \mathbb{R} \to \mathbb{R}_0^+$ and $\beta : \mathbb{R} \to \mathbb{R}_0^+$ are continuous blending functions, with

$$\alpha(s) = \max\left(0, \min\left(1, \frac{s}{\theta^-}\right)\right) + k_\epsilon, \quad \beta(s) = \max\left(0, \min\left(1, 1 - \frac{s}{\theta^+}\right)\right).$$

Concerning algebraic notation, $\nabla_{(t,x)}w(t,x)$ is a $(d+1)^2$ matrix and thus $P_\delta[\zeta]\nabla_{(t,x)}w$ represents the matrix product. We consider the Frobenius norm of matrices, given by $|A| = \sqrt{\operatorname{tr}(A^T A)}$. Suitable choices for the parameters are $\theta^+ = 0.8$ and $\theta^- = 0.0025$. For vanishing ϵ and a corresponding steepening of the slope of u, this operator basically leads to a separated diffusion on both sides of S in the relaxation of the energy.

Let us recall that the energies $E^\epsilon_{\mathrm{reg},u}, E^\epsilon_{\mathrm{phase}}$ and the term $E^\epsilon_{\mathrm{fid},u}$ are identical to those in the original Ambrosio–Tortorelli approach (see above). In addition, we ask for an optical flow field w according to the optical flow constraint encoded in $E^\epsilon_{\mathrm{fid},w}$ (cf. Figure 8.1 for a first test case). At the same time, this term implies a strong coupling of the image intensities along motion trajectories—which turns into a flow-aligned diffusion in the corresponding Euler–Lagrange equations—for the benefit of a more robust denoising and edge detection. Figure 8.2 shows an example where a completely destroyed time step in the image sequence is recovered by this enhanced diffusion along motion trajectories. Due to the regularity energy $E^\epsilon_{\mathrm{reg},w}$ this motion field is isotropically smooth away from the approximate jump set of u, and the smoothness modulus is characterized by a successively stronger anisotropy along level sets of u while approaching the approximate jump set. The energy term $E^\epsilon_{\mathrm{reg},w}$ (5.6) which we consider for the regularization of the motion field is very similar to the corresponding smoothness term in the classical approach by Nagel and Enkelmann [25], where tangential diffusion is steered by the local structure tensor. In the above multiscale approach no additional prefiltering of the image sequence in terms of a structure tensor is required.

The projection operator $P_\delta[\zeta]$ couples the smoothness of the motion field w to the image geometry, which in fact is very beneficial for the purpose of piecewise smooth motion extraction. The reverse coupling, which would try to align tangent spaces of level sets of u to the motion field, is not required and might even be misleading for our actual goal. The optical flow term in the fidelity energy $E^\epsilon_{\mathrm{fid},w}$ already couples image sequence gradients to the motion field in a direct way. Hence, we don't ask for global minimizers of the sum of all energies but formulate the phase field approximation problem as follows.

Definition 5.1 (solution of the phase field model). *Let $u_0 : D \to \mathbb{R}$ be a noisy space-time image, and let $v_\delta \in W^{1,q}(D, \mathbb{R}^d)$ be boundary data for the velocity field. A space-time image $u \in W^{1,2}(D, \mathbb{R})$, a motion field $w = (1, v + v_\delta)$, with $v \in W^{1,q}_0(D, \mathbb{R}^d)$, and a phase field $\zeta \in W^{1,2}(D, \mathbb{R})$ is called a solution of the phase* *field model if u and ζ minimize the restricted energy*

$$E_w[u,\zeta] := E^{\epsilon}_{\mathrm{fid},u}[u] + E^{\epsilon}_{\mathrm{fid},w}[u,w] + E^{\epsilon}_{\mathrm{reg},u}[u,\zeta] + E^{\epsilon}_{\mathrm{phase}}[\zeta] \tag{5.7}$$

for fixed w in $W^{1,2}(D,\mathbb{R}^{d+1})$, and if the motion field w minimizes the global energy

$$E^{\epsilon}_{\mathrm{global}}[u,w,\zeta] = E^{\epsilon}_{\mathrm{fid},u}[u] + E^{\epsilon}_{\mathrm{fid},w}[u,w] + E^{\epsilon}_{\mathrm{reg},u}[u,\zeta] + E^{\epsilon}_{\mathrm{reg},w}[w,\zeta] + E^{\epsilon}_{\mathrm{phase}}[\zeta] \tag{5.8}$$

for fixed $u,\zeta \in W^{1,2}(D,\mathbb{R})$.

In the Mumford–Shah optical flow model (4.1) the edge set S describes the discontinuities of u and w simultaneously. With the splitting introduced in the definition, we obtain a decoupling of the edge sets. Still the flow field w is smoothed along edges of u. But edges in w will not affect the phase field ζ and thus edges of u. Altogether the set of edges of w will be a subset of the edge set of u

REMARK 5.2. *The definition of u and ζ as the minimizer of a restricted functional is not only sound with respect to the applications. Indeed, a simultaneous relaxation of the global energy with respect to all unknowns is theoretically questionable. In fact, $E^{\epsilon}_{\mathrm{reg},w}$ is not convex in ζ, and we cannot expect this energy contribution to be lower semicontinuous on a suitable set of admissible functions. With the above notion of solutions the direct method in the calculus of variations can be applied, and in particular one observes compactness of the sequence of phase fields associated with a minimizing sequence of image sequences and motion fields (cf. the proof below).*

THEOREM 5.3 (existence of solutions). *Suppose that $d+1 < q < \infty$ and $\lambda_u, \lambda_w, \mu_u, \mu_w, \nu, \epsilon > 0$, and let $k_\epsilon > 0$. Then there exists a solution (u,w,ζ) of the phase field problem introduced in Definition* 5.1.

Proof. At first, we rewrite the phase field approach as an energy minimization problem, which later allows us to apply the direct method from the calculus of variations. For fixed w the energy functional $E_w[u,\zeta]$ (5.7) is strictly convex. By the direct method we obtain a unique minimizer. So let us denote by $(u[w],\zeta[w])$ this minimizer in $W^{1,2}(D,\mathbb{R}) \times W^{1,2}(D,\mathbb{R})$ of the quadratic energy functional $E_w[u,\zeta]$ for fixed $u \in W^{1,2}(D,\mathbb{R})$. The minimizing phase field is given as the weak solution of the corresponding Euler–Lagrange equation

$$-\epsilon\Delta\zeta + \frac{1}{4\epsilon}\zeta = f[u,\zeta] := \frac{1}{4\epsilon} - \frac{\mu_u}{2\nu}\left|\nabla_{(t,x)}u\right|^2\zeta. \tag{5.9}$$

Applying the weak maximum principle we observe that $\overline{\zeta} \equiv 1$ is a supersolution and $\underline{\zeta} \equiv 0$ a subsolution. Thus, $\zeta[w]$ is uniformly bounded, i.e., $0 \le \zeta[w] \le 1$.

Given $(u[w],\zeta[w])$ we consider the global energy $E^{\epsilon}_{\mathrm{global}}$ solely as a functional of the motion field $w=(1,v)$:

$$E[w] = E^{\epsilon}_{\mathrm{global}}[u[w],w,\zeta[w]]$$

on the admissible set

$$\mathcal{A} := \{w \mid w = (1, v+v_\delta),\ v \in W^{1,q}_0(D,\mathbb{R}^{d+1})\},$$

and we define $\underline{E} := \inf_{w\in\mathcal{A}} E[w]$. Testing the energy at $u \equiv 0$, $\zeta \equiv 0$, and $w = (1,v_\delta)$ we observe that $\underline{E} \le \frac{\lambda_u}{2}\left|u_0\right|^2_{L^2} + \frac{\mu_w}{q}\left|\nabla_{(t,x)}v_\delta\right|^q_{L^q} < \infty$. Let us consider a minimizing sequence $\left(w^k\right)_{k=1,\dots,\infty}$ in $\mathcal{A}$ with $E[w^k] \to \underline{E}$ for $k \to \infty$. We set $u^k = u[w^k]$ and

$\zeta^k = \zeta[w^k]$ and estimate the energy E_{global} as

$$E_{\mathrm{global}}[u,w,\zeta] \geq \frac{\lambda_u}{4}\left(|u|_{L^2}^2 - 2\,|u_0|_{L^2}^2\right) + \frac{\mu_u k_\epsilon}{2}\left|\nabla_{(t,x)}u\right|_{L^2}^2 + \frac{\mu_w k_\epsilon}{q}\left|\nabla_{(t,x)}w\right|_{L^q}^q + \frac{\nu}{4\epsilon}\left(|\zeta|_{L^2}^2 - 2\mathcal{L}(D)\right) + \nu\epsilon\left|\nabla_{(t,x)}\zeta\right|_{L^2}^2,$$

where $\mathcal{L}(D)$ denotes the Lebesgue measure of D. From this, we deduce that $(u^k)_k$ and $(\zeta^k)_k$ are bounded in $W^{1,2}(D,\mathbb{R})$ and, taking into account the boundary conditions, that $(w^k)_k$ is bounded in $W^{1,q}(D,\mathbb{R})$. Hence, we can extract a weakly converging subsequence again denoted by $(u^k,w^k,\zeta^k)_k$ having the weak limit (u,w,ζ). From the Sobolev embedding theorem and the assumption $q > d+1$ we derive that w^k strongly converges in L^∞. Furthermore, the corresponding sequence $(\zeta^k)_k$ of phase field functions $\zeta^k := \zeta[w^k]$ are weak solutions of $-\epsilon\Delta\zeta^k + \frac{1}{4\epsilon}\zeta^k = f^k$ (cf. (5.9)). From the bounds on ζ^k in L^∞ and on u^k in $W^{1,2}$ we obtain that $f^k = f[u^k,\zeta^k]$ is uniformly bounded in L^1. This observation allows us to apply a compensated compactness result to verify that $\nabla_{(t,x)}\zeta^k$ converges to $\nabla_{(t,x)}\zeta$ a.e. This is proven for the equation $-\Delta\zeta = f$ on the space $W_0^{1,2}$ in [34, Chap. I, Thm. 3.4], but can easily be generalized for equations of type $-\Delta\zeta + \zeta = f$ on $W^{1,2}$. The matrix-valued function $P_\delta[\cdot]$ is continuous and bounded. Hence, we obtain that $P_\delta[\zeta^k] \to P_\delta[\zeta]$ a.e. for $k \to \infty$. For later use, we define the constants $C_u = \sup_{k=1,\dots,\infty}\left|\nabla_{(t,x)}u^k\right|_{L^2}$ and $C_w = \sup_{k=1,\dots,\infty}\max\left\{\left|w^k\right|_{L^\infty}, \left|\nabla_{(t,x)}w^k\right|_{L^q}\right\}$.

Next, we verify that $u = u[w]$ and $\zeta = \zeta[w]$. Indeed, taking into account the lower semicontinuity of E_w and the modulus of continuity with respect to w we can estimate

$$\begin{aligned}
E_w[u,\zeta] &\leq \liminf_{k\to\infty} E_{w^k}[u^k,\zeta^k] \\
&\leq \liminf_{k\to\infty} E_{w^k}[\tilde{u},\tilde{\zeta}] \\
&\leq E_w[\tilde{u},\tilde{\zeta}] + \liminf_{k\to\infty}\left(\left|w^k\cdot\nabla_{(t,x)}\tilde{u}\right|_{L^2}^2 - \left|w\cdot\nabla_{(t,x)}\tilde{u}\right|_{L^2}^2\right) \\
&\leq E_w[\tilde{u},\tilde{\zeta}] + 2C_w\left|\nabla_{(t,x)}\tilde{u}\right|_{L^2}^2 \liminf_{k\to\infty}\left|w - w^k\right|_{L^\infty}
\end{aligned}$$

for any $\tilde{u},\tilde{\zeta} \in W^{1,2}(D,\mathbb{R})$. From the L^∞ convergence of w^k to w, we immediately obtain that $E_w[u,\zeta] \leq E_w[\tilde{u},\tilde{\zeta}]$. Thus, by definition $u = u[w]$ and $\zeta = \zeta[w]$. Based on these preliminaries, we are able to prove weak lower semicontinuity of the energy. For this we assume without loss of generality that

$$E[w^k] \leq \underline{E} + \rho, \quad \left|P_\delta[\zeta^k] - P_\delta[\zeta]\right|_{L^\infty} \leq \rho, \quad \left|w^k - w\right|_{L^\infty} \leq \rho$$

for a fixed and small constant $\rho > 0$. Applying Mazur's lemma we obtain a sequence of convex combinations

$$\left(\sum_{i=1,\dots,k}\lambda_i^k(u^i,w^i,\zeta^i)\right)_k, \quad \text{with} \sum_{i=1,\dots,k}\lambda_i^k = 1, \quad \lambda_i^k \geq 0,$$

converging strongly to (u,w,ζ) in $W^{1,2}(D,\mathbb{R}) \times W^{1,2}(D,\mathbb{R}^{d+1}) \times W^{1,2}(D,\mathbb{R})$. Finally, taking into account convexity properties of the integrands, Fatou's lemma, and the modulus of continuity of $E_{\text{fid},u}$, $E_{\text{fid},w}$, and $E_{\text{reg},w}$ with respect to w and $P_\delta[\zeta]$, respectively, we estimate (using Einstein's summation convention)

$$\begin{aligned}
E[w] &= E_{\text{global}}[u[w], w, \zeta[w]] = E_{\text{global}}[u, w, \zeta]\\
&= \int_D \frac{\lambda_u}{2}\left(\liminf_{k\to\infty} \lambda_i^k u^i - u_0\right)^2 + \frac{\lambda_w}{2}\left|\liminf_{k\to\infty} w \cdot \left(\lambda_i^k \nabla_{(t,x)} u^i\right)\right|^2 \mathrm{d}\mathcal{L}\\
&\qquad + \int_D \frac{\mu_u}{2}(\zeta^2 + k_\epsilon)\left|\liminf_{k\to\infty} \lambda_i^k \nabla_{(t,x)} u^i\right|^2 + \frac{\mu_w}{q}\left|\liminf_{k\to\infty} P_\delta[\zeta]\lambda_i^k \nabla_{(t,x)} w^i\right|^q \mathrm{d}\mathcal{L}\\
&\qquad + \int_D \left(\nu\epsilon\left|\liminf_{k\to\infty} \lambda_i^k \nabla\zeta^i\right|^2 + \frac{\nu}{4\epsilon}\left(1 - \liminf_{k\to\infty} \lambda_i^k \zeta^i\right)^2\right) \mathrm{d}\mathcal{L}\\
&\le \int_D \frac{\lambda_u}{2} \liminf_{k\to\infty} \lambda_i^k \left(\left(u^i - u_0\right)^2 + \frac{\lambda_w}{2}\left|w \cdot \left(\nabla_{(t,x)} u^i\right)\right|^2\right) \mathrm{d}\mathcal{L}\\
&\qquad + \int_D \liminf_{k\to\infty} \lambda_i^k \left(\frac{\mu_u}{2}(\zeta^2 + k_\epsilon)\left|\nabla_{(t,x)} u^i\right|^2 + \frac{\mu_w}{q}\left|P_\delta[\zeta]\nabla_{(t,x)} w^i\right|^q\right) \mathrm{d}\mathcal{L}\\
&\qquad + \int_D \liminf_{k\to\infty} \lambda_i^k \left(\nu\epsilon\left|\nabla_{(t,x)}\zeta^i\right|^2 + \frac{\nu}{4\epsilon}(1 - \zeta^i)^2\right) \mathrm{d}\mathcal{L}\\
&\le \liminf_{k\to\infty} \lambda_i^k E[u^i, w^i] + \frac{\lambda_w}{2} \sup_{i=1,\ldots,\infty} \left(\left|w^i \cdot \nabla_{(t,x)} u^i\right|_{L^2}^2 - \left|w \cdot \nabla_{(t,x)} u^i\right|_{L^2}^2\right)\\
&\qquad + \frac{\mu_w}{q} \sup_{i=1,\ldots,\infty} \left(\left|P_\delta[\zeta^i]\nabla_{(t,x)} w^i\right|_{L^q}^q - \left|P_\delta[\zeta]\nabla_{(t,x)} w^i\right|_{L^q}^q\right)\\
&\le \underline{E} + \lambda_w C_w C_u^2 \sup_{i=1,\ldots,\infty} \left|w^i - w\right|_{L^\infty} + \mu_w C_w^q \sup_{i=1,\ldots,\infty} \left(\left|P_\delta[\zeta^i] - P_\delta[\zeta]\right|_{L^\infty}\right)\\
&\le \underline{E} + \rho + \lambda_w C_w C_u^2 \rho + \mu_w C_w^q \rho.
\end{aligned}$$

This estimate holds for any $\rho \ge 0$. Thus, we obtain $E[w] \le \underline{E}$, which implies that w is a minimizer of the energy E, and hence (u, w, ζ) is a solution of our phase field problem. □

REMARK 5.4. *The above problem formulation is not only sound with respect to the actual modeling, but it will also allow a simple relaxation approach (see below). Indeed, on all tested data sets we obtain convergence in few iterations (*10–15*).*

Applying formal asymptotics, one observes that the phase field approach proposed here indeed converges to the above Mumford–Shah model. For small ϵ we expect a steepening of the gradient u on a stripe of thickness ϵ around the edge set. The phase field ζ will approximate 1 away from a shrinking neighborhood of the edge surface. For $\epsilon \to 0$ we expect to observe convergence of $E_{\text{reg},u}$ and $E_{\text{reg},w}$ to $\int_{D\setminus S} \frac{\mu_u}{2}\left|\nabla_{(t,x)} u\right|^2 + \frac{\mu_w}{q}\left|\nabla_{(t,x)} u\right|^q \mathrm{d}\mathcal{L}$ and of $E^\epsilon_{\text{phase}}$ to $\mathcal{H}^d(S)$. Under these assumptions on the qualitative behavior $\int_D (w \cdot \nabla_{(t,x)} u)^2 \,\mathrm{d}\mathcal{L}$ converges to the second term of E_{MSopt}, whereas on the edge surface one observes a concentration of energy on the jump set and which scales like $O(\epsilon^{-1})$. Thus, we observe that in the limit we reproduce our optical flow constraint $n_S \cdot (w^+ + w^-) = 0$ from the sharp-interface Mumford–Shah approach. A rigorous validation of this limit behavior in terms of Γ-convergence is still open. For results on Γ-convergence for the optical flow problem in the context of TV type models we refer to [4, 22].

6. Variations of the energy and an algorithm. In what follows, we will consider the Euler–Lagrange equations of the above energies. Thus, we need to compute the variations of the energy contributions with respect to the involved unknowns

u, w, ζ. The variation of an energy E in direction ζ with respect to a parameter function z will be denoted by $\langle \delta_z E, \zeta \rangle$. For the ease of implementation we consider the case $q = 2$. Using straightforward differentiation for sufficiently smooth u, w, ζ and initial data u_0 we obtain

$$\begin{aligned}
\langle \delta_u E^\epsilon_{\mathrm{fid},u}[u], \vartheta \rangle &= \int_D \lambda_u (u - u_0)\, \vartheta \, \mathrm{d}\mathcal{L}, \\
\langle \delta_u E^\epsilon_{\mathrm{fid},w}[u, w], \vartheta \rangle &= \int_D \lambda_w (\nabla_{(t,x)} u \cdot w)(\nabla_{(t,x)} \vartheta \cdot w) \, \mathrm{d}\mathcal{L}, \\
\langle \delta_w E^\epsilon_{\mathrm{fid}}[u, w], \psi \rangle &= \int_D \lambda_w (\nabla_{(t,x)} u \cdot w)(\nabla_{(t,x)} u \cdot \psi) \, \mathrm{d}\mathcal{L}, \\
\langle \delta_u E^\epsilon_{\mathrm{reg},u}[u, \zeta], \vartheta \rangle &= \int_D \mu_u (\zeta^2 + k_\epsilon) \nabla_{(t,x)} u \cdot \nabla_{(t,x)} \vartheta \, \mathrm{d}\mathcal{L}, \qquad (6.1)\\
\langle \delta_\zeta E^\epsilon_{\mathrm{reg},u}[u, \zeta], \xi \rangle &= \int_D \mu_u \zeta \left| \nabla_{(t,x)} u \right|^2 \xi \, \mathrm{d}\mathcal{L}, \\
\langle \delta_w E^\epsilon_{\mathrm{reg},w}[w, \zeta], \psi \rangle &= \int_D \mu_w P_\delta[\zeta] \nabla_{(t,x)} w : \nabla_{(t,x)} \psi \, \mathrm{d}\mathcal{L}, \\
\langle \delta_\zeta E^\epsilon_{\mathrm{phase}}[\zeta], \xi \rangle &= \int_D 2\nu\epsilon \nabla_{(t,x)} \zeta \cdot \nabla_{(t,x)} \xi \, \mathrm{d}\mathcal{L} + \int_D \frac{\nu}{2\epsilon} (\zeta - 1) \xi \, \mathrm{d}\mathcal{L}
\end{aligned}$$

for scalar test functions ξ, ϑ and velocity-type test functions ψ with the structure $\psi = (0, \pi)$. Here, we use the notation $A : B := \mathrm{tr}(B^T A)$. Now, summing up the different terms as in (5.7) and integrating by parts, we end up with the system of PDEs

$$-\mathrm{div}_{(t,x)} \left(\frac{\mu_u}{\lambda_u} (\zeta^2 + k_\epsilon) \nabla_{(t,x)} u + \frac{\lambda_w}{\lambda_u} w (\nabla_{(t,x)} u \cdot w) \right) + u = u_0, \tag{6.2}$$

$$-\epsilon \Delta_{(t,x)} \zeta + \left(\frac{1}{4\epsilon} + \frac{\mu_u}{2\nu} \left| \nabla_{(t,x)} u \right|^2 \right) \zeta = \frac{1}{4\epsilon}, \tag{6.3}$$

$$-\frac{\mu_w}{\lambda_w} \mathrm{div}_{(t,x)} \left(P_\delta[\zeta] \nabla_{(t,x)} v \right) + (\nabla_{(t,x)} u \cdot v) \nabla_x u = 0 \tag{6.4}$$

as the Euler–Lagrange equations characterizing the necessary conditions for a solution (u, w, ζ) of the above-stated phase field approach. Let us emphasize that with the full Euler–Lagrange equations, characterizing a global minimizer of the energy would in addition involve variations of $E_{\mathrm{reg},w}$ with respect to ζ. However, as described in section 5, we do not consider this variation, since it would add a coupling of the edges of the flow field to the edges of the image. Thus, the PDE system (6.2)–(6.4) directly corresponds to our notion of solution specified in Definition 5.1.

For Neumann boundary conditions (which we actually consider in the application) the Euler–Lagrange equation in w is not guaranteed to be coercive in $W^{1,q}$. Indeed, the optical flow term $w \cdot \nabla_{(t,x)} u$ represents a pointwise rank-1 condition, and it is not known a priori that "sufficiently many" of these conditions, in the sense of the Lebesgue measure, are assembled in the image while integrating this term. To remedy this degeneracy, we consider a gradient descent of (6.4)

$$\partial_s v - \frac{\mu_w}{\lambda_w} \mathrm{div}_{(t,x)} \left(P_\delta[\zeta] \nabla_{(t,x)} v \right) + (\nabla_{(t,x)} u \cdot w) \nabla_x u = 0. \tag{6.5}$$

Consequently, the matrices resulting from a discretization are well conditioned and the corresponding systems can be solved easily (see section 7).

Inspired by Ambrosio and Tortorelli, we propose the following iterative algorithm for the solution of the phase field problem with $q = 2$:

Step 0. Initialize $u = u_0$, $\zeta \equiv 1$, and $w \equiv (1, 0)$.
Step 1. Solve (6.2) for fixed w, ζ.
Step 2. Solve (6.3) for fixed u, w.
Step 3. Compute one step of the gradient descent (6.5) for fixed u, ζ.
Step 4. Return to Step 1.

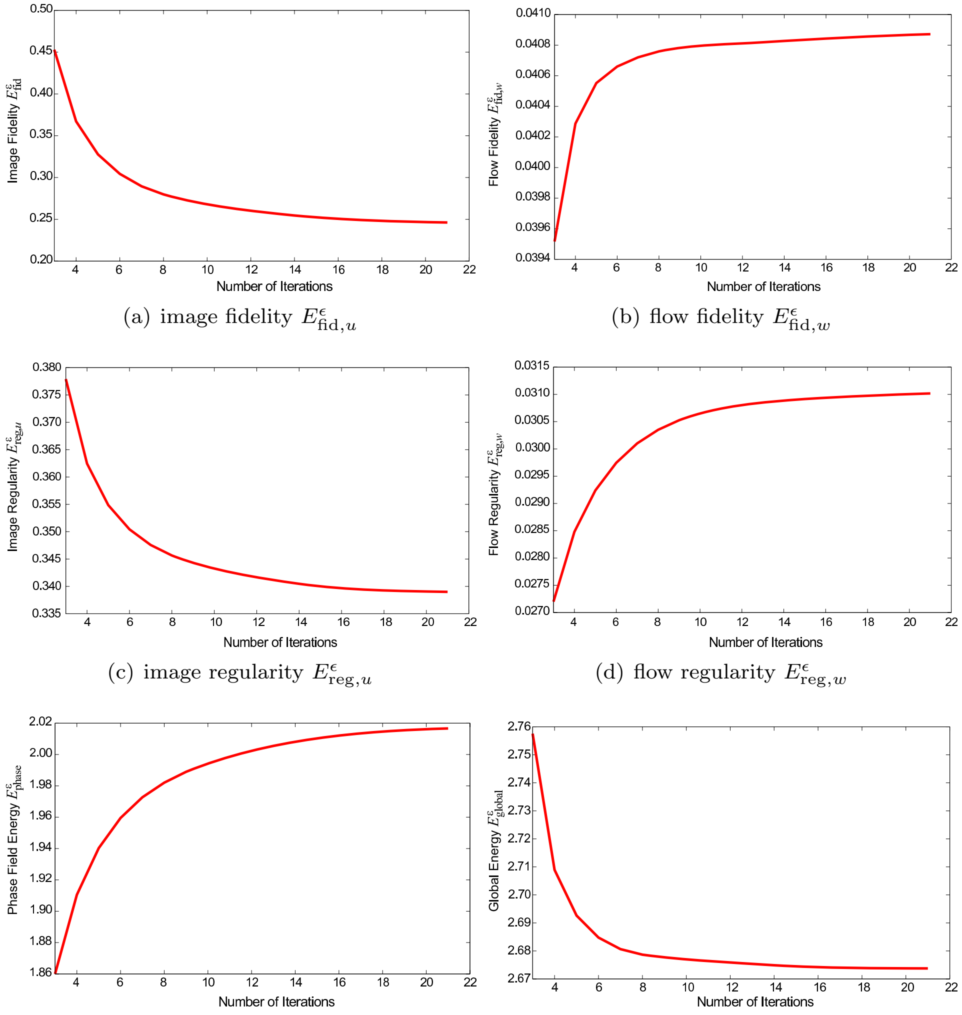
Steps 1 and 2 of the algorithm consist of a consecutive solution of linear PDEs. Let us note that we use a time step control for the gradient descent in Step 3. Alternatively we might iterate first Steps 1 and 2 until convergence, and then in another iteration we would consider the identification of the motion field w. Even though this second variation seems to be closer to our definition of solutions of the phase field problem, the above algorithm converges to the same solution in the applications we have considered. Our algorithm can be seen as a diagonal scheme, where the iteration of Steps 1 and 2 and the gradient descent iteration in Step 3 are intertwined.

7. Finite element discretization. We proceed similarly to the finite element method proposed by Bourdin and Chambolle [5, 6] for the phase field approximation of the Mumford–Shah functional, which is an extension of the approach first presented by Chambolle and Dal Maso [10].

To solve the above system of PDEs we suppose $[0, T] \times \Omega$ to be overlaid by a regular hexahedral grid. In the following, the spatial and temporal grid width are denoted by h and τ, respectively. Hence, image frames are at a distance of τ and pixels of each frame are placed on a regular mesh with grid size h.

On this hexahedral grid we consider the space of piecewise trilinear continuous functions $\mathcal{V}$ and ask for discrete functions $U, Z \in \mathcal{V}$ and $V \in \mathcal{V}^2$, such that discrete and weak counterparts of the Euler–Lagrange equations (6.2), (6.3), and (6.4) are fulfilled. This leads to the solution of systems of linear equations for the vectors of the nodal values of the unknowns U, Z, V. We refer to the appendix for a detailed description of the matrices and the resulting systems of equations. A careful implementation is required to ensure an efficient method. For a time-space volume of K time steps and images of $N \times M$ pixels, the finite element matrices for U and Z have $N\,M\,K\,C$ entries, where $C = 27$ is the number of nonzero entries per row, equal to the number of couplings of a node. The finite element matrix for V has four times more elements, as V is a two-dimensional vector. Data-sets of up to $K = 10$ frames of $N = 500$, $M = 320$ pixels can be treated by standard hardware with less than 1GB memory. The linear systems of equations are solved applying a classical conjugate gradient method. For the pedestrian sequence (Figure 8.5), one such iteration takes 47 seconds on a Pentium IV PC at 1.8 GHz running Linux. The complete method typically converges after 10–15 such iterations. To treat large video sequences, we typically consider a window of $K = 6$ frames, to avoid boundary effects, and then shift this window successively in time.

In Figure 7.1 we have depicted the progression of the various components of the energy $E^\epsilon_{\mathrm{global}}$ for the taxi sequence shown in Figure 8.6. The rapid decay of the global energy in the first steps of the algorithm is clearly visible. While the image fidelity $E^\epsilon_{\mathrm{fid}}$ and its regularity $E^\epsilon_{\mathrm{reg},u}$ decay, the other parts of the energy increase. Obviously this is the case, because we are starting with constant initial values $\zeta = 0$ and $w = (1, 0)$.



(a) image fidelity $E^{\epsilon}_{\mathrm{fid},u}$

(b) flow fidelity $E^{\epsilon}_{\mathrm{fid},w}$

(c) image regularity $E^{\epsilon}_{\mathrm{reg},u}$

(d) flow regularity $E^{\epsilon}_{\mathrm{reg},w}$

(e) phase field energy $E^{\epsilon}_{\mathrm{phase}}$

(f) global energy $E^{\epsilon}_{\mathrm{global}}$

FIG. 7.1. *For the example presented in Figure* 8.6 *(bottom row) we show the progression of the various energy contributions during the solution iteration. The decay of the global energy can be seen in the lower right plot* (f)*.*

8. Results and discussion. We present here several results of the proposed method for two-dimensional image sequences. In the considered examples, the parameter setting $\epsilon = h/4$, $\mu_u = h^{-2}$, $\mu_w = \lambda_u = 1$, $\lambda_w = 10^5 h^{-2}$, and $k_\epsilon = \epsilon$, $\delta = \epsilon$ has proven to give good results. We first consider a simple example of a white disk moving with constant speed $v = (1, 1)$ on a vaguely textured, low-contrast, dark background (Figure 8.1). Let us first consider the top row in Figure 8.1, which corresponds to the energy formulation without the projection component. A limited amount of smoothing results from the regularization energy $E^{\epsilon}_{\mathrm{reg},u}$ (Figure 8.1(a)), which is desirable to ensure robustness in the resulting optical flow term $\nabla_{(t,x)}u \cdot w$ and removes noisy artifacts in real-world videos; see, e.g., Figures 8.4 and 8.5. The phase field clearly captures the moving object's contour. The optical flow is depicted in Figure 8.1(c) by color coding the vector directions as shown by the lower right color wheel. Clearly, the method is able to extract the uniform motion of the disc's boundary, which has a

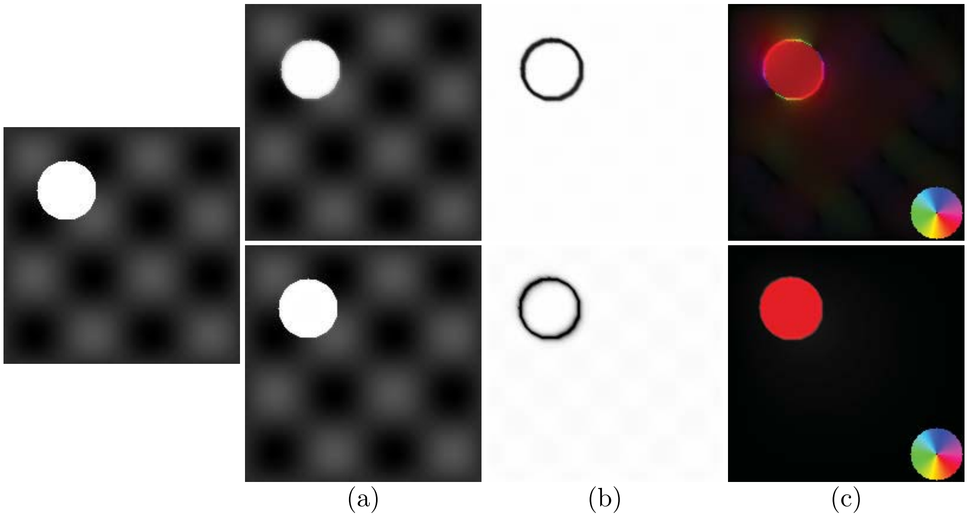


Fig. 8.1. *One frame of the test sequence (left) and corresponding smoothed images* (a)*, phase field* (b)*, optical flow (color coded)* (c)*. Top row: Energy formulation without projection. Bottom row: energy formulation with projection.*

high image contrast. The optical flow information, available only on the motion edges (black in Figure 8.1(b)), is propagated only to a limited extent into the informationless area inside the moving disk. Indeed, we notice that the model with the standard regularity term for w (5.5) is not able to diffuse the optical flow information, concentrated on the motion edges, in order to completely and uniformly fill in the moving circle.

In the bottom row of Figure 8.1, the same example is shown, this time run with the energy formulation including the projection term. We now clearly see a perfect reconstruction of the optical flow (Figure 8.1(c), bottom row) also inside the nontextured moving disc.

In the next example we revisit this simple image sequence of the moving circle. This time we have added noise to the sequence. At the same time we have completely destroyed the information of one frame of the sequence. In Figure 8.2 we show the results for frames 3 and 9–11, where frame 10 has been completely destroyed. From the images we see that the phase field detects the missing circle in the destroyed frame as a temporal edge surface in the sequence. Indeed the ζ drops down to zero in the temporal vicinity of the destroyed frame. This is still visible in the previous and the next time steps, shown in the second and third rows. But it does not hamper the restoration of the correct optical flow field shown in the fourth column. This is due to the anisotropic smoothing of information from the surrounding frames into the destroyed frame. For this example we have chosen $\epsilon = 0.4h$.

Another synthetic example is shown in Figure 8.3. This example is from the publicly available data-set collection at [11]. Here, a textured sphere spins on a textured background (Figure 8.3(a)). Again, the method is able to clearly segment the moving object from the background, even though the object does not change position. We used a phase field parameter $\epsilon = 0.15h$. The extracted optical flow clearly shows the spinning motion (Figure 8.3(d)) and the discontinuous motion field.

A first example on real video data is shown in Figure 8.4. The video shows a table tennis player whose body moves to the right while the hand goes down as he strikes the ball. This motion is well captured in the flow field (Figure 8.4(c)).

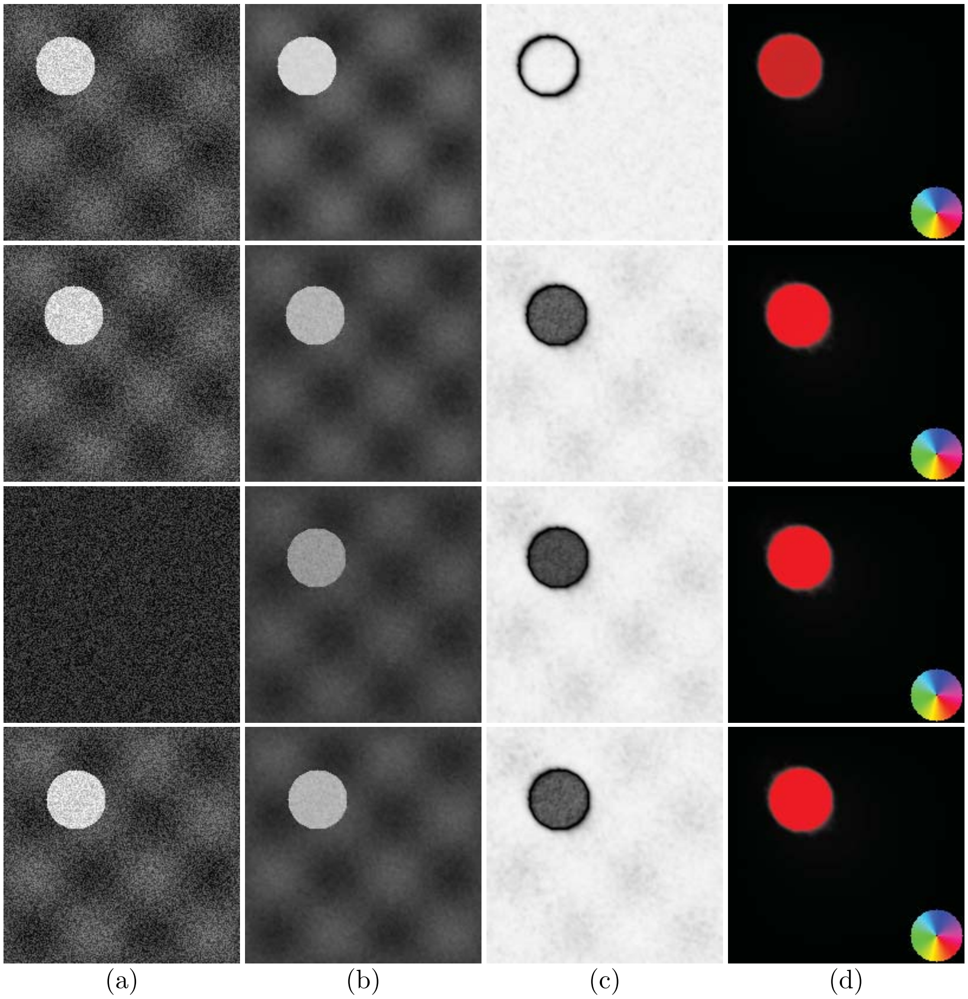


Fig. 8.2. *Noisy test sequence: From top to bottom frames* 3 *and* 9–11 *are shown.* (a) *Original image sequence,* (b) *smoothed images,* (c) *phase field,* (d) *estimated motion (color coded).*

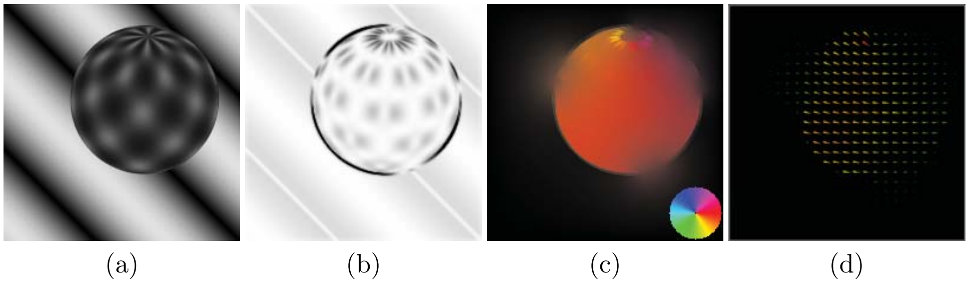
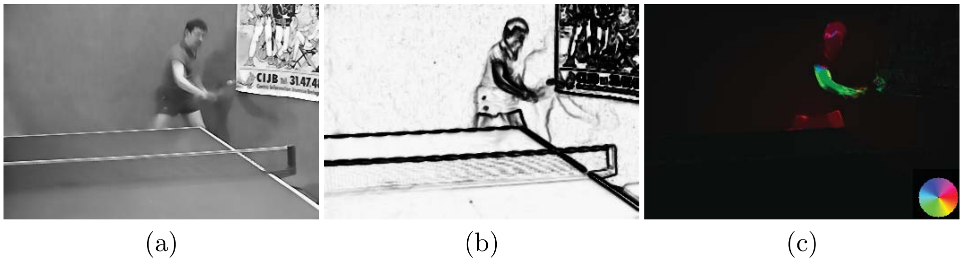


Fig. 8.3. *Rotating sphere: smoothed image* (a), *phase field* (b), *optical flow (color coded)* (c), *optical flow (vector plot, color coded magnitude)* (d).

Furthermore, we consider a complex, higher resolution video sequence, taken under outdoor conditions by a monochrome video camera. The sequence shows a group of walking pedestrians (Figure 8.5 (top)). The human silhouettes are well extracted



(a) (b) (c)

Fig. 8.4. *Table tennis sequence: smoothed image* (a), *phase field* (b), *and optical flow* (c).

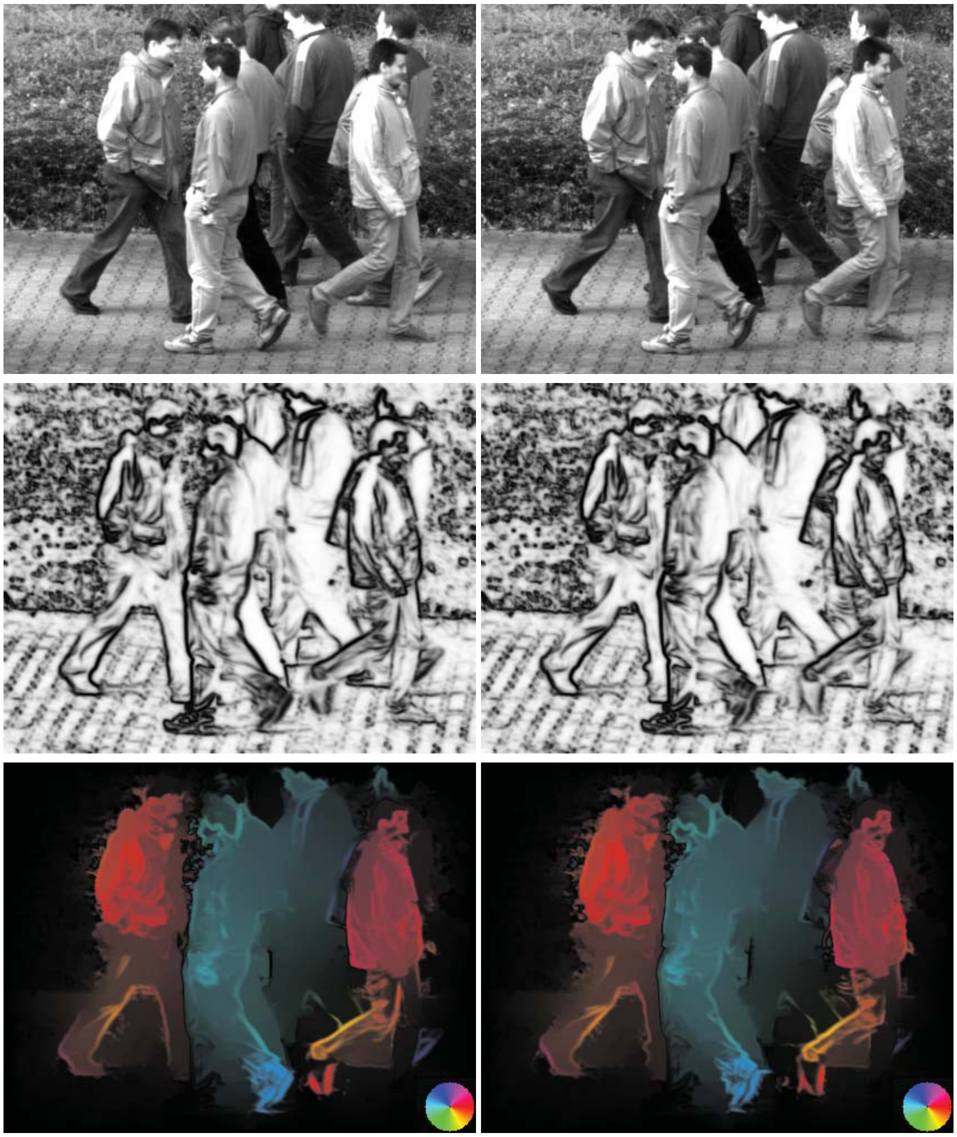


Fig. 8.5. *Pedestrian video: frames from original sequence (top), phase field (middle), and optical flow, color coded (bottom).*

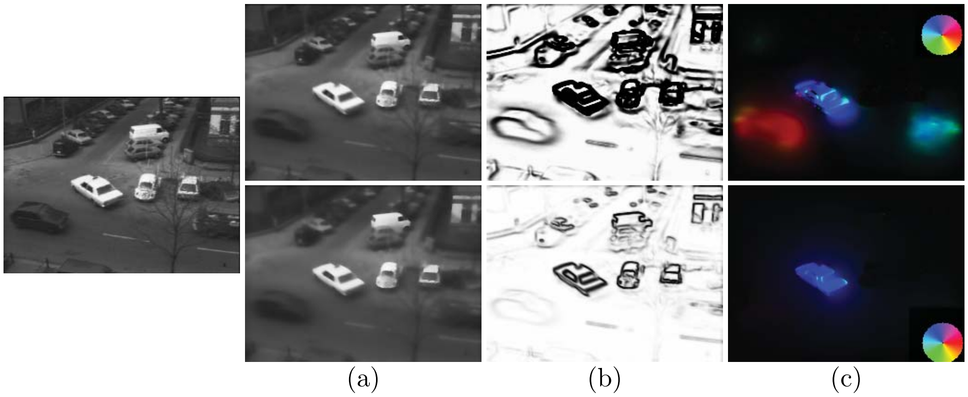


Fig. 8.6. *The taxi sequence. Original image (left). Flow extraction without the projection operator (top row) and with projection (bottom row). Smoothed image* (a)*, phase field* (b)*, and optical flow, color coded* (c)*.*

and captured by the phase field (Figure 8.5 (middle)). We do not show a vector plot of the optical flow, as it is hard to interpret visually at the video sequence resolution of 640 × 480 pixels. However, the color-coded optical flow plot (Figure 8.5 (bottom)) shows how the method is able to extract the moving limbs of the pedestrians. The overall red and blue color corresponds to the walking directions of the pedestrians. The estimated motion is smooth inside the areas of the individual pedestrians and not smeared across the motion boundaries. In addition, the algorithm nicely segments the different moving persons. The cluttered background poses no big problem to the segmentation, nor do the edges of occluding and overlapping pedestrians, who are moving at almost the same speed.

Finally, let us note a limitation of the approach we have presented above: Let us consider the well-known Hamburg taxi video sequence, which is available from [18]. Figure 8.6 shows the taxi sequence processed both with the classical AT energy component (top row) and with our projection operator (bottom row). The progression of the various energy contributions is shown in Figure 7.1. Here we start with $u = 0$, i.e., a black image, and a zero velocity field $v = 0$. In this sequence, cars of differing image contrasts are moving. When the projection operator P_δ in our model is used (bottom row), only the central, high-contrast moving car is captured. When the operator is not used (top row), motion edges corresponding to low-contrast image edges also determine the phase field; hence the other oppositely moving cars in the bottom part of the image are captured as well and the corresponding optical flow is extracted. For all the cars in this example, the motion field is determined largely by the low-contrast shading and not only by high-contrast image edges, as was the case in the synthetic example in Figure 8.1.

Appendix. Algorithmic building blocks. In this appendix we would like to focus on the discrete version of the Euler–Lagrange equations resulting from (5.7). Let us denote by $\{\Psi_i\}_{i=1,\dots,N}$ the usual nodal basis of $\mathcal{V}$ (cf. section 7). The corresponding basis of the vector-valued discrete functions $\Psi \in \mathcal{V}^2$ is given by $\{\Psi_i e_1\}_i \cup \{\Psi_i e_2\}_i$, where $e_{1,2}$ are the standard basis vectors of $\mathbb{R}^2$: $e_1 = (1, 0)$, $e_2 = (0, 1)$. For any discrete function $Q \in \mathcal{V}$ we denote by $\overline{Q}$ the corresponding nodal vector. For discrete vector-valued functions we order the coefficients such that the e_1 coefficients are followed by the e_2 coefficients. Hence, the systems of discrete equations to be solved in the above algorithm are given in matrix vector notion as follows. We ask for solution vectors $\overline{U}, \overline{Z} \in \mathbb{R}^N$ and $\overline{V} \in \mathbb{R}^{2N}$, such that denoting $\overline{W} = (1, \overline{V})$ we have

$$(\mathbf{L}_u[W,\zeta] + \mathbf{M})\overline{U} = R_u, \tag{A.1}$$

$$(\mathbf{L}_\zeta + \mathbf{M}_\zeta[U])\overline{Z} = R_\zeta, \tag{A.2}$$

$$(\mathbf{L}_w[Z] + \mathbf{M}_w[U])\overline{V} = R_w. \tag{A.3}$$

These systems contain the matrices $\mathbf{L}_u[W,Z], \mathbf{L}_\zeta, \mathbf{M}, \mathbf{M}_\zeta[U] \in \mathbb{R}^{N\times N}$, $\mathbf{L}_w[Z], \mathbf{M}_w[U] \in \mathbb{R}^{2N\times 2N}$, $R_u, R_\zeta \in \mathbb{R}^n$, and finally $R_w \in \mathbb{R}^{2N}$, which can easily be derived from the variations of the energy (5.8). We have

$$\begin{aligned}
(\mathbf{L}_u[W,Z])_{ij} &= \int_D \frac{\mu_u}{\lambda_u}(Z^2 + k_\epsilon)\nabla_{(t,x)}\Psi_i \cdot \nabla_{(t,x)}\Psi_j \\
&\qquad + \frac{\lambda_w}{\lambda_u}(\nabla_{(t,x)}\Psi_i \cdot W)(\nabla_{(t,x)}\Psi_j \cdot W)\,\mathrm{d}\mathcal{L}, \\
\mathbf{M}_{ij} &= \int_D \Psi_i\Psi_j\,\mathrm{d}\mathcal{L}, \\
R_u &= \mathbf{M}\overline{\mathcal{I}_h u_0},
\end{aligned}$$

as well as

$$\begin{aligned}
(\mathbf{L}_\zeta)_{ij} &= \epsilon\int_D \nabla_{(t,x)}\Psi_i \cdot \nabla_{(t,x)}\Psi_j\,\mathrm{d}\mathcal{L}, \\
(\mathbf{M}_\zeta[U])_{ij} &= \int_D \left(\frac{\mu_u}{2\nu}\left|\nabla_{(t,x)}U\right|^2 + \frac{1}{4\epsilon}\right)\Psi_i\Psi_j\,\mathrm{d}\mathcal{L}, \\
(R_\zeta)_i &= \frac{1}{4\epsilon}\int_D \Psi_i\,\mathrm{d}\mathcal{L}
\end{aligned}$$

and

$$\begin{aligned}
(\mathbf{L}_w[Z])_{ikjl} &= \int_D \mu_w P_\delta[Z]\nabla_{(t,x)}\Psi_i \cdot \nabla_{(t,x)}\Psi_j\delta_{kl}\,\mathrm{d}\mathcal{L}, \\
(\mathbf{M}_w[U])_{ikjl} &= \int_D \lambda_w\partial_{x_k}U\partial_{x_l}U\,\Psi_i\Psi_j\,\mathrm{d}\mathcal{L}, \\
(R_w)_{ik} &= -\int_D \lambda_w\partial_t U\partial_{x_k}U\Psi_i\,\mathrm{d}\mathcal{L}.
\end{aligned}$$

Here, δ_{kl} is the usual Kronecker symbol, which is 1 if $k = l$ and otherwise 0. Let us remark that the integrands are piecewise polynomials of degree ≤ 2. We use a suitable quadrature rule on the hexahedra, which ensures exact integration.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, New York, 2000.

[2] L. Ambrosio and V. M. Tortorelli, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B (7), 6 (1992), pp. 105–123.

[3] G. Aubert, R. Deriche, and P. Kornprobst, *Computing optical flow via variational techniques*, SIAM J. Appl. Math., 60 (1999), pp. 156–182.

[4] G. Aubert and P. Kornprobst, *A mathematical study of the relaxed optical flow problem in the space $BV(\Omega)$*, SIAM J. Math. Anal., 30 (1999), pp. 1282–1308.

[5] B. Bourdin, *Image segmentation with a finite element method*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 229–244.

[6] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math., 85 (2000), pp. 609–646.

[7] T. Brox, A. Bruhn, and J. Weickert, *Variational motion segmentation with level sets*, in Computer Vision – ECCV 2006, Lecture Notes in Comput. Sci. 3951, H. Bischof, A. Leonardis, and A. Pinz, eds., Springer, Berlin, 2006, pp. 471–483.

[8] A. Bruhn and J. Weickert, *A confidence measure for variational optic flow methods*, in Geometric Properties from Incomplete Data, R. Klette, R. Kozera, L. Noakes, and J. Weickert, eds., Springer, Dordrecht, The Netherlands, 2006, pp. 283–297.

[9] V. Caselles and B. Coll, *Snakes in movement*, SIAM J. Numer. Anal., 33 (1996), pp. 2445–2456.

[10] A. Chambolle and G. Dal Maso, *Discrete approximation of the Mumford-Shah functional in dimension two*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 651–672.

[11] Computer Vision Research Group, *Optical Flow Datasets*, University of Otago, New Zealand, 2005; http://www.cs.otago.ac.nz/research/vision.

[12] D. Cremers, T. Kohlberger, and C. Schnörr, *Nonlinear shape statistics in Mumford-Shah based segmentation*, in Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Lecture Notes in Comput. Sci. 2351, A. Heyden, P. Johansen, M. Nielsen, and G. Sparr, eds., Springer, Berlin, 2002, pp. 93–108.

[13] D. Cremers and S. Soatto, *Motion competition: A variational approach to piecewise parametric motion segmentation*, Internat. J. Comput. Vision, 62 (2005), pp. 249–265.

[14] C. A. Davatzikos, R. N. Bryan, and J. L. Prince, *Image registration based on boundary mapping*, IEEE Trans. Medical Imaging, 15 (1996), pp. 112–115.

[15] M. Droske and W. Ring, *A Mumford–Shah level-set approach for geometric image registration*, SIAM J. Appl. Math., 66 (2006), pp. 2127–2148.

[16] M. Droske, W. Ring, and M. Rumpf, *Mumford-Shah based registration*, Comput. Vis. Sci., to appear.

[17] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[18] *Hamburg Taxi Sequence*, http://i21www.ira.uka.de/image_sequences.

[19] B. Horn and B. Schunk, *Determining optical flow*, Artificial Intelligence, 17 (1981), pp. 185–204.

[20] T. Kapur, L. Yezzi, and L. Zöllei, *A variational framework for joint segmentation and registration*, in the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001), IEEE Computer Society, 2001, pp. 44–51.

[21] S. L. Keeling and W. Ring, *Medical image registration and interpolation by optical flow with maximal rigidity*, J. Math. Imaging Vision, 23 (2005), pp. 47–65.

[22] P. Kornprobst, R. Deriche, and G. Aubert, *Image sequence analysis via partial differential equations*, J. Math. Imaging Vision, 11 (1999), pp. 5–26.

[23] E. Memin and P. Perez, *A multigrid approach for hierarchical motion estimation*, in Proceedings of the International Conference on Computer Vision (ICCV), 1998, pp. 933–938.

[24] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.

[25] H.-H. Nagel and W. Enkelmann, *An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences*, IEEE Trans. Pattern Anal. Mach. Intell., 8 (1986), pp. 565–593.

[26] P. Nesi, *Variational approach to optical flow estimation managing discontinuities*, Image Vision Comput., 11 (1993), pp. 419–439.

[27] J.-M. Odobez and P. Bouthemy, *Robust multiresolution estimation of parametric motion models*, J. Visual Commun. Image Representation, 6 (1995), pp. 348–365.

[28] J.-M. Odobez and P. Bouthemy, *Direct incremental model-based image motion segmentation for video analysis*, Signal Process., 66 (1998), pp. 143–155.

[29] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert, *Highly accurate optic flow computation with theoretically justified warping*, Internat. J. Comput. Vision, 67 (2006), pp. 141–158.

[30] N. Paragios and R. Deriche, *Geodesic active contours and level sets for the detection and tracking of moving objects*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 266–280.

[31] O. Penrose and P. C. Fife, *Thermodynamically consistent models of phase-field type for the kinetics of phase transitions*, Phys. D, 43 (1990), pp. 44–62.

[32] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi, *Particle filtering for geometric active contours with application to tracking moving and deforming objects*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, IEEE Computer Society, 2005, pp. 2–9.

[33] C. Schnörr, *Segmentation of visual motion by minimizing convex non-quadratic functionals*, in Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 1994.

[34] M. Struwe, *Variational Methods: Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*, Springer, Berlin, 2000.

[35] G. Unal, G. Slabaugh, A. Yezzi, and J. Tyan, *Joint Segmentation and Non-Rigid Registration without Shape Priors*, Tech. report scr-04-tr-7495, Siemens Corporate Research, Princeton, NJ, 2004.

[36] B. Vemuri, J. Ye, Y. Chen, and C. Leonard, *Image registration via level-set motion: Applications to atlas-based segmentation*, Medical Image Anal., 7 (2003), pp. 1–20.

[37] J. Y. A. Wang and E. H. Adelson, *Representing moving images with layers*, IEEE Trans. Image Process., 3 (1994), pp. 625–638.

[38] S.-L. Wang, R. F. Sekerka, A. A. Wheeler, B. T. Murray, S. R. Coriell, R. J. Braun, and G. B. McFadden, *Thermodynamically-consistent phase-field models for solidification*, Phys. D, 69 (1993), pp. 189–200.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 619–628

HOMOGENEOUS BRANCHED-CHAIN EXPLOSIONS[*]

LUIS L. BONILLA[†], MANUEL CARRETERO[†], AND J. B. KELLER[‡]

Abstract. A model of homogeneous explosions with competing branching and recombination processes due to Kapila is analyzed by singular perturbation methods. In this model, the concentration of radicals is very low during a long induction period that ends with a rapid radical-growth stage in which all the reactants are consumed as the radicals reach their peak concentrations. The sudden jump in radical concentration is then followed by a long period of chain termination. Based on an exact relation between the fuel concentration and a slowly varying combination of fuel and radicals, we find a composite of two matched asymptotic expansions providing very good agreement with the numerical solution. This approximation is compared to another composite obtained by the method of multiple self-adjusting scales. Both approximations seem to be similarly accurate provided the induction time is calculated beyond leading order.

Key words. jump phenomena, chain-branched homogeneous explosions, induction time, matched asymptotic expansions, multiple scale methods

AMS subject classifications. 34E15, 80A32

DOI. 10.1137/070692911

1. Introduction.

Jump phenomena are characterized by large amplitude dynamic responses to small amplitude disturbances and typically involve different time scales: the system may evolve slowly during long time intervals which are separated by fast transition layers during which the system changes abruptly [6, 13]. Polymer flow in a capillary [7], jump-to-contact instabilities in ultra-thin film lubrication [10], vibration in mechanical systems [11], instabilities of the current in semiconductors [2, 12], saltatory motion of wave fronts in discrete systems [4] and branched-chain explosions [8, 9], and overdriven detonations [14] in combustion theory are examples of jump phenomena. Their multilayer structure makes it difficult to find a uniform description of jump phenomena. In [3], we introduced a method of self-adjusting time scales to describe homogeneous branched-chain explosions, whose main ingredient is a fast time scale which is a nonlinear function of one of the system variables. This method is not standard in that it requires two different solvability conditions depending on whether time is smaller or larger than the very large induction time. An approximate solution valid for all times was obtained by patching two different approximations at the induction time.

In this paper, we introduce an alternative method (based on an exact relation between the fuel concentration and a slowly varying combination of fuel and radicals) to approximate the solutions of the explosion problem before and after the induction time, and match them to find a uniform approximation. We also find an exact expression for the induction time in terms of an integral whose leading order approximation

[*] Received by the editors May 24, 2007; accepted for publication (in revised form) September 5, 2007; published electronically December 7, 2007. This research was supported by MEC grant MAT2005-05730-C02-01 and by the Autonomous Region of Madrid under grant S-0505/ENE/0229 (COMLIMAMS).

http://www.siam.org/journals/siap/68-3/69291.html

[†] Gregorio Millán Institute for Fluid Dynamics, Nanoscience and Industrial Mathematics, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain (bonilla@ing.uc3m.es, manili@math.uc3m.es).

[‡] Department of Mathematics, Stanford University, Stanford, CA 94305 (keller@math.stanford.edu).

coincides with the induction time provided by the inner expansion. Comparison between our approximate solution and the numerical solution of the model equations shows discrepancies of order ϵ that can be tracked to the fact that the leading order expression for the induction time is not sufficiently accurate. Much better agreement between our theory and the numerical solution of the model is obtained using the numerically calculated integral expression for the induction time. We have also found a matching procedure to improve the method of self-adjusting scales. The resulting composite solution yields a similarly accurate approximation to the numerical solution of the model, provided the induction time is calculated to first order. While the simplicity of the boundary layer method found in this paper makes it preferable, the method of self-adjusting scales may be more widely applicable, in particular when more complex chemical schemes are used.

2. Model problem and solution. A straightforward manipulation of the kinetic rate equations for a homogeneous branched-chain explosion at constant pressure, similar to that presented in [9], leads to the following dimensionless problem [3]:

$$\frac{dx}{dt} = \exp\left[\frac{\beta\theta}{1+\theta}\right] x\, f - \epsilon\, x, \tag{2.1}$$

$$\frac{df}{dt} = -\exp\left[\frac{\beta\theta}{1+\theta}\right] x\, f, \tag{2.2}$$

$$\frac{d\theta}{dt} = q\,\epsilon\, x, \tag{2.3}$$

to be solved with the initial conditions

$$x(0) = \nu, \quad f(0) = 1, \quad \theta(0) = 0. \tag{2.4}$$

Equations (2.1)–(2.3) correspond to a two-step reaction scheme without an initiation stage in which a small amount of intermediate radical X (chain carrier) is present from the beginning. Production of the radical X is enhanced through the chain-branching process F + X → 2X that uses up the fuel F. This process is terminated when the radical X is fully consumed through the production of a product P in the radical recombination reaction X → P. All of the heat is generated through the termination step. The chain-branching rate is assumed to have Arrhenius form with a constant prefactor and constant activation temperature. The termination rate depends only weakly on temperature and we assume that it is constant. In (2.1)–(2.3), $f(t)$ and $x(t)$ are the normalized concentration of fuel and of radical, respectively, and θ is the temperature at time t. The nondimensional parameters q, β, ν, and ϵ represent a heat release, the activation energy of the branching reaction, the initial radical concentration, and a chemical-time ratio, respectively. In realistic applications $q = O(1)$, $\beta = O(1)$, and $0 < \nu \ll \epsilon \ll 1$. See [3] for the specific definitions of the nondimensional variables and parameters we use in terms of dimensional variables, rate constants, heat release, and initial concentrations.

Linear combinations of (2.1)–(2.3) lead to

$$\frac{d}{dt}[\theta + q(x+f)] = 0,$$

an equation that can be readily integrated using the condition (2.4) to give

$$x = 1 + \nu - f - q^{-1}\theta. \tag{2.5}$$

This expression, which replaces (2.3) in the following analysis, reveals in particular that θ reaches a maximum value $\theta = q(1+\nu)$ as radicals and fuel are depleted. On the other hand by adding (2.1) and (2.2), it is easy to see that the sum $x+f$ that appears in (2.5) varies only slowly in time, and would be conserved in the absence of radical recombination, i.e., if $\epsilon = 0$. The use of combined variables of this type is always convenient in combustion problems with multiple chemical scales. For instance, by introducing in this case the alternative variable

$$y = x + f, \tag{2.6}$$

we can isolate the effect of radical branching. The problem reduces then to that of integrating

$$\frac{df}{dt} = -(y - f)\, f\, e^{a(y)}\,, \tag{2.7}$$

$$\frac{dy}{dt} = -\epsilon\,(y - f), \tag{2.8}$$

$$a(y) = \frac{\beta q\,(1 + \nu - y)}{1 + q\,(1 + \nu - y)}\,, \tag{2.9}$$

with initial conditions

$$f(0) = 1, \quad y(0) = 1 + \nu. \tag{2.10}$$

Thus x and θ are given by

$$x = y - f, \tag{2.11}$$

$$\theta = q\,(1 + \nu - y) \tag{2.12}$$

as follows from (2.5) and (2.6).

Equations (2.7) and (2.8) yield

$$\frac{df}{dy} = \frac{f}{\epsilon}\, e^{a(y)}\,, \tag{2.13}$$

which, together with (2.10), lead to the relation

$$f(y) = \exp\left(-\frac{1}{\epsilon}\int_y^{1+\nu} e^{a(s)}\, ds\right)\,. \tag{2.14}$$

By using (2.14) for $f(y)$ in (2.8), we obtain the following integro-differential equation for $y(t)$:

$$\frac{dy}{dt} = -\epsilon\,[y - f(y)] = -\epsilon\left[y - \exp\left(-\frac{1}{\epsilon}\int_y^{1+\nu} e^{a(s)} ds\right)\right]. \tag{2.15}$$

The solution of (2.15) with $y(0) = 1 + \nu$ is given by

$$\epsilon t = \int_y^{1+\nu} \frac{ds}{s - f(s)} = \int_y^{1+\nu} \frac{ds}{s - \exp\left(-\frac{1}{\epsilon}\int_s^{1+\nu} e^{a(r)}\, dr\right)}. \tag{2.16}$$

The solution $y(t)$ can be obtained by evaluating the integral in (2.16) numerically. However, the asymptotic forms of y, both for t large, and for ϵ small and t finite, can be determined analytically.

3. Asymptotic form of $y(t)$ for t large. The integrand in (2.16) becomes infinite, and the integral diverges, when the denominator of the integrand vanishes. This occurs at y_∞, determined by

$$y_\infty = \exp\left(-\frac{1}{\epsilon}\int_{y_\infty}^{1+\nu} e^{a(s)}\,ds\right). \tag{3.1}$$

Thus $t \to \infty$ as $y \to y_\infty$. Note that y_∞ is the exponentially small amount of fuel that remains unused after the termination step causes the radical concentration to become zero. At this point, x, f, and θ in (2.1)–(2.3) have reached their final stationary values, $x = 0$, $f = y_\infty$, and $\theta = q\,(1+\nu-y_\infty)$. To obtain $y(t)$ for t large we note that the denominator in (2.16) can be written as

$$y - f(y) = [1 - f'(y_\infty)]\,(y - y_\infty) + O[(y-y_\infty)^2]. \tag{3.2}$$

We add and subtract from the integrand in (2.16) the reciprocal of the first term on the right-hand side of (3.2). Thus we obtain

$$\epsilon t = \int_y^{1+\nu}\left[\frac{1}{s-f(s)} - \frac{(s-y_\infty)^{-1}}{1-f'(y_\infty)}\right]ds + \int_y^{1+\nu}\frac{(s-y_\infty)^{-1}}{1-f'(y_\infty)}\,ds. \tag{3.3}$$

The second integral in (3.3) can be evaluated explicitly. The first integral has a finite limit as $y \to y_\infty$, which we denote as ϵt_o. Then for t large, (3.3) yields

$$\epsilon t \sim \epsilon t_o + [1 - f'(y_\infty)]^{-1}\ln\left(\frac{1+\nu-y_\infty}{y-y_\infty}\right), \quad t \gg 1. \tag{3.4}$$

Here t_o is defined by

$$\epsilon t_o = \int_{y_\infty}^{1+\nu}\left[\frac{1}{s-f(s)} - \frac{(s-y_\infty)^{-1}}{1-f'(y_\infty)}\right]ds. \tag{3.5}$$

Upon solving (3.4) for y, we get

$$y(t) \sim y_\infty + (1+\nu-y_\infty)\,e^{-\epsilon\,[1-f'(y_\infty)](t-t_o)}, \quad t \gg 1. \tag{3.6}$$

Thus for $t > t_o$, y decays exponentially to the asymptotic value y_∞.

When ϵ is small, (3.1) shows that y_∞ is exponentially small. Then $f'(y_\infty)$ is also exponentially small, and (3.1) becomes simply

$$y(t) \sim (1+\nu)\,e^{-\epsilon\,(t-t_o)}, \quad t \gg 1, \quad \epsilon \ll 1. \tag{3.7}$$

The integral (3.5) for t_o can be evaluated for ϵ small. The result is

$$t_o \sim t_{o,0} = \frac{\ln\nu^{-1}}{1+\nu}, \quad \epsilon \ll 1. \tag{3.8}$$

4. Asymptotic form of $y(t)$ for ϵ small. For ϵ small and $t = O(1)$, we write

$$y(t) = y_0(t) + \epsilon y_1(t) + O(\epsilon^2). \tag{4.1}$$

We use (4.1) in (2.8) and in the initial condition $y(0) = 1+\nu$. Then we equate coefficients of ϵ^0 to get

$$\frac{dy_0}{dt} = 0, \quad y_0(0) = 1+\nu. \tag{4.2}$$

Solving (4.2) gives

$$y_0 = 1 + \nu. \tag{4.3}$$

Now we use (4.1) and (4.3) in (2.8) and in the initial condition with $f(y)$ given by (2.15). From the coefficients of ϵ, we get

$$\frac{dy_1}{dt} = -(1+\nu) + e^{y_1}, \quad y_1(0) = 0. \tag{4.4}$$

Separating variables in (4.4) and integrating them yields

$$t = -\int_0^{y_1} \frac{ds}{1+\nu-e^s} = -\frac{y_1}{1+\nu} + \frac{\ln(1+\nu-e^{y_1})}{1+\nu} - \frac{\ln\nu}{1+\nu}. \tag{4.5}$$

Solving (4.5) for y_1 and exponentiating both sides of the resulting equation yields

$$e^{y_1} = e^{-(1+\nu)t}\nu^{-1}(1+\nu-e^{y_1}). \tag{4.6}$$

Now we solve (4.6) for e^{y_1} and take the logarithm of the result to get

$$y_1 = \ln\left[\frac{1+\nu}{1+\nu e^{(1+\nu)t}}\right]. \tag{4.7}$$

Finally (4.1), (4.3), and (4.7) yield

$$y(t) = 1+\nu+\epsilon\ln\left[\frac{1+\nu}{1+\nu e^{(1+\nu)t}}\right] + O(\epsilon^2). \tag{4.8}$$

5. Matching and the composite expansion. For $t \gg 1$, the inner expansion (4.8) becomes, when (3.8) is used,

$$y(t) \sim 1+\nu-\epsilon\,[(1+\nu)\,(t-t_o)], \quad \epsilon \ll 1, \quad t \gg 1. \tag{5.1}$$

On the other hand, for $\epsilon\,(t-t_o) \ll 1$, the outer expansion (3.7) becomes

$$y(t) \sim (1+\nu) - \epsilon\,(1+\nu)\,(t-t_o), \quad 0 < \epsilon, \quad 0 \le (t-t_o) \ll 1. \tag{5.2}$$

These two expansions match.

A composite expansion can be formed by adding together the inner expansion (4.8) and the outer expansion (3.7), and subtracting the common part, given by the right-hand side of (5.1) or (5.2). However, (5.2) holds only for $\epsilon(t-t_o)$ positive. Therefore we multiply the outer expansion minus its inner form (5.2) by the Heaviside function $H(t-t_o)$. Thus the composite expansion for $\epsilon \ll 1$ is

$$\begin{aligned} y_c(t) \sim{}& 1+\nu-\epsilon\,\ln\left(\frac{1+e^{(1+\nu)\,(t-t_o)}}{1+\nu}\right) \\ &+ H(t-t_o)\left[(1+\nu)\,e^{-\epsilon(t-t_o)} - \{1+\nu-\epsilon\,(1+\nu)\,(t-t_o)\}\right], \end{aligned} \tag{5.3}$$

where t_o is given by (3.8). By inserting this approximate y_c in (2.14), (2.11), and (2.12), we find the approximations for f, x, and θ, respectively:

$$f_c(t) = \exp\left(-\frac{1}{\epsilon}\int_{y_c(t)}^{1+\nu} e^{a(s)}\,ds\right), \tag{5.4}$$

$$x_c(t) = y_c(t) - f_c(t), \tag{5.5}$$

$$\theta_c(t) = q\,[1+\nu-y_c(t)]. \tag{5.6}$$

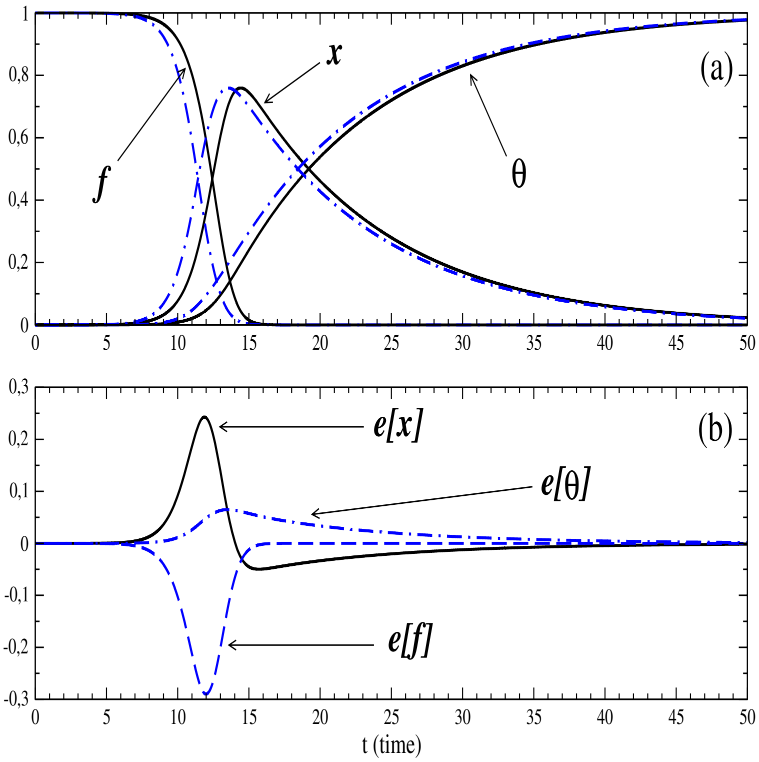


FIG. 1. (a) *Time evolution of* x, f, *and* θ *for* $\epsilon = 0.1$, $\nu = 10^{-5}$, $\beta = 5$, $q = 1$ *obtained by numerical integration of* (2.1) *to* (2.4) *(solid lines), and by using* f_c, x_c, *and* θ_c *in* (5.4), (5.5), *and* (5.6) *(dashed lines).* (b) *Differences* $e[x]$, $e[f]$, *and* $e[\theta]$ *between the approximations* (5.4), (5.5) *and* (5.6), *and the numerical solutions of the model for the same parameter values as in* (a).

The results (5.3) to (5.6) with the approximate induction time (3.8) are compared to the numerical solution of the complete problem in Figure 1. Note that the errors $e[x]$ and $e[f]$ are approximately equal in magnitude: $e[x]+e[f] \approx 0$. Our approximate solutions capture rather well the behavior of the solution of the model equations. However, we observe appreciable differences that are essentially due to the fact that the induction time (3.8) is not a good approximation to the real value. Better agreement with the numerical solution of the model is obtained if an improved induction time is calculated by retaining $O(\epsilon)$ terms in $t_o(\epsilon)$ as defined by (3.5):

$$t_o(\epsilon) = t_{o,0} + \epsilon\, t_{o,1} + O(\epsilon^2), \tag{5.7}$$

$$t_{o,1} = \frac{\ln(\nu^{-1}+1)}{(1+\nu)^2} - \frac{\beta q(1+\nu)-1}{(1+\nu)^2} \int_0^\infty \frac{\sigma\, e^{-\sigma}}{1+\nu-e^{-\sigma}}\, d\sigma, \tag{5.8}$$

where $t_{o,0}$ is given by (3.8). For the parameters used in Figure 1, $t_{o,0} = 11.51$, whereas (5.7) yields $t_o(\epsilon) = 12.01 + O(\epsilon^2)$. We have found that the differences $e[x]$ and $e[f]$ from the numerical solution are still larger than $\epsilon = 0.1$. This means that the induction time is not well resolved even if we keep terms of order ϵ. Therefore we calculate numerically the integral (3.5) and use the resulting value of $t_o \approx 12.31$ in (5.3)–(5.6). The much more satisfactory result is depicted in Figure 2, and we observe that the differences $e[x]$ and $e[f]$ seem now to be of order ϵ^2.

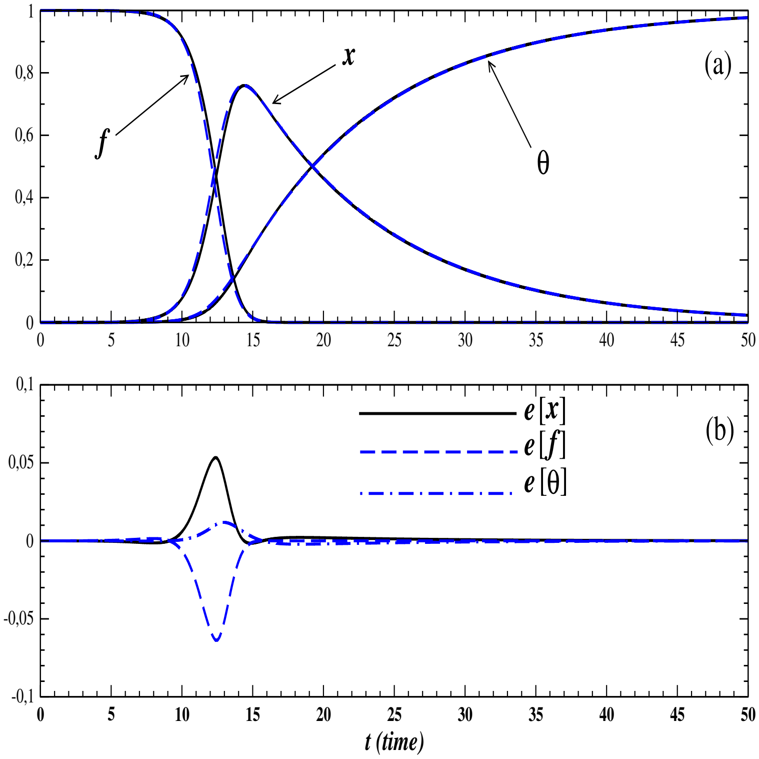


FIG. 2. (a) *Time evolution of x, f, and θ for the same parameter values as in Figure* 1. *The solid lines are the numerical solutions of the model whereas the dashed lines are given by* (5.3)–(5.6) *with t_o given by numerical evaluation of* (3.5). (b) *Differences $e[x]$, $e[f]$, and $e[\theta]$ of the approximations in* (a) *from the numerical solution of the model.*

6. Comparison with the method of self-adjusting time scales. To leading order, the method of self-adjusting scales gave approximated x and f that were somewhat better than (5.3)–(5.5), as a comparison of Figure 1(a) to Figure 3(a) of [3] (both figures have the same parameter values) readily shows. This method used patching of two different asymptotic expansions at the induction time, and this patching implies that the temperature given by the method of self-adjusting scales is a worse approximation than (5.6). In the appendix, we indicate how to improve the method by matching two terms of the expansion valid before the induction time to one term of the expansion for later times. As with the boundary layer method, the quality of the composite expansion given by self-adjusting scales depends on the accuracy with which we calculate the induction time. Inserting the leading order approximation to the induction time, we observe differences of order ϵ between the composite and the numerical solution of the model, similar to those in Figure 1. Calculating the induction time with the method of self-adjusting scales to order ϵ yields a composite expansion whose differences from the numerical solution for x, f, and θ are of order ϵ^2, as shown in Figure 3. Thus, the difference between the composite (5.3) with induction time given by the numerical evaluation of (3.5), and the composite (A.4)–(A.6) (with induction time including terms of order ϵ) given by the method of self-adjusting scales is quite small: see Figures 2(b) and 3(b). To order ϵ, both methods seem to provide similarly accurate approximations although the greater simplicity of the boundary layer method makes it preferable.

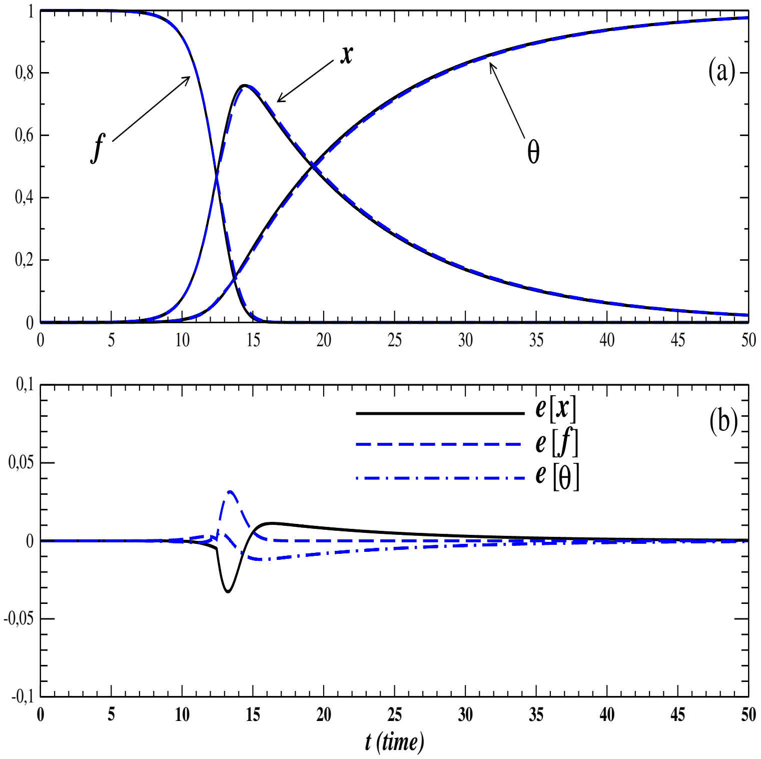


FIG. 3. *Time evolution of x, f, and θ for the same parameter values as in Figure* 1. (a) *Comparison of* (A.3)–(A.6) *(dashed line) to the numerical solution of the model (solid line).* (b) *Differences of the approximations in* (a) *from the numerical solution of the model.*

In more complex models of homogeneous chain-branched explosions, there is an additional induction stage in which radicals are generated by chemical reactions (whose rate constants may contain different Arrhenius factors than those in the branching stage), more than one radical may be acting, etc. Examples are the Blythe, Kapila and Short three-stage branched-chain explosions [1] and the Del Alamo and Williams calculation of ignition times of branched-chain explosions [5]. To extend the ideas in this paper to those more complicated schemes remains a challenge for the future.

Appendix. Matching in the method of self-adjusting scales. We can correct the effects of patching in the method of self-adjusting scales described in [3] by using ideas similar to those used in section 5. For $\tau < \tau_o$ ($\tau = \epsilon t$), the two-term approximation of y before the induction time is

$$y = 1 + \nu - \epsilon \ln\left(\frac{1 + e^{\eta - \eta_o}}{1 + \nu}\right) + o(\epsilon), \tag{A.1}$$

as obtained from (B.24) in [3]. If we let $(\eta - \eta_o) \to +\infty$ in this expression, then we obtain

$$\begin{aligned} y \sim 1 + \nu + \epsilon \ln(1 + \nu) - \epsilon\,(\eta - \eta_o) &= 1 + \nu - (1 + \nu)\tau + \epsilon\tau + \epsilon \ln(1 + \nu^{-1}) \\ &= 1 + \nu - (1 + \nu - \epsilon)[\tau - \epsilon \ln(1 + \nu^{-1})/(1 + \nu - \epsilon)], \end{aligned}$$

where (3.32), (3.36), and (3.42) of [3] have been used. Provided $\tau_o = \epsilon \ln(\nu^{-1}+1)/(1+\nu-\epsilon)$, the previous expression is $y \sim 1+\nu-(1+\nu-\epsilon)(\tau-\tau_o)$. This matches the one-term expansion $y = (1+\nu)\, e^{-(\tau-\tau_o)}$ for $\tau > \tau_o$ [3], except for a higher order term $\epsilon(\tau-\tau_o)$. Thus we obtain the composite expansion:

$$\text{(A.2)} \qquad y_{c,sas} = 1+\nu-\epsilon\,\ln\left(\frac{1+e^{\eta-\eta_o}}{1+\nu}\right) - \{[1+\nu+\epsilon\ln(1+\nu)]\,(1-e^{-(\tau-\tau o)}) \\ -\,\epsilon\,(\eta-\eta_o)\}\,H(\tau-\tau_o), \qquad \tau_o = \frac{\epsilon\ln(1+\nu^{-1})}{1+\nu-\epsilon}.$$

In the method of self-adjusting time scales, the radical concentration is given to leading order by

$$\text{(A.3)} \qquad x = \frac{y}{1+e^{-(\eta-\eta_o)}}, \quad f = y-x, \quad \theta = q(1+\nu-y),$$

according to (3.47), (3.49), and (3.50) of [3].

The composite (A.2)–(A.3) better approximates the temperature and fuel concentration than the leading order approximation of the method of self-adjusting scales with patching, but the differences from the numerical solution of the model are still of order ϵ. Further improvement comes from calculating the induction time better. Equation (B.21) of [3] gives the induction time including first order corrections according to the method of self-adjusting scales:

$$\text{(A.4)} \qquad t_{o,c} = \int_{\ln\nu}^{0} \frac{dh}{1+\nu-\epsilon+\epsilon\,[\beta q\,(1+\nu)-1]\,\ln\left(\frac{1+e^h}{1+\nu}\right)}\,,$$

which is somewhat smaller than τ_o/ϵ (for the parameter values in Figure 1, $\tau_o/\epsilon = 12.79$, $t_{o,c} = 12.44$). By replacing $\epsilon t_{o,c}$ instead of τ_o in (A.2), we obtain a better approximation:

$$\text{(A.5)} \qquad y_{c,sas} = 1+\nu-\epsilon\,\ln\left(\frac{1+e^{\eta-\eta_o}}{1+\nu}\right) - \{[1+\nu+\epsilon\ln(1+\nu)]\,(1-e^{-\epsilon\,(t-to,c)}) \\ -\,\epsilon\,(\eta-\eta_o)\}\,H(t-t_{o,c}),$$

$$\text{(A.6)} \quad t-t_{o,c} = \int_{0}^{\eta-\eta_o} \frac{dh}{1+\nu-\epsilon+\epsilon[q\beta(1+\nu)-1]\,\ln\left(\frac{1+e^h}{1+\nu}\right)}.$$

The relation (A.6) between $\eta-\eta_o$ and $t-t_{o,c}$ is obtained from (B.20) and (B.21) in [3]. Figure 3 shows that the corresponding x, f, and θ in (A.3) provide essentially an approximation of the solution throughout the whole time interval which is accurate to $O(\epsilon^2)$.

REFERENCES

[1] P. A. Blythe, A. K. Kapila, and M. Short, *Homogeneous ignition for a three-step chain-branching reaction model*, J. Engrg. Math., 56 (2006), pp. 105–128.

[2] L. L. Bonilla and H. T. Grahn, *Nonlinear dynamics of semiconductor superlattices*, Rep. Prog. Phys., 68 (2005), pp. 577–683.

[3] L. L. Bonilla, A. L. Sánchez, and M. Carretero, *The description of homogeneous branched-chain explosions with slow radical recombination by self-adjusting time scales*, SIAM J. Appl. Math., 61 (2000), pp. 528–550.

[4] A. Carpio and L. L. Bonilla, *Wave front depinning transition in discrete one-dimensional reaction-diffusion systems*, Phys. Rev. Lett., 86 (2001), pp. 6034–6037.

[5] G. Del Alamo and F. A. Williams, *Thermal-runaway approximation for ignition times of branched-chain explosions*, AIAA J., 43 (2005), pp. 2599–2605.

[6] R. Haberman, *Slowly varying jump and transition phenomena associated with algebraic bifurcation problems*, SIAM J. Appl. Math., 37 (1979), pp. 69–106.

[7] S. G. Hatzikiriakos and J. M. Dealy, *Role of slip and fracture in the oscillating flow of HDPE in a capillary*, J. Rheology, 36 (1992), pp. 845–884.

[8] G. Joulin, A. Liñán, G. S. S. Ludford, N. Peters, and C. Schmidt-Lainé, *Flames with chain-branching/chain-breaking kinetics*, SIAM J. Appl. Math., 45 (1985), pp. 420–434.

[9] A. K. Kapila, *Homogeneous branched-chain explosion: Initiation to completion*, J. Engrg. Math., 12 (1978), pp. 221–235.

[10] U. Landman, W. D. Luedtke, and J. P. Gao, *Atomic-scale issues in tribology: Interfacial junctions and nano-elastohydrodynamic*, Langmuir, 12 (1996), pp. 4514–4528.

[11] J. Moon and J. A. Wickert, *Non-linear vibration of power transmission belts*, J. Sound Vibration, 200 (1997), pp. 419–431.

[12] F.-J. Niedernostheide, ed., *Nonlinear Dynamics and Pattern Formation in Semiconductor and Devices*, Springer Proceedings in Physics 79, Springer, Berlin, 1995.

[13] E. L. Reiss, *A new asymptotic method for jump phenomena*, SIAM J. Appl. Math., 39 (1980), pp. 440–455.

[14] A. L. Sánchez, M. Carretero, P. Clavin, and F. Williams, *One-dimensional overdriven detonations with branched-chain kinetics*, Phys. Fluids, 13 (2001), pp. 776–792.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 629–647

SPECTRAL THEORY FOR AN ELASTIC THIN PLATE FLOATING ON WATER OF FINITE DEPTH[*]

CHRISTOPHE HAZARD[†] AND MICHAEL H. MEYLAN[‡]

Abstract. The spectral theory for a two-dimensional elastic plate floating on water of finite depth is developed (this reduces to a floating rigid body or a fixed body under certain limits). Two spectral theories are presented based on the first-order and second-order formulations of the problem. The first-order theory is valid only for a massless plate, while the second-order theory applies for a plate with mass. The spectral theory is based on an inner product (different for the first- and second-order formulations) in which the evolution operator is self-adjoint. This allows the time-dependent solution to be expanded in the eigenfunctions of the self-adjoint operator which are nothing more than the single frequency solutions. We present results which show that the solution is the same as those found previously when the water depth is shallow, and show the effect of increasing the water depth and the plate mass.

Key words. linear water waves, elastic plate, spectral expansion

AMS subject classifications. 76B15, 74B05, 35P25

DOI. 10.1137/060665208

1. Introduction.

This paper concerns the application of spectral theory to the offshore engineering problem of the linear wave problem in the presence of a floating elastic plate. We show how the time-dependent solution can be expressed as an expansion over the single frequency incoming wave solutions. As well as presenting the theory, we use the derived expressions to make practical calculations of the time-dependent motion of an elastic plate.

The time-dependent linear water wave problem has received considerable attention, and various solution methods have been developed. The simplest method is to use a time stepping method to advance the solution, combined with a solution for the Dirichlet-to-Neumann map [10]. The solution method is numerically demanding, and there will be significant error growth for long time calculations. A more mathematically sophisticated time stepping procedure is based on a kernel function derived from the single frequency solutions. This method is described in detail in [11] and has been recently applied to the solution of the time-dependent motion of a floating elastic plate [7, 19]. This method, while using the single frequency solutions, requires that the solution be stepped forward in time. The convolution of the kernel function is used, and for this reason the method is sometimes called the *memory effect*. This method requires only integration over the wetted surface of the body. There are other methods, including a method based on the time-dependent Green function [18, 22], which also requires time stepping and simultaneous solution of the body equations of motion. The spectral solution presented here is based on the single frequency solutions, but does not require a time stepping solution; instead the solution is calculated for all time using the fast Fourier transform (FFT).

[*] Received by the editors July 17, 2006; accepted for publication (in revised form) September 6, 2007; published electronically December 12, 2007.

http://www.siam.org/journals/siap/68-3/66520.html

[†] Laboratoire de Propagation des Ondes: Etude Mathématique et Simulation (UMR 2706 du CNRS): ENSTA / POems, 32 Boulevard Victor, 75015 Paris, France (hazard@ensta.fr).

[‡] Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand (meylan@math.auckland.ac.nz).

Spectral theory for finite-dimensional or compact operators is well known, for example, the calculation of the modes and frequencies of vibration of an elastic plate. For the operators in linear wave theory, which are self-adjoint but not compact, the spectral theory is much more complicated. This theory has been developed for fixed bodies [1, 20, 3, 4] in infinitely deep water. The focus of these works was on developing the theory, and no calculations were made. Only [12], which considered the problem of a floating massless elastic plate in the case of shallow water (which greatly simplified the equations of motion), has presented any calculations using a spectral method for a linear wave problem. The solution in [12] was found by both a spectral expansion and by Lax–Philips theory. The present work can be seen as an extension of the spectral solution of [12] to the considerably more complicated situation when the water can no longer be approximated as shallow.

The floating elastic plate is a very natural problem to consider when attempting to use spectral theory to make practical calculations of time-dependent motion. The floating elastic plate is amongst the best-studied problems in hydroelasticity and has been used to model floating breakwaters [18], ice floes [17], and very large floating structures [6, 21]. The single frequency (time harmonic) response has been well studied. While more complicated to study than the rigid body, the elastic body reduces to a rigid floating body and to a rigid dock in various limits as the stiffness and mass tend to infinity. Therefore, by providing a solution to the elastic plate problem, we are also providing a solution to the rigid floating body and dock problem. Furthermore, [19] has developed a solution to the time-dependent problem for a floating elastic plate on water of infinite depth.

As mentioned previously, the single frequency response for a floating elastic plate has been well studied. We will give a brief summary of the research in the case of the two-dimensional problem. The problem is mentioned in [18] where a solution for the case of shallow water is presented. The first solutions to the problem, without the assumptions of shallow water or a small plate effect, were by [13] and [15]. These solutions are based on using the Green function methods which had been developed for rigid bodies. The elastic plate problem (because of the assumption of shallow draft) actually has an eigenfunction expansion in the plate covered and non–plate covered regions. This property was exploited by [2] but only for the case of a semi-infinite plate. The matching using inner products was recently considered by [9], and a solution method was derived using residue calculus, but only for a semi-infinite plate. A fast method for multiple plates has recently been developed by [8].

In this paper, we present the time-dependent solution to the floating elastic plate on water of finite depth using spectral theory. The solution is based on finding an inner product in which the non–time-dependent operator which occurs in the time-dependent equation is self-adjoint. We show in section 2 that there are two formulations which can be developed; the first is based on a first-order equation in time and is closely related to the method used in [12]. However, this formulation is valid only for a massless plate. The second formulation is based on the second-order time-dependent equation and does allow for the plate to have mass. The spectral approach that we follow to solve the time-dependent problem has a quite natural physical interpretation: it consists of expanding the time-dependent solution in a basis of time-harmonic solutions. The main difficulty mathematically is to normalize the generalized eigenfunctions (time-harmonic solutions), and we use a perturbation technique based on the generalized eigenfunctions for the pure hydrodynamic problem, that is, in the absence of the plate (section 3). We then have to determine the associated scattered waves in the presence of the plate (section 4): we thus obtain a basis of generalized eigenfunctions for the coupled problem. Finally, we present in section 5 some numerical results: they agree with those found earlier by [12] where the equivalent theory was developed for shallow water, and show the effect of water depth on the time-dependent solution.

2. Statement of the problem and mathematical formulations.

2.1. Governing equations.

The plate is infinite in the y direction, so that only the x and z directions are considered. The x direction is horizontal, the positive z axis points vertically up, and the plate covers the region $-b \leqslant x \leqslant b$. The water is of uniform depth h. The amplitudes are assumed small enough that the linear theory is appropriate, and the plate is sufficiently thin that the shallow draft approximation may be made [21].

The mathematical description of the problem follows from [18]. The kinematic condition is

$$\partial_t \zeta = \partial_n \Phi, \qquad z = 0,$$

where ζ is the displacement of the water surface or the plate (from the shallow draft approximation), ∂_n is the outward normal derivative, and Φ is the velocity potential of the water, which satisfies

$$\begin{aligned} \Delta\Phi &= 0, \qquad -h < z < 0, \\ \partial_n \Phi &= 0, \qquad z = -h. \end{aligned}$$

On the other hand, the dynamic condition obtained by matching the pressure at the free surface is

$$-\rho g \zeta - \rho \partial_t \Phi = \begin{cases} 0, & x \notin (-b, b), \\ D\partial_x^4 \zeta + \rho' d\, \partial_t^2 \zeta, & x \in (-b, b), \end{cases} \qquad z = 0,$$

where D is the bending rigidity of the plate per unit length, ρ is the density of water, ρ' is the density of the plate, d is the plate thickness, and g is the acceleration due to gravity. At the ends of the plate the free edge boundary conditions

$$\lim_{x\downarrow -b} \partial_x^2 \zeta = \lim_{x\uparrow b} \partial_x^2 \zeta = \lim_{x\downarrow -b} \partial_x^3 \zeta = \lim_{x\uparrow b} \partial_x^3 \zeta = 0$$

are applied.

Nondimensional variables are now introduced using a length parameter L for the space variables and $\sqrt{L/g}$ for the time variable. We leave the choice of the length parameter arbitrary, since there are two natural length parameters, the water depth and the characteristic length $(D/\rho g)^{1/4}$. It also means that we can present results in our nondimensional variables in which the plate properties are kept constant and the water depth is varied. Hence the nondimensional surface displacement and velocity potential satisfy the following coupled equations, where the overbar denotes nondimensional variables,

$$\partial_{\bar t} \bar\zeta = \partial_{\bar n} \bar\Phi, \qquad \bar z = 0, \tag{2.1}$$

$$\bar\Delta \bar\Phi = 0, \qquad -\bar h < \bar z < 0, \tag{2.2}$$

$$\partial_{\bar n} \bar\Phi = 0, \qquad \bar z = -\bar h, \tag{2.3}$$

$$-\bar\zeta - \partial_{\bar t} \bar\Phi = \begin{cases} 0, & \bar x \notin (-\bar b, \bar b), \\ \beta \partial_{\bar x}^4 \bar\zeta + \gamma \partial_{\bar t}^2 \bar\zeta, & \bar x \in (-\bar b, \bar b), \end{cases} \qquad \bar z = 0, \tag{2.4}$$

plus the free edge boundary conditions

$$\lim_{\bar{x}\downarrow -\bar{b}} \partial_{\bar{x}}^2 \bar{\zeta} = \lim_{\bar{x}\uparrow \bar{b}} \partial_{\bar{x}}^2 \bar{\zeta} = \lim_{\bar{x}\downarrow -\bar{b}} \partial_{\bar{x}}^3 \bar{\zeta} = \lim_{\bar{x}\uparrow \bar{b}} \partial_{\bar{x}}^3 \bar{\zeta} = 0, \tag{2.5}$$

where $\beta = D/(\rho g h^4)$ and $\gamma = \rho' d/(\rho h)$. Note that if we consider the limit as $\beta \to \infty$, we obtain a rigid floating body (with negligible submergence), and if we consider the limit as $\beta \to \infty$ and $\gamma \to \infty$, then we obtain the dock boundary condition ($\eta = 0$, $\bar{x} \in (-\bar{b}, \bar{b})$). For clarity the overbar is dropped from now on.

Our aim is to exhibit a spectral expansion of the solution to these equations together with suitable initial conditions at time $t = 0$. To do so, we shall rewrite this system in an abstract form which involves a *self-adjoint* operator: such a property is essential for the application of the spectral approach. We propose below two such mathematical formulations corresponding to two different choices of inner products.

2.2. Two component energy inner product for a massless plate. We first introduce an inner product based on the usual mechanical energy, to which both potential and displacement contribute: this energy consists of the kinetic energy of the water ($\propto |\nabla\Phi|^2$), the potential energy of the water ($\propto |\zeta|^2$), and the energy of the plate. This approach is a generalization to finite depth of the inner product used in [12]. As was the case in [12], this inner product is based on the assumption that the plate is massless. This means that we consider only the stiffness of the plate, not its inertia, in calculating the equations of motion. This is obviously an approximation, but it is actually valid for many practical situations. This approximation is discussed in [12] and [14].

The starting point of the abstract formulation lies in the following remark: if we know the velocity potential only at the surface of the water, say $\phi(x,t) = \Phi(x,0,t)$, then (2.2) and (2.3) determine Φ everywhere. Consider then the operator $\mathbf{G}$ which maps ϕ onto $\Phi = \mathbf{G}\phi$: this is the harmonic lifting which solves the boundary value problem

$$\begin{aligned} &\Delta\Phi = 0, && -h < \bar{z} < 0,\\ &\Phi = \phi, && z = 0,\\ &\partial_n \Phi = 0, && z = -h. \end{aligned} \tag{2.6}$$

This operator allows us to rewrite our system (2.1)–(2.5) with $\gamma = 0$ as a problem set only on $z = 0$:

$$\begin{aligned} \partial_t \zeta &= \partial_n \mathbf{G}\phi,\\ -\zeta - \partial_t \phi &= \chi_P \beta \partial_x^4 \zeta, \end{aligned}$$

where χ_P is the characteristic function for the plate covered region $P = (-b, +b)$ (i.e., $\chi_P(x) = 1$ if $x \in P$; else $\chi_P(x) = 0$). The edge conditions (2.5) are omitted for simplicity.

If we combine ζ and ϕ in a two component vector $U(x,t)$ given by

$$U(x,t) = \begin{pmatrix} \phi(x,t) \\ \mathrm{i}\zeta(x,t) \end{pmatrix}, \tag{2.7}$$

the above coupled equations turn into a Schrödinger-type equation

$$i\partial_t U = \mathcal{P}U, \quad \text{with} \quad \mathcal{P} = \begin{pmatrix} 0 & 1 + \chi_P \beta \partial_x^4 \\ \partial_n \mathbf{G} & 0 \end{pmatrix}. \tag{2.8}$$

The problem is to find an inner product $\langle \cdot\,, \cdot \rangle_{\mathcal{V}}$, defining a Hilbert space $\mathcal{V}$, for which $\mathcal{P}$ becomes self-adjoint. We shall actually verify only the symmetry property:

$$\langle \mathcal{P}V, V' \rangle_{\mathcal{V}} = \langle V, \mathcal{P}V' \rangle_{\mathcal{V}}. \tag{2.9}$$

The questions related to the domain of $\mathcal{P}$ are essential in the mathematical definition of self-adjointness but not in the numerical implementation of the method which is the focus of the present work. The works of [1, 20, 3] which are focused on the mathematics do treat the domain rigorously, however only for the case of a rigid body. A mathematically rigorous treatment in the case of an elastic body remains undeveloped.

A convenient way to interpret the symmetry property of $\mathcal{P}$ is to exhibit the associated bilinear form (we should say more precisely "sesquilinear"):

$$\langle \mathcal{P}V, V' \rangle_{\mathcal{V}} = p(V, V'). \tag{2.10}$$

It is then clear that (2.9) holds if $p(\cdot\,, \cdot)$ is symmetric, that is,

$$p(V', V) = \overline{p(V, V')}. \tag{2.11}$$

In order to exhibit $p(\cdot\,, \cdot)$ in our case, first consider the operator $\partial_n \mathbf{G}$ involved in the definition of $\mathcal{P}$. Choose the usual inner product of $L^2(\mathbb{R})$, i.e.,

$$\langle \psi, \psi' \rangle_{\mathbb{R}} = \int_{\mathbb{R}} \psi(x)\, \overline{\psi'(x)}\, dx.$$

From (2.6), Green's formula yields

$$\langle \partial_n \mathbf{G}\psi, \psi' \rangle_{\mathbb{R}} = \int_{\mathbb{R}} \int_{-h}^{0} \nabla(\mathbf{G}\psi) \cdot \overline{\nabla(\mathbf{G}\psi')}\, dx\, dz, \tag{2.12}$$

where the right-hand side clearly defines a positive symmetric form: the operator $\partial_n \mathbf{G}$ is thus positive and self-adjoint in $L^2(\mathbb{R})$. On the other hand, the operator $\beta\partial_x^4$ is positive and self-adjoint in $L^2(P)$. Indeed, if ζ satisfies the edge conditions (2.5), integrating by parts twice gives

$$\langle \beta\partial_x^4 \zeta, \zeta' \rangle_P = \beta \langle \partial_x^2 \zeta, \partial_x^2 \zeta' \rangle_P. \tag{2.13}$$

Then it may seem natural to consider $\mathcal{P}$ in $L^2(\mathbb{R}) \times L^2(\mathbb{R})$. But $\mathcal{P}$ would not be self-adjoint. The symmetry property actually occurs in a subspace, defined by the inner product

$$\langle V, V' \rangle_{\mathcal{V}} = \int_{\mathbb{R}} \int_{-h}^{0} \nabla(\mathbf{G}\psi) \cdot \overline{\nabla(\mathbf{G}\psi')}\, dx\, dz + \langle \xi, \xi' \rangle_{\mathbb{R}} + \beta \langle \partial_x^2 \xi, \partial_x^2 \xi' \rangle_P \tag{2.14}$$

for all complex vector functions

$$V = \begin{pmatrix} \psi \\ \xi \end{pmatrix} \quad \text{and} \quad V' = \begin{pmatrix} \psi' \\ \xi' \end{pmatrix}. \tag{2.15}$$

To see this, we simply have to use (2.12) and (2.13) in the relation

$$\begin{aligned} \langle \mathcal{P}V, V' \rangle_{\mathcal{V}} = \int_{\mathbb{R}} \int_{-h}^{0} & \nabla(\mathbf{G}(1 + \chi_P \beta\partial_x^4)\xi) \cdot \overline{\nabla(\mathbf{G}\psi')}\, dx\, dz \\ & + \langle \partial_n \mathbf{G}\psi, \xi' \rangle_{\mathbb{R}} + \beta \langle \partial_x^2(\partial_n \mathbf{G}\psi), \partial_x^2 \xi' \rangle_P, \end{aligned}$$

which yields

$$\langle \mathcal{P}V, V' \rangle_{\mathcal{V}} = \left\langle (1 + \chi_P \beta \partial_x^4)\xi, \partial_n \mathbf{G}\psi' \right\rangle_{\mathbb{R}} + \left\langle \partial_n \mathbf{G}\psi, (1 + \chi_P \beta \partial_x^4)\xi' \right\rangle_{\mathbb{R}},$$

where ξ is implicitly assumed to satisfy the edge conditions (2.5). The bilinear form on the right-hand side is clearly symmetric, and thus $\mathcal{P}$ is self-adjoint (but not positive).

At least formally, we can express the solution to (2.8) as

$$U(t) = e^{i\mathcal{P}t} U_0 \quad \text{with} \quad U_0(x) = U(x,0) = \begin{pmatrix} \phi_0(x) \\ i\zeta_0(x) \end{pmatrix}, \tag{2.16}$$

where $\exp(i\mathcal{P}t)$ is a unitary operator that we will make explicit by a spectral expansion.

2.3. Single component inner product which allows nonzero mass. In this section we will derive an abstract formulation which allows the case of nonzero mass. The inner product involved in this formulation is not based on the total energy and is closer to the standard L^2 inner product.

We introduce a new variable $\Psi = -\partial_t \Phi$, i.e., the opposite of the acceleration potential. Our system (2.1)–(2.4) then becomes

$$\partial_t^2 \zeta + \partial_n \Psi = 0, \qquad z = 0, \tag{2.17}$$

$$\Delta \Psi = 0, \qquad -h < z < 0, \tag{2.18}$$

$$\partial_n \Psi = 0, \qquad z = -h, \tag{2.19}$$

$$-\zeta + \Psi = \begin{cases} 0, & x \notin (-b,b), \\ \beta \partial_x^4 \zeta + \gamma \partial_t^2 \zeta, & x \in (-b,b), \end{cases} \qquad z = 0. \tag{2.20}$$

Instead of the harmonic lifting $\mathbf{G}$ introduced in the first-order formulation, we consider the operator $\mathbf{H}$ which maps ζ onto Ψ, where the latter solves the boundary value problem

$$\begin{aligned} &\Delta \Psi = 0, \qquad -h < z < 0, \\ &\partial_n \Psi = 0, \qquad z = -h, \\ &-\zeta + \Psi = \begin{cases} 0, & x \notin (-b,b), \\ \beta \partial_x^4 \zeta + \gamma \partial_n \Psi, & x \in (-b,b), \end{cases} \qquad z = 0. \end{aligned}$$

Our system (2.17)–(2.20) thus turns into the following second-order equation:

$$\partial_t^2 \zeta + \partial_n \mathbf{H} \zeta = 0. \tag{2.21}$$

The operator $\partial_n \mathbf{H}$ is positive and self-adjoint in the Hilbert space $\mathcal{H}$ with inner product

$$\langle \zeta, \zeta' \rangle_{\mathcal{H}} = \langle \zeta, \zeta' \rangle_{\mathbb{R}} + \beta \langle \partial_x^2 \zeta, \partial_x^2 \zeta' \rangle_P. \tag{2.22}$$

Indeed, let $\Psi = \mathbf{H}\zeta$ and $\Psi' = \mathbf{H}\zeta'$ for arbitrary ζ and ζ' satisfying (2.5). Using Green's formula and the above definition of $\mathbf{H}$, we have

$$\begin{aligned} &\int_{\mathbb{R}} \int_{-h}^{0} \nabla \Psi \cdot \overline{\nabla \Psi'} \, dx \, dz \\ &\qquad = \langle \partial_n \Psi, \zeta' \rangle_{\mathbb{R}} + \beta \langle \partial_n \Psi, \partial_x^4 \zeta' \rangle_P - \gamma \langle \partial_n \Psi, \partial_n \Psi' \rangle_P \\ &\qquad = \langle \partial_n \Psi, \zeta' \rangle_{\mathbb{R}} + \beta \langle \partial_x^2 (\partial_n \Psi), \partial_x^2 \zeta' \rangle_P - \gamma^{-1} \langle \zeta + \beta \partial_x^4 \zeta - \Psi, \zeta' + \beta \partial_x^4 \zeta' - \Psi' \rangle_P. \end{aligned}$$

Substituting this relation into

$$\langle \partial_n \mathbf{H}\zeta, \zeta' \rangle_{\mathcal{H}} = \langle \partial_n \mathbf{H}\zeta, \zeta' \rangle_{\mathbb{R}} + \beta \langle \partial_x^2 (\partial_n \mathbf{H}\zeta), \partial_x^2 \zeta' \rangle_P$$

yields

$$\langle \partial_n \mathbf{H}\zeta, \zeta' \rangle_{\mathcal{H}} = \int_{\mathbb{R}} \int_{-h}^{0} \nabla(\mathbf{H}\zeta) \cdot \overline{\nabla(\mathbf{H}\zeta')} \, dx \, dz + \gamma^{-1} \langle \zeta + \beta \partial_x^4 \zeta - \mathbf{H}\zeta, \zeta' + \beta \partial_x^4 \zeta' - \mathbf{H}\zeta' \rangle_P .$$

The right-hand side defines a positive symmetric bilinear form, which shows that $\partial_n \mathbf{H}$ is a positive self-adjoint operator in $\mathcal{H}$.

Thanks to this property, we shall construct the spectral expansion of the solution to (2.21) together with the initial conditions

$$\zeta(x,0) = \zeta_0(x) \quad \text{and} \quad \partial_t \zeta(x,0) = \theta_0(x), \tag{2.23}$$

where θ_0 is related to ϕ_0 in (2.16) by the relation $\theta_0 = \partial_n \mathbf{G}\phi_0$. This solution is

$$\zeta(t) = \cos((\partial_n \mathbf{H})^{1/2} t) \, \zeta_0 + (\partial_n \mathbf{H})^{-1/2} \sin((\partial_n \mathbf{H})^{1/2} t) \, \theta_0,$$

or equivalently, in the condensed form

$$\zeta(t) = \mathrm{Re}\{\exp(i(\partial_n \mathbf{H})^{1/2} t)\eta_0\} \quad \text{with} \quad \eta_0 = \zeta_0 - i(\partial_n \mathbf{H})^{-1/2} \theta_0. \tag{2.24}$$

3. Spectral decompositions without the floating plate. We consider in this section the *free* problem associated with our scattering problem, obtained by removing the floating plate. The basic ideas of the spectral approach that we propose for solving the preceding equations are described here. The solution is simply the solution found using the Fourier transform explained in the context of *generalized eigenfunctions* expansion. We will use the normalization results for the simpler free problem to derive the normalization for the problem with the plate.

3.1. Spectral representation for the first-order equation. If we remove the plate, the first-order problem (2.8) then simplifies to

$$i\partial_t U = \tilde{\mathcal{P}} U, \quad \text{with} \quad \tilde{\mathcal{P}} = \begin{pmatrix} 0 & 1 \\ \partial_n \mathbf{G} & 0 \end{pmatrix}, \tag{3.1}$$

where the tilde symbol refers to the free problem. Similarly to $\mathcal{P}$, the operator $\tilde{\mathcal{P}}$ is self-adjoint in the free energy space $\tilde{\mathcal{V}}$ defined by the inner product (see (2.14))

$$\langle V, V' \rangle_{\tilde{\mathcal{V}}} = \int_{\mathbb{R}} \int_{-h}^{0} \nabla(\mathbf{G}\psi) \cdot \overline{\nabla(\mathbf{G}\psi')} \, dx \, dz + \langle \xi, \xi' \rangle_{\mathbb{R}} \tag{3.2}$$

for all V and V' defined by (2.15). The solution to (3.1) can be formally expressed as a function of the initial state $U(0) = U_0$:

$$U(t) = e^{i\tilde{\mathcal{P}} t} U_0. \tag{3.3}$$

Spectral theory offers a way to express $\exp(i\tilde{\mathcal{P}} t)$ by means of the eigenelements of $\tilde{\mathcal{P}}$, that is, the solutions to the eigenvalue problem

$$(\tilde{\mathcal{P}} - \lambda) V = 0 \quad \text{with } V \neq 0 \text{ and } \lambda \in \mathbb{R}. \tag{3.4}$$

For our free problem $\tilde{\mathcal{P}}$ possesses a *continuous spectrum.* (Recall that a point $\lambda \in \mathbb{R}$ belongs to the continuous spectrum if $\mathcal{P} - \lambda$ is injective but not surjective.)

The generalized eigenfunctions of $\tilde{\mathcal{P}}$ can be found easily and are given by

$$\tilde{U}_{\lambda,\kappa}(x) = \exp(i\kappa k(\lambda^2)x) \begin{pmatrix} \lambda^{-1} \\ 1 \end{pmatrix}, \tag{3.5}$$

where $\kappa = \pm 1$ and $k(\lambda^2)$ is the positive root of the dispersion equation

$$k \tanh kh = \lambda^2. \tag{3.6}$$

We can easily show using the well-known orthogonality relation for the Fourier transform that

$$\begin{aligned} \left\langle \tilde{U}_{\lambda,\kappa}, \tilde{U}_{\lambda',\kappa'} \right\rangle_{\tilde{\mathcal{V}}} &= \int_{\mathbb{R}} \int_{-h}^{0} \frac{1}{\lambda\lambda'} \nabla(\mathbf{G}e^{i\kappa kx}) \cdot \overline{\nabla(\mathbf{G}e^{i\kappa' k' x})}\, dx\, dz + \left\langle e^{i\kappa kx}, e^{i\kappa' k' x} \right\rangle_{\mathbb{R}} \\ &= \frac{k \tanh kh}{\lambda\lambda'} \int_{\mathbb{R}} \frac{1}{\lambda\lambda'} e^{i\kappa kx} \overline{e^{i\kappa' k' x}}\, dx + \left\langle e^{i\kappa kx}, e^{i\kappa' k' x} \right\rangle_{\mathbb{R}} \\ &= 4\pi \left| \frac{d\lambda}{dk} \right| \delta_{\kappa\kappa'} \delta\left(\lambda - \lambda'\right). \end{aligned}$$

Note that we have used the property of the free linear waves that

$$\partial_n \mathbf{G} e^{i\kappa kx} = k \tanh kh e^{i\kappa kx}.$$

Therefore the solution to the free Schrödinger-type equation (3.1) for an initial state U_0 is given by

$$U(t) = \frac{1}{4\pi} \int_{\mathbb{R}} e^{i\lambda t} \sum_{\kappa=\pm 1} \langle U_0, \tilde{U}_{\lambda,\kappa} \rangle_{\tilde{\mathcal{V}}} \tilde{U}_{\lambda,\kappa} \left| \frac{dk}{d\lambda} \right| d\lambda. \tag{3.7}$$

3.2. Spectral representation for the second-order equation $\partial_n \mathbf{G}$. On the other hand, the second-order wave equation (2.21) becomes

$$\partial_t^2 \zeta + \partial_n \mathbf{G} \zeta = 0, \tag{3.8}$$

where $\partial_n \mathbf{G}$ is positive and self-adjoint in $L^2(\mathbb{R})$ (see (2.12)). We denote here by ν instead of λ the spectral variable for $\partial_n \mathbf{G}$: from (3.8), it represents a frequency squared, whereas λ stands for a signed frequency in the previous section.

It can easily be shown that the generalized eigenfunctions of $\partial_n \mathbf{G}$ are given by the linear wave solutions for a fixed frequency given by

$$\tilde{\zeta}_{\nu,\kappa}(x) = e^{i\kappa\, k(\nu)x} \quad \text{for } \nu \in \mathbb{R}^+ \text{ and } \kappa = \pm 1, \tag{3.9}$$

where $k(\nu)$ is the positive root of the dispersion equation (3.6). From the properties of the Fourier transform we know that

$$\langle \tilde{\zeta}_{\nu,\kappa}\, , \tilde{\zeta}_{\nu',\kappa'} \rangle_{\mathbb{R}} = 2\pi \frac{d\nu}{dk}\, \delta_{\kappa,\kappa'}\, \delta(\nu - \nu'). \tag{3.10}$$

This means that the solution to (3.8) together with initial conditions of the form (2.23) is

$$\begin{aligned} \zeta(t) &= \mathrm{Re}\{\exp(i(\partial_n \mathbf{G})^{1/2} t)\eta_0\}, \quad \text{where } \eta_0 = \zeta_0 - i(\partial_n \mathbf{H})^{-1/2}\theta_0, \\ &= \mathrm{Re}\left\{ \frac{1}{2\pi} \int_{\mathbb{R}^+} e^{i\sqrt{\nu}t} \sum_{\kappa=\pm 1} \langle \eta_0, \tilde{\zeta}_{\nu,\kappa} \rangle_{\mathcal{H}}\, \tilde{\zeta}_{\nu,\kappa}\, \frac{dk}{d\nu}\, d\nu \right\}. \end{aligned}$$

4. Spectral decompositions of the hydroelastic problem. In this section we will write down the spectral expansions for the plate-water system. These expansions will be based closely on the expansions we have just derived for the case when the plate is absent. We find the generalized eigenfunctions from the solution for a unit incident wave together with the associated scattered waves generated by the coupling with the plate: their superpositions (incident + scattered) yield a generalized spectral basis for the coupled problem. Here we choose *outgoing* scattered waves, but the same result holds for scattered waves that are incoming.

4.1. The first-order formulation. We denote by $U_{\lambda,\kappa}$ the eigenfunctions in the presence of the plate, which correspond to the incident waves $\tilde{U}_{\lambda,\kappa}$ coming respectively from the right ($\kappa = \operatorname{sgn}(\lambda)$) and the left ($\kappa = -\operatorname{sgn}(\lambda)$). These are single frequency solutions: they satisfy the equation

$$\begin{pmatrix} 0 & 1+\chi_P\beta\partial_x^4 \\ \partial_n\mathbf{G} & 0 \end{pmatrix} U_{\lambda,\kappa} = \lambda U_{\lambda,\kappa},$$

and have the asymptotics

$$U_{\lambda,+1}(x) \sim \left(e^{+ikx} + S_{11}\,e^{-ikx}\right)\begin{pmatrix} \lambda^{-1} \\ 1 \end{pmatrix} \quad \text{as } x \to +\infty, \tag{4.1}$$
$$U_{\lambda,+1}(x) \sim \left(S_{12}\,e^{+ikx}\right)\begin{pmatrix} \lambda^{-1} \\ 1 \end{pmatrix} \quad \text{as } x \to -\infty,$$

and

$$U_{\lambda,-1}(x) \sim \left(S_{22}\,e^{-ikx}\right)\begin{pmatrix} \lambda^{-1} \\ 1 \end{pmatrix} \quad \text{as } x \to +\infty, \tag{4.2}$$
$$U_{\lambda,-1}(x) \sim \left(e^{-ikx} + S_{21}\,e^{+ikx}\right)\begin{pmatrix} \lambda^{-1} \\ 1 \end{pmatrix} \quad \text{as } x \to -\infty,$$

where S_{11}, S_{12}, S_{21}, and S_{22} are the reflection and transmission coefficients (which must be determined). We define the components of $U_{\lambda,\kappa}$ as

$$U_{\lambda,\kappa}(x) = \begin{pmatrix} \phi_{\lambda,\kappa}(x) \\ \xi_{\lambda,\kappa}(x) \end{pmatrix}.$$

Setting $\Phi_{\lambda,\kappa} = \mathbf{G}\phi_{\lambda,\kappa}$, the above eigenvalue problem amounts to the boundary value problem

$$\begin{aligned} &\Delta\Phi_{\lambda,\kappa} = 0, \qquad -h < z < 0, \\ &\partial_n\Phi_{\lambda,\kappa} = 0, \qquad z = -h, \\ &\lambda\Phi_{\lambda,\kappa} = \left(1+\chi_P\beta\partial_x^4\right)\xi_{\lambda,\kappa}, \qquad z = 0, \\ &\partial_n\Phi_{\lambda,\kappa} = \lambda\,\xi_{\lambda,\kappa}, \qquad z = 0, \end{aligned} \tag{4.3}$$

subject to the appropriate radiation conditions given by (4.1) and (4.2) plus the free edge boundary conditions (2.5). There are a number of methods which can be used to solve this problem. The equation was solved in [13] using a Green function for the water and plate. It was solved in [15] using a Green function for the water and an expansion in modes for the plate. The solution method is described in the appendix. Note that we are assuming that there is no point spectrum, i.e., that the floating elastic plate problem does not admit any trapped modes. While there is no proof that trapped modes do not exist, all evidence points to this conclusion. The present theory could be extended to include trapped modes, following [16, 5, 4]; it seems sensible not to do this since we strongly believe they do not exist.

Having found the generalized eigenfunctions, we need to determine their normalization; that is, we know that

$$\langle U_{\lambda,\kappa}, U_{\lambda',\kappa'}\rangle = q_\lambda^{-1}\,\delta(\lambda-\lambda')\,\delta_{\kappa\kappa'}, \tag{4.4}$$

but we have no easy way to determine q_λ. We can use a formal argument as was done in [12]. However, we can find this normalization using a very powerful result from spectral theory, namely that the normalization of the perturbed eigenfunctions is identical to the normalization of the free problem (i.e., the problem without the plate discussed previously). We do not present a proof of this here, but note that this has been done in many related situations. The first proofs were for the Schrödinger equation [16, 5] and for the Helmholtz equation [23]. Recently, a proof was given for water waves for a rigid floating body in infinite depth [3, 4]. The proof for the plate can be found following an argument similar to that given in [4]. In particular the eigenfunctions satisfy the following orthogonality conditions:

$$\langle U_{\lambda,\kappa}, U_{\lambda',\kappa'}\rangle_{\mathcal{V}} = 4\pi\left|\frac{d\lambda}{dk}\right|\delta_{\kappa\kappa'}\delta\left(\lambda-\lambda'\right).$$

This normalization agrees in the appropriate limit with the normalizations of [3, 12] for infinite depth and shallow water, respectively.

We can express the solution to (2.8) as a spectral expansion as we did for the free problem, except that we use the eigenfunctions $U_{\lambda,\kappa}$ defined by (4.3) and the space $\mathcal{V}$, i.e.,

$$U(t) = \frac{1}{4\pi}\int_{\mathbb{R}} e^{i\lambda t}\sum_{\kappa=\pm1}\langle U_0, U_{\lambda,\kappa}\rangle_{\mathcal{V}}\, U_{\lambda,\kappa}\left|\frac{dk}{d\lambda}\right| d\lambda, \quad \text{where} \quad U_0 = \begin{pmatrix}\phi_0\\ i\zeta_0\end{pmatrix}. \tag{4.5}$$

The calculation of the inner products can be simplified by observing that

$$\begin{aligned}\langle U_0, U_{\lambda,\kappa}\rangle_{\mathcal{V}} &= \langle \partial_n\mathbf{G}\phi_0, \phi_{\lambda,\kappa}\rangle_{\mathbb{R}_x} + \left\langle\left(1+\chi_P\beta\partial_x^4\right)i\zeta_0, \xi_{\lambda,\kappa}\right\rangle_{\mathbb{R}_x}\\ &= \langle \phi_0, \partial_n\mathbf{G}\phi_{\lambda,\kappa}\rangle_{\mathbb{R}_x} + \left\langle i\zeta_0, \left(1+\chi_P\beta\partial_x^4\right)\xi_{\lambda,\kappa}\right\rangle_{\mathbb{R}_x}\\ &= \langle \phi_0, \lambda\,\xi_{\lambda,\kappa}\rangle_{\mathbb{R}_x} + \langle i\zeta_0, \lambda\,\phi_{\lambda,\kappa}\rangle_{\mathbb{R}_x}\\ &= \lambda\left(\langle \phi_0, \xi_{\lambda,\kappa}\rangle_{\mathbb{R}_x} + \langle i\zeta_0, \phi_{\lambda,\kappa}\rangle_{\mathbb{R}_x}\right).\end{aligned} \tag{4.6}$$

4.2. The second-order formulation. The second-order spectral expansion follows from (2.21) and (3.10) exactly as with the first-order formulation. The equation for the eigenfunctions is

$$\partial_n\mathbf{H}\zeta_{\nu,\kappa} = \nu\zeta_{\nu,\kappa},$$

where we define the asymptotics so that $\zeta_{\nu,\kappa}$ corresponds to the incident wave free wave $\tilde{\zeta}_{\nu,\kappa}$ defined in (3.9), i.e.,

$$\begin{aligned}&\zeta_{\lambda,+1}(x) \sim e^{+ikx} + S_{11}\,e^{-ikx} \quad \text{as } x\to+\infty,\\ &\zeta_{\lambda,+1}(x) \sim S_{12}\,e^{+ikx} \quad \text{as } x\to-\infty,\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}&\zeta_{\lambda,-1}(x) \sim S_{22}\, e^{-ikx} \quad \text{as } x \to -\infty,\\ &\zeta_{\lambda,-1}(x) \sim e^{-ikx} + S_{21}\, e^{+ikx} \quad \text{as } x \to +\infty.\end{aligned} \tag{4.8}$$

Setting $\Psi_{\nu,\kappa} = \mathbf{H}\zeta_{\nu,\kappa}$, the boundary value problem to solve is

$$\begin{aligned}&\Delta\Psi_{\nu,\kappa} = 0, \qquad -h < z < 0,\\ &\partial_n \Psi_{\nu,\kappa} = 0, \qquad z = -h,\\ &-\zeta_{\nu,\kappa} + \Psi_{\nu,\kappa} = \chi_P \left(\beta\partial_x^4\zeta_{\nu,\kappa} + \gamma\partial_n\Psi_{\nu,\kappa}\right), \qquad z = 0,\\ &\partial_n\Psi_{\nu,\kappa} = \nu\zeta_{\nu,\kappa}, \qquad z = 0,\end{aligned} \tag{4.9}$$

subject to the radiation conditions (4.7) and (4.8) and the free edge conditions (2.5). We solve this system by the method described in [8]. Note that (4.9) is equivalent to the boundary value problem (4.3) when the mass γ is zero, taking $\nu = \lambda^2$ and

$$\xi_{\lambda,\kappa} = i\,\zeta_{\lambda^2,\kappa} \quad \text{and} \quad \Phi_{\lambda,\kappa} = i\lambda^{-1}\,\Psi_{\lambda^2,\kappa}. \tag{4.10}$$

The orthogonality relations follow from those for the free problem given by (3.10), and therefore

$$\langle \zeta_{\nu,\kappa}, \zeta_{\nu',\kappa'}\rangle_{\mathcal{H}} = 2\pi \frac{d\nu}{dk}\delta_{\kappa\kappa'}\delta\left(\nu - \nu'\right).$$

The explicit expression of the time-dependent solution (2.24) to (2.21) is

$$\begin{aligned}\zeta(t) &= \mathrm{Re}\{\exp(i(\partial_n\mathbf{H})^{1/2}t)\eta_0\}\\ &= \mathrm{Re}\left\{\frac{1}{2\pi}\int_{\mathbb{R}^+} e^{i\sqrt{\nu}t}\sum_{\kappa=\pm1}\langle\eta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}}\,\zeta_{\nu,\kappa}\,\frac{dk}{d\nu}\,d\nu\right\},\end{aligned} \tag{4.11}$$

where we recall that $\eta_0 = \zeta_0 - i(\partial_n\mathbf{H})^{-1/2}\theta_0$ and $\theta_0 = \partial_n\mathbf{G}\phi_0$. The real part in this expression takes the form

$$\zeta(t) = \frac{1}{2\pi}\int_{\mathbb{R}^+}\sum_{\kappa=\pm1}\left(\cos(\sqrt{\nu}t)\,\langle\zeta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}} + \frac{\sin(\sqrt{\nu}t)}{\sqrt{\nu}}\,\langle\theta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}}\right)\zeta_{\nu,\kappa}\,\frac{dk}{d\nu}\,d\nu. \tag{4.12}$$

4.3. Agreement of the two expansions when $\gamma = 0$. The expression (4.12) coincides with (4.5) in the case where $\gamma = 0$. To see this, note that

$$\begin{aligned}\langle\zeta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}} &= \langle\zeta_0, (1 + \chi_P\beta\partial_x^4)\zeta_{\nu,\kappa}\rangle_{\mathbb{R}}\\ &= \langle\zeta_0, \psi_{\nu,\kappa}\rangle_{\mathbb{R}} \quad \text{with } \psi_{\nu,\kappa} = \Psi_{\nu,\kappa}\big|_{z=0}\end{aligned} \tag{4.13}$$

(where we used (4.9) with $\gamma = 0$), and similarly

$$\begin{aligned}\langle\theta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}} &= \langle\partial_n\mathbf{G}\phi_0, \psi_{\nu,\kappa}\rangle_{\mathbb{R}}\\ &= \langle\phi_0, \partial_n\mathbf{G}\psi_{\nu,\kappa}\rangle_{\mathbb{R}}\\ &= \nu\,\langle\phi_0, \zeta_{\nu,\kappa}\rangle_{\mathbb{R}}.\end{aligned} \tag{4.14}$$

Hence from (4.10), the inner product (4.6) becomes

$$\langle U_0, U_{\lambda,\kappa}\rangle_{\mathcal{V}} = \langle\zeta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}} - i\lambda^{-1}\langle\theta_0, \zeta_{\nu,\kappa}\rangle_{\mathcal{H}}.$$

Substituting this expression into (4.5) and considering the displacement only, we get

$$\zeta(t) = \frac{1}{4\pi}\int_{\mathbb{R}^+} e^{i\lambda t} \sum_{\kappa=\pm1} \left(\langle\zeta_0,\zeta_{\nu,\kappa}\rangle_{\mathcal{H}} - i\lambda^{-1}\langle\theta_0,\zeta_{\nu,\kappa}\rangle_{\mathcal{H}}\right)\zeta_{\nu,\kappa}\left|\frac{dk}{d\lambda}\right| d\lambda$$
$$+ \frac{1}{4\pi}\int_{\mathbb{R}^+} e^{-i\lambda t} \sum_{\kappa=\pm1} \left(\langle\zeta_0,\zeta_{\nu,\kappa}\rangle_{\mathcal{H}} + i\lambda^{-1}\langle\theta_0,\zeta_{\nu,\kappa}\rangle_{\mathcal{H}}\right)\zeta_{\nu,\kappa}\left|\frac{dk}{d\lambda}\right| d\lambda,$$

which is nothing but (4.12) using the change of variable $\lambda = \sqrt{\nu}$.

5. Numerical results. The numerical implementation of (4.5) and (4.11) is relatively straightforward once the eigenfunctions have been calculated by solving (4.3) and (4.9). We represent the solution to (4.3) and (4.9) at discrete points. We use this discrete representation to calculate the inner products using the expressions given by (4.6) and (4.13)–(4.14) and numerical quadrature. The final integral in (4.5) and (4.11) is calculated using the well-known FFT algorithm. The number of points in the frequency domain which are required to compute the solution depends strongly on the initial conditions; i.e., if the initial data is smooth (has few high frequency components), then fewer frequency calculations will be required. Furthermore, we require more points when the initial condition is under the plate because of the high frequencies of the free plate vibration (for our value of β) for a given initial smoothness. The case of an incident disturbance which is nonzero far from the plate also allows significant simplification, and this case is discussed below. We concentrate on making calculations similar to those in [12] for the purpose of comparison and checking. Of course the solution in [12] was valid only for shallow water, so we will extend the calculations to water of finite depth. The plate length is $b = 50$ and stiffness is $\beta = 2 \times 10^4$.

In order to compare with the results obtained in [12], we first consider a wave which is incoming from the left with initial potential given by

$$\phi_0\,(x) = e^{-(x+125)^2/350}$$

and with the corresponding displacement so that the pulse is traveling to the right in the absence of the plate. This means that the initial state $U_0 = (\phi_0, i\zeta_0)^T$ must have zero spectral components for the plane waves which propagate to the left, that is,

$$U(t) = \frac{1}{4\pi}\int_{\mathbb{R}} e^{i\lambda t}\,\langle U_0, U_{\lambda,-1}\rangle_{\mathcal{V}}\,U_{\lambda,-1}\left|\frac{dk}{d\lambda}\right| d\lambda, \quad \text{where} \quad U_0 = \begin{pmatrix}\phi_0\\ i\zeta_0\end{pmatrix}, \tag{5.1}$$

and we can deduce from the asymptotic form of the eigenfunctions that

$$\langle U_0, U_{\lambda,-1}\rangle = 2\int_{\mathbb{R}} e^{-ik(\lambda)x}\,\phi_0(x)\,dx.$$

This considerably simplifies the calculation of (5.1) and we can calculate it using the FFT of $\phi_0\,(x)$ and then an inverse FFT of (5.1).

The potential is shown in Figure 1 for water depth $h = 1$. This is identical to the equivalent figure in [12]. We now consider the effect of increasing the water depth so that the shallow plate approximation is no longer valid. Figures 2 and 3 show the evolution of the potential for the water depths $h = 5$ and $h = 20$, respectively.

In all these figures we used 512 values for λ from $-\pi$ to π to calculate $U_{\lambda,\kappa}$, and we used 4096 points (truncated with zeros) to compute the inverse FFT. Note that

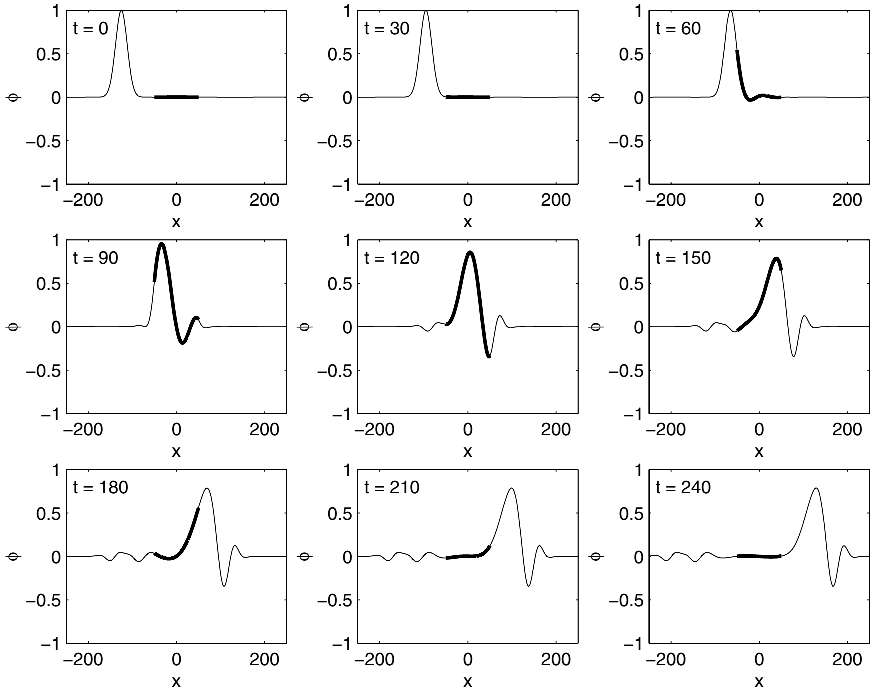


FIG. 1. *The evolution of the potential due to a pulse traveling to the right for the times shown. The plate is shown by the bold line.* $\beta = 2 \times 10^4$*,* $b = 50$*,* $\gamma = 0$*, and* $h = 1$*.*

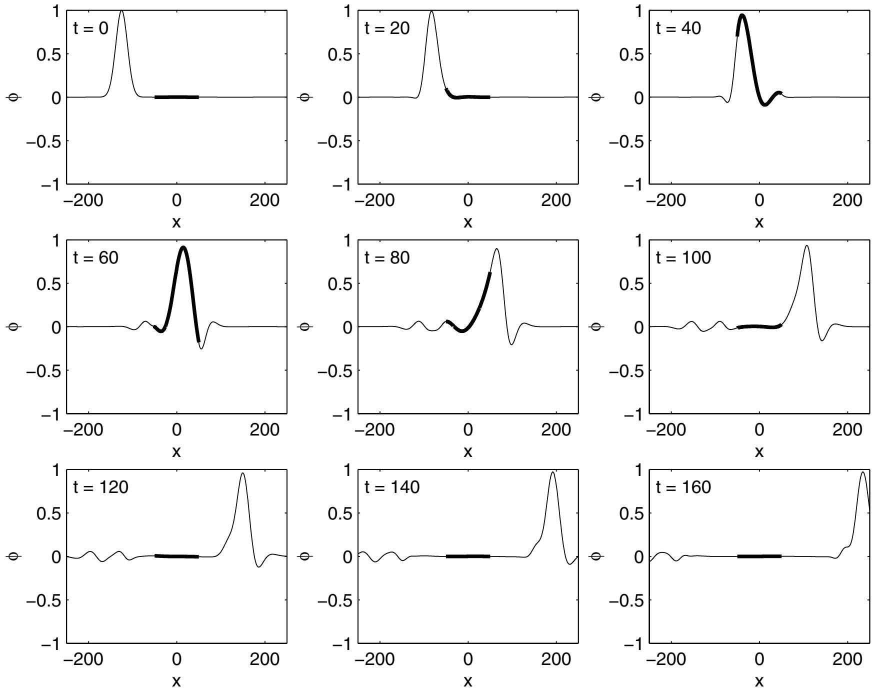


FIG. 2. *As for Figure* 1 *except that* $h = 5$*.*

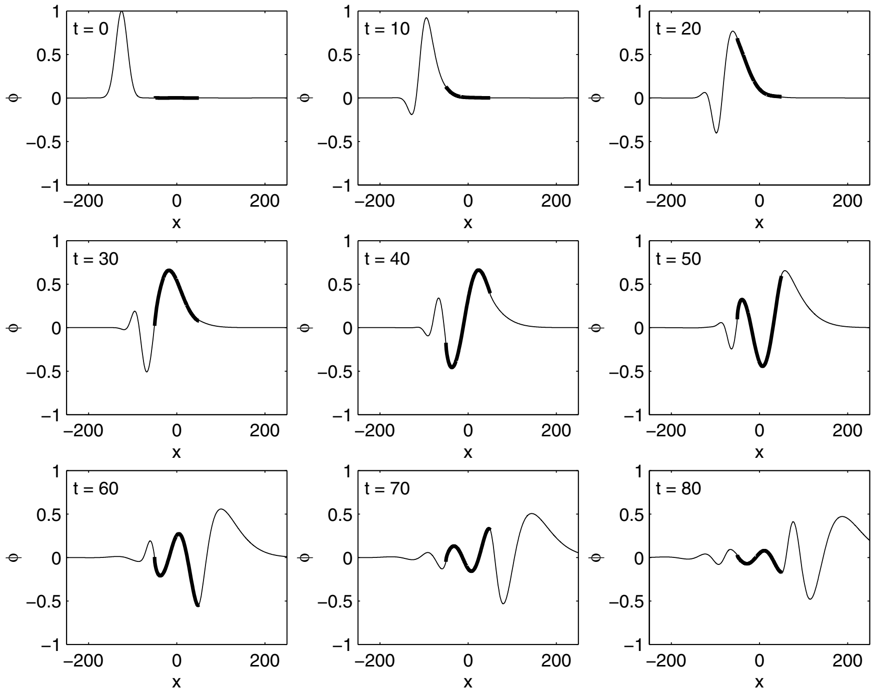


FIG. 3. *As for Figure* 1 *except that* $h = 20$.

we need to work with the variable λ to be able to use the FFT algorithm to calculate the solution in time.

We now consider the evolution of the plate released from an initial displacement. We use the same values as before; plate length is $b = 50$ and stiffness is $\beta = 2 \times 10^4$. The initial plate potential and displacement are given by

$$U_0 = \begin{pmatrix} 0 \\ ie^{-x^2/350} \end{pmatrix} \quad \text{and} \quad U_0 = \begin{pmatrix} 0 \\ ie^{-(x-50)^2/350} \end{pmatrix}$$

in Figures 4 and 5, respectively. The results for $h = 100$ agree closely with results calculated using the method of [19] valid for water of infinite depth.

Finally we investigate the effect of the parameter γ. Figure 6 shows the evolution of the symmetric displacement for $h = 5$ with $\gamma = 0$, 1, and 10. As expected, the effect of γ is small, even for the value $\gamma = 10$, which is highly unphysical.

In these figures we need to use the full expression of (4.5) and (4.11). Owing to the requirement for higher frequencies, we use 1024 values for λ spaced between -2π and 2π, and we again use 4096 points truncated with zeros for the inverse FFT.

6. Summary. We have presented the spectral theory for the two-dimensional linear wave problem of floating elastic plate on water of finite depth. Two theories, based on the first-order and second-order formulations of the problem, have been developed. The first-order theory was valid only when the plate was assumed massless, while the second-order theory allowed for the plate to have arbitrary mass. Both theories depended on different inner products in which the appropriate evolution operators were self-adjoint. The spectral theory solution was found by an expansion

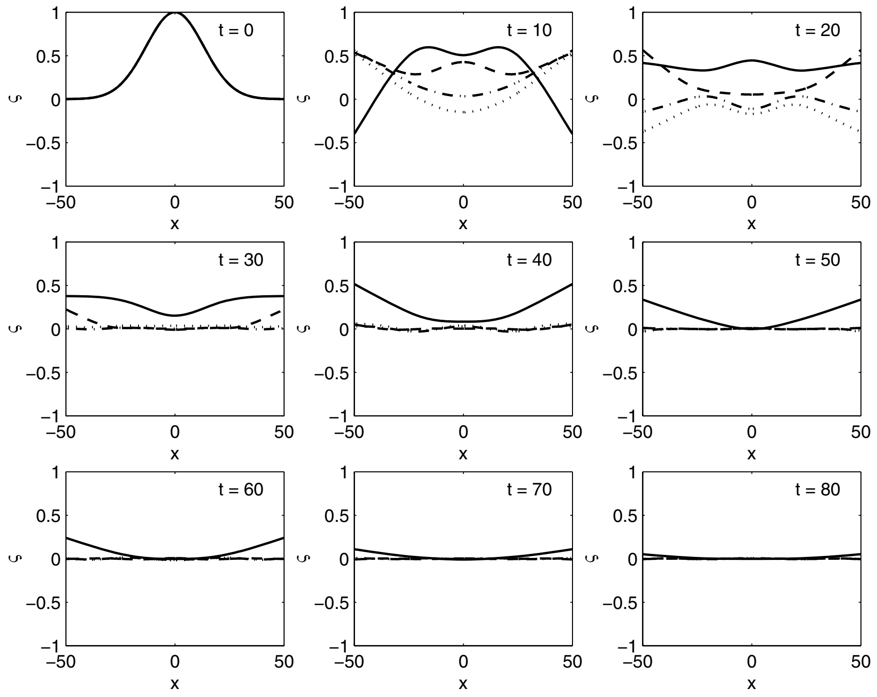


FIG. 4. *The evolution of a symmetric displacement for the times shown. $\beta = 2 \times 10^4$, $b = 50$, and $\gamma = 0$. The solid line is for $h = 1$, the dashed line for $h = 5$, the chained line for $h = 20$, and the dotted line for $h = 100$.*

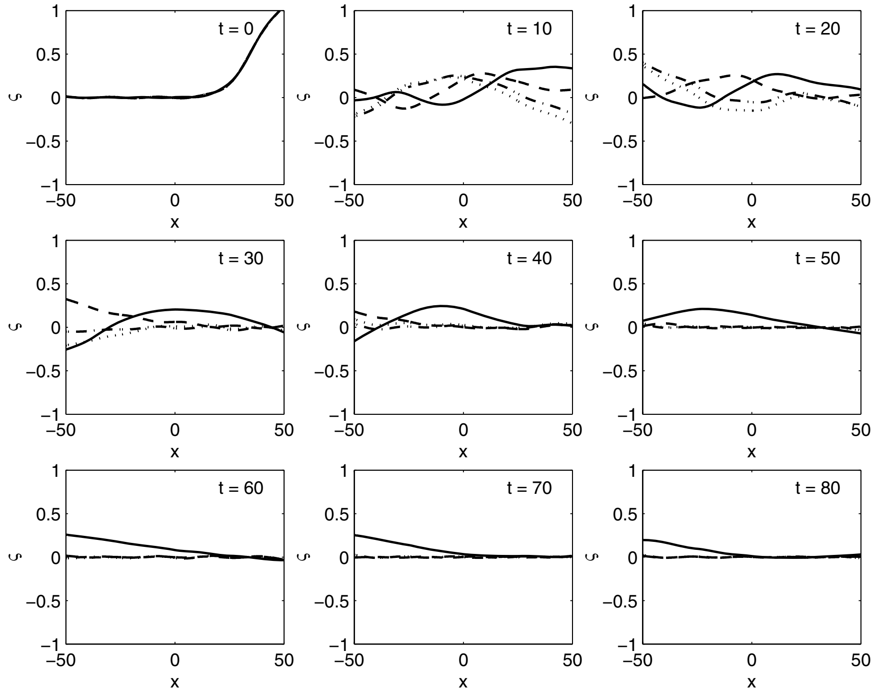


FIG. 5. *As for Figure 4 except the initial displacement is nonsymmetric.*

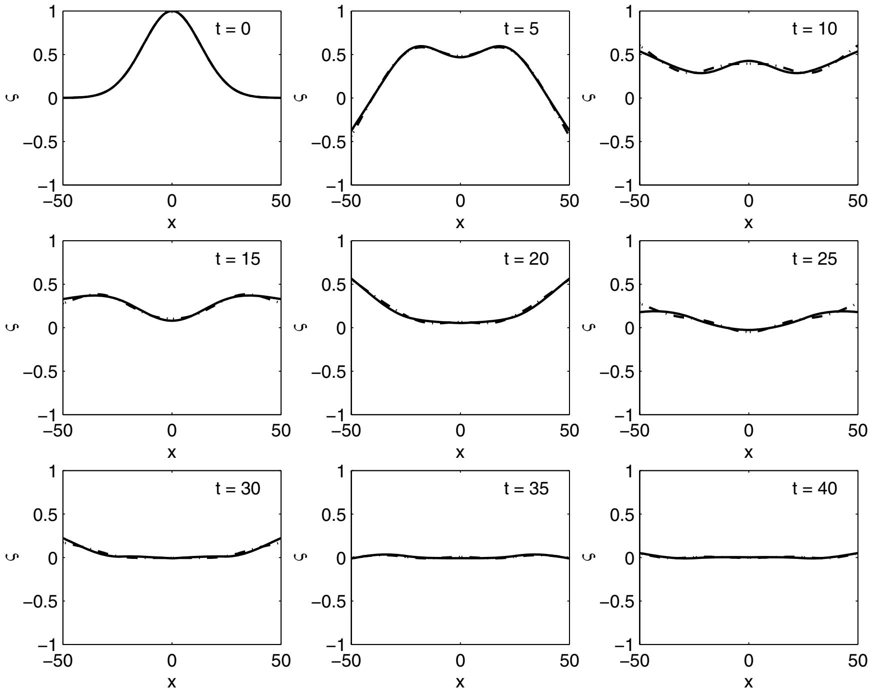


Fig. 6. *The evolution of a symmetric displacement for the times shown.* $\beta = 2 \times 10^4$*,* $b = 50$*, and* $h = 5$*. The solid line is for* $\gamma = 0$*, the dashed line for* $\gamma = 1$*, and the chained line for* $\gamma = 10$*.*

in the time-harmonic (single frequency solutions). We presented solutions for some simple forcings and showed that the solutions agreed with those found earlier by [12] and showed the effect of increasing the water depth and the mass on the solutions.

Appendix. Solution method to find the eigenfunctions. In this appendix the solution method used to calculate the single frequency solution is described. For the high values of the plate stiffness used here we require very high frequency solutions, and the numerical solution is challenging. We used a mode matching method based on [9] where the solution was found for a semi-infinite plate. However, unlike [9], where the equations were solved by the residue calculus method, our equations are solved by matrix inversion. This method was used recently in [8] to solve for multiple elastic plates. We will show the method only for the case of a wave which is incident from the right. We express the solution as

$$\text{(A.1)}\quad \Phi \text{ or } \Psi = \begin{cases} e^{ikx}\cosh k_0\left(z+h\right) + \displaystyle\sum_{n=0}^{\infty} a_n e^{ik_n x}\cosh k_n\left(z+h\right), & x > b, \\ \displaystyle\sum_{n=-2}^{\infty} b_n e^{i\kappa_n x}\cosh \kappa_n\left(z+h\right) + c_n e^{-i\kappa_n x}\cosh \kappa_n\left(z+h\right), & -b < x < b, \\ \displaystyle\sum_{n=0}^{\infty} d_n e^{-ik_n x}\cosh k_n\left(z+h\right), & x < -b. \end{cases}$$

In (A.1), k_0 is the negative real solution to the dispersion equation

$$k\tanh kh = \nu = \lambda^2 \tag{A.2}$$

(so $k_0 = -k(\nu)$), and k_n for $n > 0$ are the positive imaginary solutions to (A.2) ordered by increasing imaginary part. Under the plate there is a new dispersion equation given by

$$\left(\beta\kappa^5 + (1-\gamma\lambda^2)\right)\tanh \kappa h = \lambda^2. \tag{A.3}$$

We define κ_0 to be the negative real solution to (A.3), κ_n for $n > 0$ are the positive imaginary solutions to (A.3) ordered by increasing imaginary part, and κ_{-2} and κ_{-1} are the complex solutions with positive imaginary parts. This dispersion equation is discussed in detail in [2]. We should note that $a_0 = S_{11}$ and $d_0 = S_{12}$.

We calculate the solution numerically by matching these solutions at $x = \pm b$ and by imposing the edge conditions. First we truncate the sum at N, which gives a system with $4N + 8$ unknowns. We will obtain $4N + 4$ equations by matching the potential at $x = \pm b$ and taking the trivial inner product with respect to the N vertical eigenfunctions outside the plate, matching both the potential and its derivative. The final four equations will come from the free edge conditions.

If we match the potential at $x = b$ and take inner products with respect to the vertical eigenfunctions outside the plate, we obtain

$$d_{00}e^{ikb} + \mathbf{Da} = \mathbf{Mb} + \mathbf{Nc},$$

where $\mathbf{D}$ is the diagonal matrix (diagonal because the vertical eigenfunctions are orthogonal with respect to the trivial inner product) with entries

$$d_{ii} = \int_{-h}^{0} e^{ik_i b}\cosh^2 k_i\,(z+h)\,dz, \quad 0 \le i \le N,$$

and the entries of the matrices $\mathbf{M}$ and $\mathbf{N}$ (which are not square) are given by

$$m_{ij} = \int_{-h}^{0} e^{i\kappa_i b}\cosh\kappa_j\,(z+h)\cosh k_i\,(z+h)\,dz, \quad 0 \le i \le N, \quad -2 \le j \le N,$$

and

$$n_{ij} = \int_{-h}^{0} e^{-i\kappa_i b}\cosh\kappa_j\,(z+h)\cosh k_i\,(z+h)\,dz, \quad 0 \le i \le N, \quad -2 \le j \le N,$$

and $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the vectors whose entries are a_n, b_n, and c_n, respectively. We use a nonstandard numbering of the matrix entries which corresponds to the numbering of the roots of the dispersion equation. Obviously the integrals above can be found analytically, but the expressions for these are not included here. There are also ways of improving the stability of the matrix equations by an appropriate normalization, and these issues are discussed in [9]. We then match the derivative of the potential at $x = b$ and again take an inner product with respect to the vertical eigenfunctions outside the plate region, obtaining

$$ik\hat{d}_{00}e^{ikb} + \hat{\mathbf{D}}\mathbf{a} = \hat{\mathbf{M}}\mathbf{b} + \hat{\mathbf{N}}\mathbf{c},$$

where $\hat{\mathbf{D}}$ is the diagonal matrix with entries

$$\hat{d}_{ii} = \int_{-h}^{0} ik_i e^{ik_i b}\cosh^2 k_i\,(z+h)\,dz, \quad 0 \le i \le N,$$

and the entries of the matrices $\hat{\mathbf{M}}$ and $\hat{\mathbf{N}}$ are given by

$$\hat{m}_{ij} = \int_{-h}^{0} i\kappa_i e^{i\kappa_i b} \cosh \kappa_j (z+h) \cosh k_i (z+h) \, dz, \quad 0 \le i \le N, \quad -2 \le j \le N,$$

and

$$\hat{n}_{ij} = \int_{-h}^{0} -i\kappa_i e^{-i\kappa_i b} \cosh \kappa_j (z+h) \cosh k_i (z+h) \, dz, \quad 0 \le i \le N, \quad -2 \le j \le N.$$

A similar set of equations is derived by matching and taking an inner product at $x = -b$. The final four equations are obtained by imposing the free edge conditions, for example the conditions that

$$\lim_{x \uparrow b} \partial_x^2 \zeta = 0$$

implies that

$$\sum_{n=-2}^{N} \kappa_n^3 b_n e^{i\kappa_n b} \sinh \kappa_n h + \kappa_n^3 c_n e^{-i\kappa_n b} \sinh \kappa_n h = 0.$$

Some further simplifications are possible, symmetry arguments can be used to reduce the number of unknowns by a factor of two, and we can also use a wide spacing approximation if the frequency is high enough. These are not described here for the sake of brevity.

REFERENCES

[1] J. T. Beale, *Eigenfunction expansions for objects floating in an open sea*, Comm. Pure Appl. Math., 30 (1977), pp. 283–313.

[2] C. Fox and V. A. Squire, *On the oblique reflexion and transmission of ocean waves at shore fast sea ice*, Philos. Trans. Roy. Soc. London Ser. A, 347 (1994), pp. 185–218.

[3] C. Hazard and M. Lenoir, *Surface water waves*, in Scattering, R. Pike and P. Sabatier, eds., Academic Press, San Diego, CA, 2002, pp. 618–636.

[4] C. Hazard and F. Loret, *Generalized eigenfunction expansions for scattering problems with an application to water waves*, Proc. Roy. Soc. Edinburgh Sect. A, 137 (2007), pp. 995–1035.

[5] T. Ikebe, *Eigenfunction expansions associated with the Schroedinger operators and their applications to scattering theory*, Arch. Rational Mech. Anal., 5 (1960), pp. 1–34.

[6] M. Kashiwagi, *Research on hydroelastic response of VLFS: Recent progress and future work*, Int. J. Offshore and Polar Engineering, 10 (2000), pp. 81–90.

[7] M. Kashiwagi, *A time-domain mode-expansion method for calculating transient elastic responses of a pontoon-type VLFS*, J. Marine Sci. Tech., 5 (2000), pp. 89–100.

[8] A. Kohout, M. H Meylan, S. Sakai, K. Hanai, P. Leman, and D. Brossard, *Linear water wave propagation through multiple floating elastic plates of variable properties*, J. Fluids Structures, 23 (2007), pp. 649–663.

[9] C. M. Linton and H. Chung, *Reflection and transmission at the ocean/sea-ice boundary*, Wave Motion, 38 (2003), pp. 43–52.

[10] P. McIver, M. McIver, and J. Zhang, *Excitation of trapped water waves by the forced motion of structures*, J. Fluid Mech., 494 (2003), pp. 141–162.

[11] C. C. Mei, *The Applied Dynamics of Ocean Surface Waves*, World Scientific, River Edge, NJ, 1989.

[12] M. H Meylan, *Spectral solution of time dependent shallow water hydroelasticity*, J. Fluid Mech., 454 (2002), pp. 387–402.

[13] M. H. Meylan and V. A. Squire, *The response of ice floes to ocean waves*, J. Geophys. Res., 99 (1994), pp. 891–900.

[14] Y. Namba and M. Ohkusu, *Hydroelastic behaviour of floating artificial islands in waves*, Int. J. Offshore and Polar Engineering, 9 (1999), pp. 39–47.

[15] J. N. Newman, *Wave effects on deformable bodies*, Appl. Ocean Res., 16 (1994), pp. 45–101.

[16] A. Ya. Povzner, *On the expansions of arbitrary functions in terms of the eigenfunctions of the operator* $-\Delta u + cu$, Mat. Sbornik N.S., 32 (1953), pp. 109–156 (in Russian).

[17] V. A. Squire, J. P. Dugan, P. Wadhams, P. J. Rottier, and A. J. Liu, *Of ocean waves and sea ice*, in Annual Review of Fluid Mechanics, Vol. 27, Annual Reviews, Palo Alto, CA, 1995, pp. 115–168.

[18] J. J. Stoker, *Water Waves: The Mathematical Theory with Applications*, Interscience, New York, 1957.

[19] I .V. Sturova, *Unsteady behavior of an elastic beam floating on the surface of an infinitely deep fluid*, J. Appl. Mech. Tech. Phys., 47 (2006), pp. 71–78.

[20] M. Vullierme-Ledard, *The limiting amplitude principle applied to the motion of floating bodies*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 125–170.

[21] E. Watanabe, T. Utsunomiya, and C. M. Wang, *Hydroelastic analysis of pontoon-type VLFS: A literature survey*, Eng. Struct., 26 (2004), pp. 245–256.

[22] J. V. Wehausen and E. V. Laitone, *Surface waves*, in Fluid Dynamics III, S. Flügge and C. Truesdell, eds., Handbuch der Physik 9, Springer-Verlag, New York, 1960, pp. 446–778.

[23] C. H. Wilcox, *Scattering Theory for the d'Alembert Equation in Exterior Domains*, Springer-Verlag, New York, 1975.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 648–665

HYDRODYNAMIC LIMIT OF A FOKKER–PLANCK EQUATION DESCRIBING FIBER LAY-DOWN PROCESSES[*]

L. L. BONILLA[†], T. GÖTZ[‡], A. KLAR[§], N. MARHEINEKE[‡], AND R. WEGENER[¶]

Abstract. In this paper, a stochastic model for the turbulent fiber lay-down in the industrial production of nonwoven materials is extended by including a moving conveyor belt. In the hydrodynamic limit corresponding to large noise values, the transient and stationary joint probability distributions are determined using the method of multiple scales and the Chapman–Enskog method. Moreover, exponential convergence towards the stationary solution is proven for the reduced problem. For special choices of the industrial parameters, the stochastic limit process is an Ornstein–Uhlenbeck process. It is a good approximation of the fiber motion even for moderate noise values. Moreover, as shown by Monte-Carlo simulations, the limiting process can be used to assess the quality of nonwoven materials in the industrial application by determining distributions of functionals of the process.

Key words. stochastic differential equations, Fokker–Planck equation, asymptotic expansion, Ornstein–Uhlenbeck process

AMS subject classifications. 37H10, 34E13, 60H30, 65C05

DOI. 10.1137/070692728

1. Introduction.

Nonwoven materials/fleece are webs of long flexible fibers that are used for composite materials (filters) as well as in the hygiene and textile industries. They are produced in melt-spinning operations: hundreds of individual endless fibers are obtained by the continuous extrusion of a molten polymer through narrow nozzles that are densely and equidistantly placed in a row at a spinning beam. The viscous/viscoelastic fibers are stretched and spun until they solidify due to cooling air streams. Before the elastic fibers lay down on a moving conveyor belt to form a web, they become entangled and form loops due to the highly turbulent air flows. The homogeneity and load capacity of the fiber web are the most important textile properties for quality assessment of industrial nonwoven fabrics. The optimization and control of the fleece quality require modeling and simulation of fiber dynamics and lay-down; in addition, it is necessary to determine the distribution of fiber mass and directional arrangement in the web.

The software FIDYST, developed on the basis of the mathematical model of [9] at the Fraunhofer ITWM, Kaiserslautern, enables numerical simulation of the spinning and deposition regime in the nonwoven production processes; cf. Figure 1. The interaction of the fiber with the turbulent air flows is described by a stochastic force in the momentum equation, which is derived, analyzed, and experimentally validated in [11, 12]. The resulting force model depends on the flow velocity which is split into mean and random parts following Reynolds' idea for the averaged Navier–Stokes

[*] Received by the editors May 22, 2007; accepted for publication (in revised form) September 19, 2007; published electronically December 12, 2007. The authors acknowledge support from the Rheinland–Pfalz Excellence Cluster "Dependable Adaptive Systems and Mathematical Modelling."
http://www.siam.org/journals/siap/68-3/69272.html

[†] G. Millán Institute for Modeling, Simulation and Industrial Mathematics, Universidad Carlos III Madrid, Leganes 28911, Spain (bonilla@ing.uc3m.es).

[‡] Fachbereich Mathematik, Technische Universität Kaiserslautern, 67653 Kaiserslautern, Germany (goetz@mathematik.uni-kl.de, marheineke@mathematik.uni-kl.de).

[§] Fachbereich Mathematik, Technische Universität Kaiserslautern, 67653 Kaiserslautern, Germany, and Fraunhofer ITWM, 67663 Kaiserslautern, Germany (klar@itwm.fhg.de).

[¶] Fraunhofer ITWM, 67663 Kaiserslautern, Germany (wegener@itwm.fhg.de).

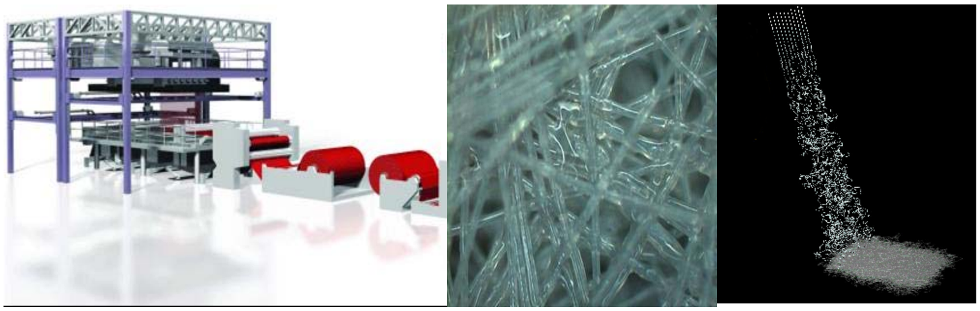


FIG. 1. *Production of nonwoven materials. Left to right: plant and fleece (Neumag, www.neumag.saurer.com), simulated process (computation by Fraunhofer ITWM (FIDYST), and visualization by Fraunhofer IGD).*

equations. The random force is modeled as white noise with a fluctuation-dependent amplitude that carries information of the kinetic turbulent energy, dissipation rate, and correlation lengths. Due to the huge amount of physical details incorporated in FIDYST, the simulations of the fiber spinning and lay-down usually require an extremely large computational effort and high memory storage. Hence, the optimization and control of the full process, and particularly of fleece quality, are difficult. Thus, a simplified stochastic model for the fiber lay-down process is presented in [5]. Under the assumption of a nonmoving conveyor belt, this model describes the position of the fiber on the transport belt by a stochastic differential system containing parameters that characterize the process. For example, the effect of air turbulence has to be identified from the full model and adapted to be used in the reduced one. Parameter identification can be obtained from a FIDYST-simulation of a single, relatively short fiber whose computation time is short even using the more complex model. Then, the reduced model can be used to calculate fast and efficiently the performance of hundreds of long fibers for fleece production. In [5] the associated Fokker–Planck equation and stationary solution are investigated for the case of a nonmoving conveyor belt. In this case, the model without noise is conservative and its equations, Hamiltonian. For small turbulence noise, stochastic averaging can be used to derive a stochastic equation for the energy and related functionals of the stochastic process. Moreover, their distributions can be analyzed. An analytic investigation of the corresponding Fokker–Planck equation has been performed in [7], ergodicity of the process has been proven, and explicit rates for the convergence to the stationary solution have been obtained.

In this paper, we extend the stochastic model of [5] to a more realistic fiber lay-down model with a moving transport belt; cf. section 2. In this case, the model equations are no longer Hamiltonian for zero noise. For both moving and nonmoving conveyor belts, we consider the case of large turbulence noise, $A \to \infty$, in which the probability density of the fiber becomes rapidly independent of the angle between the fiber and the direction of the conveyor's motion and the angle between the fiber and the position vector of its tip, respectively. In the case of a nonmoving belt, section 3 describes how to use the method of multiple scales in order to determine explicitly a reduced Smoluchowski equation for the fiber probability density, the stationary distribution, and the transient joint probability distributions, all from the associated Fokker–Planck equation. For a moving belt, the same magnitudes are determined using the Chapman–Enskog method [4, 1] in section 4. To leading order, the stationary distributions are of Gaussian type; in particular, for special choices of the process parameters, Ornstein–Uhlenbeck processes turn out to be the limit solutions. In section 5, exponential convergence towards the stationary solution of the reduced Fokker–Planck equation is proved by classical arguments. The numerical results in section 6 show that direct Monte-Carlo simulations of the fiber process agree quite well with the theoretical results even for moderate values of the noise strength A. In addition, certain functionals of the fiber (i.e., mass distributions) are essential for the quality assessment of nonwoven materials. We compare their distributions with the corresponding functionals for the limiting Ornstein–Uhlenbeck process.

2. The model. Consider a slender, elastic, nonextensible, and endless fiber in a lay-down regime. Let the fiber be produced with the spinning speed v_{spin}, excited into motion by a surrounding highly turbulent air flow and laid down on a conveyor belt moving with the velocity v_{belt}. Due to its slenderness, the fiber laid on the two-dimensional transport belt is described as a curve $\eta : \mathbb{R}_0^+ \to \mathbb{R}^2$. Choosing arc-length parameterization, the nonextensibility condition $\|d\eta/dt\| = 1$ holds by setting

$$d\eta = (\cos\alpha, \sin\alpha)\, dt,$$

where α denotes the angle of the fiber relative to the direction of motion e_1 of the transport belt. The reference point of the spinning process determined by the position of the nozzle moves in the coordinate system of the transport belt in the direction $-e_1$. Thus,

$$\xi(t) = \eta(t) - (-\kappa t e_1)$$

describes the deviation of the fiber from the reference point as a function of the arc-length parameter t, where $\kappa = v_{belt}/v_{spin} \in [0, 1]$ is the ratio between the belt and spinning speeds. Generalizing [5], we model (ξ, α) by the following stochastic differential system:

$$d\xi_1 = (\cos\alpha + \kappa)\, dt, \tag{2.1a}$$

$$d\xi_2 = \sin\alpha\, dt, \tag{2.1b}$$

$$d\alpha = c(\xi)\,(\xi_1 \sin\alpha - \xi_2 \cos\alpha)\, dt + A\, dW_t. \tag{2.1c}$$

Here, the change of the angle α is characterized by the deterministic buckling/coiling c of the fiber (which tends to turn it back to its reference point) and by the random fluctuations $A\, dW_t$ due to the interaction of the fiber with the external turbulent air flow; W denotes a one-dimensional Wiener process.

REMARK 2.1. *The general deterministic coiling behavior of flexible fibers has been studied, for example, in* [10, 8]. *The function c in our model prescribes its amplitude that depends on the lay-down process. c is a scalar-valued function for isotropic processes and a matrix-valued one for anisotropic processes* [5]. *For reasons that will become clear later on (cf.* (4.9)*), physically reasonable solutions can be expected only if* $\exp(-B(\xi) - k\xi_1)$ *is integrable for $k \in \mathbb{R}$, where $\partial_{\xi_i} B(\xi) = c(\xi)\xi_i$. A typical example satisfying this condition is $c(\xi) = 1$ since then $B(\xi) = (\xi_1^2 + \xi_2^2)/2$.*

REMARK 2.2. *The isotropic model considered here can be treated as dimensionless with $c(e_1) = 1$, for anisotropic lay-down processes with $1/2\,\mathrm{tr}(c(e_1)) = 1$. This corresponds to a scaled throwing (lay-down) range of order one. Consequently, the noise amplitude A characterizes the relation between stochastic and deterministic rates in the behavior of the system.*

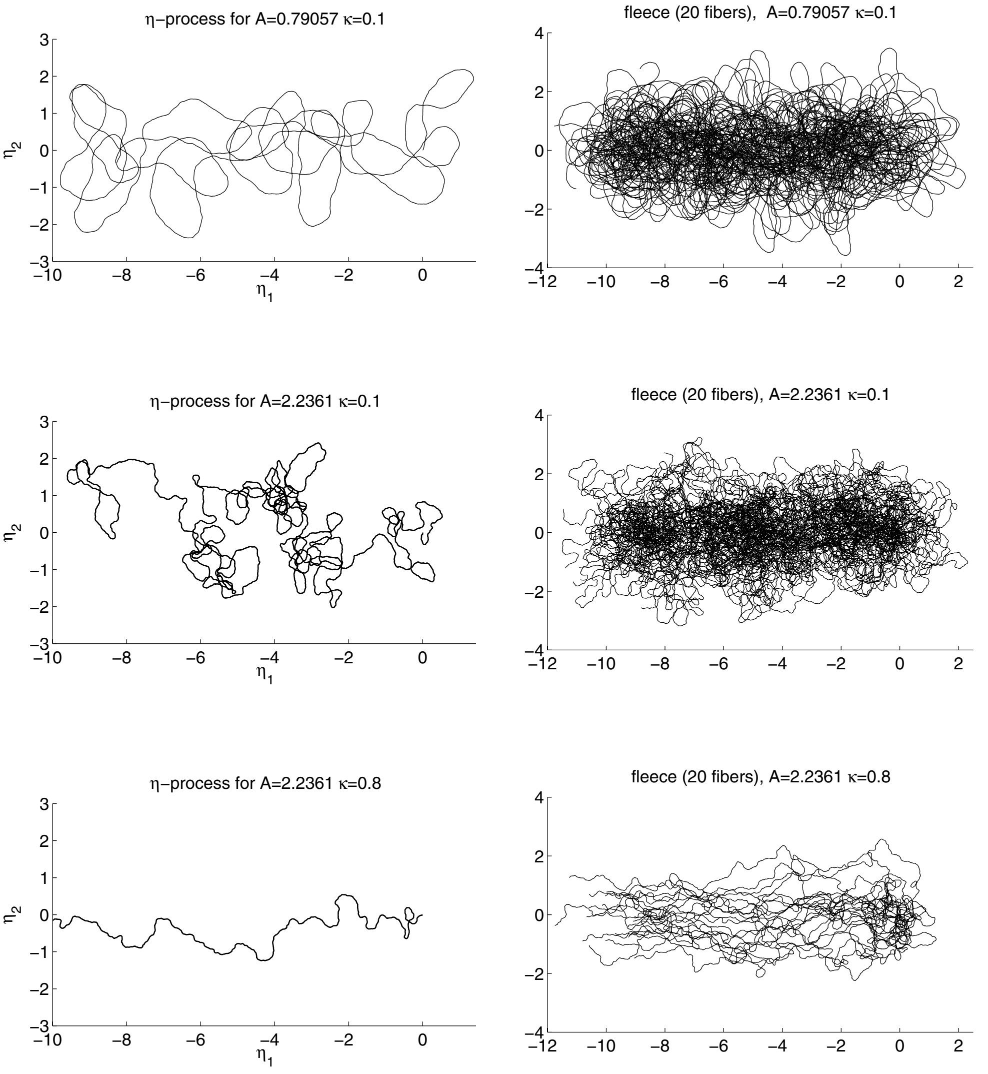


FIG. 2. *Left: η-path. Right: Associated fleece (*20 *fibers). Top to bottom:* $(A, \kappa) = \{(0.79, 0.1), (2.23, 0.1), (2.23, 0.8)\}$.

To illustrate our previous considerations, realizations of the processes η and ξ are exemplified in Figures 2 (left) and 3, respectively, where the parameters (A, κ) are selected in the set $(A, \kappa) = \{(0.79, 0.1), (2.23, 0.1), (2.23, 0.8)\}$, and $c(\xi) = 1$ is fixed. Superposing many fibers, i.e., η-paths, generates a nonwoven material whose properties depend on the industrial control parameters A, κ, and c; see Figure 2 (right) for 20 fibers. In this figure, the distance between two neighboring spinning nozzles is $d_{spin} = 2.5 \cdot 10^{-3}$, fleece length is $L_{fleece} = 10$, and fiber length is $T = L_{fleece}/\kappa$. For $\kappa \to 1$ the belt velocity coincides with the spinning speed such that the fibers lay down almost straight independent of turbulence noise. The smaller κ is, the more fiber material (length) can become entangled and form loops. The size of the

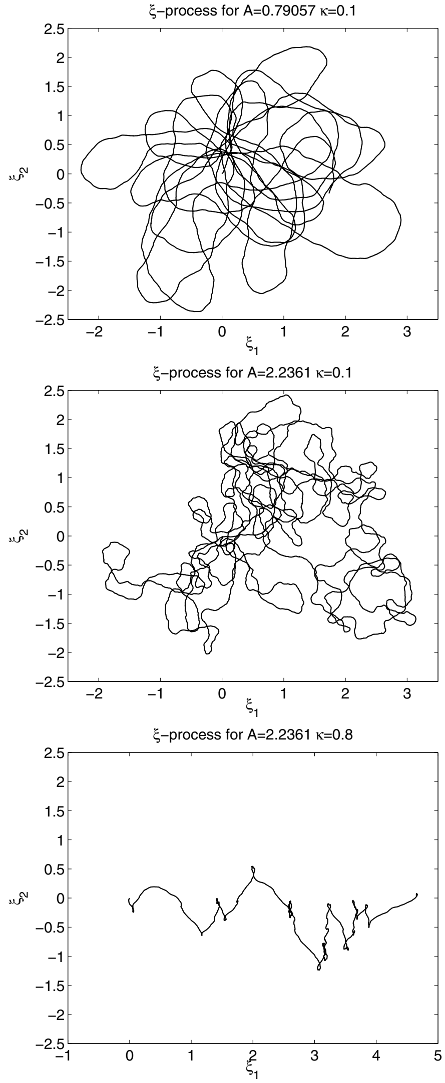


FIG. 3. *ξ-path, corresponding to Figure* 2. *Top to bottom:* $(A, \kappa) = \{(0.79, 0.1), (2.23, 0.1), (2.23, 0.8)\}$.

loops is thereby determined by the amplitude of the turbulence noise A. For small A the deterministic coiling/buckling radius dominates the fiber behavior, whereas a finer entanglement on various scales arises for large A. For the industrial application, nonwoven materials with a homogeneous distribution of mass and fiber orientation are desirable, and they typically have these characteristics for small κ and larger A. To get a deeper insight into the probability density of the underlying ξ-process (2.1), $p = p(\xi_1, \xi_2, \alpha, t)$, we consider its associated Fokker–Planck equation

$$(2.2) \quad \partial_t p + (\cos\alpha + \kappa)\, \partial_{\xi_1} p + \sin\alpha \partial_{\xi_2} p - \partial_\alpha \left[c(\xi)(-\xi_1 \sin\alpha + \xi_2 \cos\alpha) p\right] = \frac{A^2}{2} \partial_\alpha^2 p.$$

REMARK 2.3. *In the case of a nonmoving conveyor belt ($\kappa = 0$), the processes η*

and ξ coincide. Then, it is advantageous to introduce polar coordinates $\xi_1 = r\cos\varphi$, $\xi_2 = r\sin\varphi$, and $\beta = \alpha - \varphi$, and to define $b(r) = \|\xi\|\, c(\|\xi\|)$ as done in [5]. *The resulting system then reduces to two dimensions and the associated Fokker–Planck equation for (r,β) reads*

$$\partial_t p + \cos\beta\partial_r p + \left(b(r) - \frac{1}{r}\right)\partial_\beta\left(p\sin\beta\right) = \frac{A^2}{2}\partial_\beta^2 p. \tag{2.3}$$

In the following we determine the evolution and the stationary solution of the Fokker–Planck equations (2.2), (2.3) in the limit as $A \to \infty$. Note, since we embed our model in the context of dynamical systems and stochastic processes, we refer occasionally to the notation and interpretation of time for the fiber arc-length t.

3. The nonmoving conveyor belt. We start our investigation with the case of a nonmoving belt. This case is quite instructive and allows us to introduce the main ideas to also tackle the case of a moving belt. Let $\varepsilon = 1/A^2 \ll 1$. As already mentioned above, we introduce polar coordinates and obtain the Fokker–Planck equation

$$\partial_t p + \cos\beta\partial_r p + \left(b(r) - \frac{1}{r}\right)\partial_\beta\left(p\sin\beta\right) = \frac{1}{2\varepsilon}\partial_\beta^2 p \tag{3.1a}$$

for the density distribution $p(r,\beta,t)$ subject to the normalization condition

$$\int_{\mathbb{R}_+\times[-\pi,\pi]} p(r,\beta,t)\,dr\,d\beta = 1 \tag{3.1b}$$

and the initial condition

$$p(r,\beta,0) = p_0(r,\beta). \tag{3.1c}$$

Note that the stochastic term appears only in the angular coordinate. Hence, for dominating stochastic forcing, i.e., $\varepsilon \ll 1$, we expect a fast averaging over the β-coordinate. Dominant balance between diffusion and the time derivative of p implies a fast time scale $\tau = t/\varepsilon$. The relaxation to the stationary distribution will take much longer.

To capture the fast averaging over β and the slower convergence to the stationary solution, we use the method of multiple scales. Let us introduce two time scales: the fast scale $\tau = t/\varepsilon$ and a slow scale $T = \varepsilon t$. For the distribution function $p = p(r,\beta,t;\varepsilon)$ (which is 2π-periodic in β), we propose the following ansatz:

$$p = p^{(0)}(r,\beta,\tau,T) + \varepsilon p^{(1)}(r,\beta,\tau,T) + \varepsilon^2 p^{(2)}(r,\beta,\tau,T) + \cdots. \tag{3.2}$$

Inserting (3.2) into (3.1) and equating equal powers of ϵ in the resulting equations, we obtain a hierarchy of problems for the $p^{(m)}$. As we shall see, secular terms appear only in the equation for $p^{(2)}$, and their elimination requires the introduction of the slow scale $T = \epsilon t$. To leading order, we have to solve

$$Lp^{(0)} = 0, \tag{3.3a}$$

$$\int_{\mathbb{R}_+\times[-\pi,\pi]} p^{(0)}\,dr\,d\beta = 1, \tag{3.3b}$$

$$p^{(0)}(r,\beta,0,0) = p_0(r,\beta), \tag{3.3c}$$

where $L = \partial_\tau - \partial_\beta^2/2$ denotes the diffusion operator in the angular direction. Solving the parabolic equation (3.3a) yields

$$p^{(0)}(r,\beta,\tau,T) = \frac{1}{2\pi}\mathcal{P}(r,T) + \sum_{j\in\mathbb{Z}\setminus\{0\}} e^{ij\beta - j^2\tau/2} C_j(r), \tag{3.4a}$$

where

$$C_j(r) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ij\beta} p_0(r,\beta)\, d\beta \tag{3.4b}$$

and

$$\mathcal{P}(r,0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} p_0(r,\beta)\, d\beta \tag{3.4c}$$

are the Fourier coefficients of the initial condition.

In the case of a rotationally symmetric initial distribution $p_0 = p_0(r)$, all the coefficients C_j vanish identically. If the initial distribution is not symmetric, then the angular components $C_j e^{ij\beta - j^2\tau/2}$ are exponentially decaying with τ, i.e., the angular dependence of p is averaged out on the fast time scale τ. The relaxation to the stationary solution is determined by the behavior of $\mathcal{P}(r,T)$ on the long time scale T. Therefore, we will neglect the exponentially small terms $C_j e^{ij\beta - j^2\tau/2}$ in the following.

To determine the stationary solution $\mathcal{P}$, we proceed with the next terms of the expansion (3.2). The $\mathcal{O}(\varepsilon)$-problem reads as

$$\begin{aligned} Lp^{(1)} &= -\frac{\cos\beta}{2\pi}\left[\partial_r\mathcal{P} + \left(b(r) - \frac{1}{r}\right)\mathcal{P}\right], \\ \int_{\mathbb{R}_+\times[-\pi,\pi]} p^{(1)}\, dr\, d\beta &= 0, \\ p^{(1)}(r,\beta,0,0) &= 0. \end{aligned}$$

Again, solving the above parabolic problem yields

$$p^{(1)} = \frac{\mathcal{A}(r,T)}{2\pi} - \frac{\cos\beta}{\pi}\left[\partial_r\mathcal{P}(r,T) + \left(b - \frac{1}{r}\right)\mathcal{P}(r,T)\right],$$

where $\mathcal{A}(r,T)$ is a solution of the homogeneous problem, $L\mathcal{A} = 0$, such that $\int_0^\infty \mathcal{A}(r,T)\, dr = 0$ (normalization condition). At this order, we have two functions, $\mathcal{P}$ and $\mathcal{A}$, not yet determined. Hence, we proceed to the second order

$$\begin{aligned} Lp^{(2)} &= \frac{\cos^2\beta}{\pi}\partial_r\left[\partial_r\mathcal{P} + \left(b - \frac{1}{r}\right)\mathcal{P}\right] \\ &\quad + \left(b - \frac{1}{r}\right)\left[\partial_r\mathcal{P} + \left(b - \frac{1}{r}\right)\mathcal{P}\right]\partial_\beta\frac{\sin\beta\cos\beta}{\pi} - \frac{1}{2\pi}\partial_T\mathcal{P} \\ &\quad - \frac{\cos\beta}{2\pi}\left[\partial_r\mathcal{A} + \left(b(r) - \frac{1}{r}\right)\mathcal{A}\right] \\ &= \frac{1+\cos 2\beta}{2\pi}\partial_r\left[\partial_r\mathcal{P} + \left(b - \frac{1}{r}\right)\mathcal{P}\right] - \frac{1}{2\pi}\partial_T\mathcal{P} \end{aligned}$$

$$+\left(b-\frac{1}{r}\right)\left[\partial_r\mathcal{P}+\left(b-\frac{1}{r}\right)\mathcal{P}\right]\frac{\cos 2\beta}{\pi}$$
$$-\frac{\cos\beta}{2\pi}\left[\partial_r\mathcal{A}+\left(b(r)-\frac{1}{r}\right)\mathcal{A}\right].$$

To ensure the boundedness of $p^{(2)}$, the average of the right-hand side of the preceding equation over β should vanish. Otherwise a secular term proportional to T would be part of the solution $p^{(2)}$. This solvability condition yields

$$\partial_T\mathcal{P}=\partial_r\left[\partial_r\mathcal{P}+\left(b(r)-\frac{1}{r}\right)\mathcal{P}\right], \tag{3.5a}$$

where $\mathcal{P}$ also satisfies the normalization condition

$$\int_{\mathbb{R}_+}\mathcal{P}(r,T)\,dr=1, \tag{3.5b}$$

the initial condition (3.4c), and

$$\mathcal{P}(0,T)=\mathcal{P}(\infty,T)=0. \tag{3.5c}$$

Equation (3.5) is the reduced Fokker–Planck (Smoluchowski) equation, which determines the leading order approximation to the solution of the system (3.1), in the limit as $\varepsilon\to 0$, i.e., for dominating stochastic forcing.

The stationary solution $\mathcal{P}_s(r)$ satisfying (3.5) is given by

$$\mathcal{P}_s(r)=k\,re^{-B(r)}, \tag{3.6}$$

where $B'(r)=b(r)$ and k is the normalization constant. Note that $\mathcal{P}_s(r)$ is independent of the noise strength A, and is also the stationary solution of the full Fokker–Planck equation (3.1). The limiting stochastic differential equation (SDE) associated to (3.5) reads

$$dr=-\left(b(r)-\frac{1}{r}\right)dT+\sqrt{2}\,dW_T.$$

Remark 3.1. *In the generic case $b(r)=r$, we obtain $B(r)=r^2/2$ and the stationary solution*

$$\mathcal{P}_s(r)=re^{-r^2/2},$$

which is a rotational symmetric Gaussian distribution centered at the origin with variance 1. *The solution of its associated SDE*

$$dr=-\left(r-\frac{1}{r}\right)dT+\sqrt{2}\,dW_T$$

is a radially symmetric Ornstein–Uhlenbeck process. This can be concluded from the Fokker–Planck equation of the reduced process (3.5). *Defining the function $\tilde{\mathcal{P}}(\xi)=\mathcal{P}(r)/r$ for $\xi=(\xi_1,\xi_2)$ and $r=\sqrt{\xi_1^2+\xi_2^2}$, we obtain*

$$\partial_T\tilde{\mathcal{P}}=\nabla_\xi\cdot(\nabla_\xi+c(\xi)\xi)\,\tilde{\mathcal{P}} \tag{3.7}$$

with the associated SDE

$$d\xi = -c(\xi)\xi\, dT + \sqrt{2}\, dW_T.$$

For our special case $c(\xi) = 1$, the solution is the Ornstein–Uhlenbeck process. The probability density for this case can be calculated explicitly, as we will do in the next section.

Remark 3.2. *A direct solution of* (3.5) *for $b(r) = r$ can be performed in terms of a series expansion in Laguerre polynomials. For a normalized initial distribution we obtain*

$$\mathcal{P} = re^{-r^2/2} + a_1 e^{-2T} r \left(1 - \frac{r^2}{2}\right) e^{-r^2/2} + \sum_{\nu=2}^{\infty} a_\nu e^{-2\nu T} r e^{-r^2/2} L_\nu \left(\frac{r^2}{2}\right),$$

where the expansion coefficients are determined by the initial distribution

$$a_\nu = \frac{\int_{\mathbb{R}_+ \times [-\pi,\pi]} p_0(r,\beta)\, L_\nu(\frac{r^2}{2})\, dr\, d\beta}{2\pi \int_0^\infty e^{-x} [L_\nu(x)]^2\, dx}.$$

Remark 3.3. *We consider the full Fokker–Planck equation* (3.1)*. Even with a rotationally symmetric initial condition and the rotationally symmetric stationary solution* (3.6)*, terms depending on the angle β appear at intermediate times. This can be seen by computing the next term in the expansion* (3.2)

$$p^{(1)} = p^{(1)}(r, \beta, T) = -\frac{\cos\beta}{\pi} \left[\partial_r \mathcal{P} + \left(b(r) - \frac{1}{r}\right) \mathcal{P}\right],$$

which depends on β even though the initial condition and the stationary solution do not. This could have been already anticipated from the full Fokker–Planck equation, which does not admit time-dependent rotationally symmetric solutions.

4. The case of a moving conveyor belt. In the case of a moving belt, the Fokker–Planck equation (2.2) reads as

$$\partial_t p + ((s + \kappa e_1) \cdot \nabla_\xi)\, p - \partial_\alpha \left[c(\xi)\, (n \cdot \xi)\, p\right] = \frac{1}{2\varepsilon} \partial_\alpha^2 p, \tag{4.1}$$

where $s = (\cos\alpha, \sin\alpha)$ and $n = \partial_\alpha s = (-\sin\alpha, \cos\alpha)$ as well as $\varepsilon = 1/A^2$ are introduced to simplify the notations. The density distribution p satisfies the normalization condition

$$\int_{\mathbb{R}^2 \times [-\pi,\pi]} p(\xi, \alpha, t)\, d\xi\, d\alpha = 1.$$

Additionally we have the initial condition

$$p(\xi, \alpha, 0) = p_0(\xi, \alpha).$$

In the case of strong stochastic influence, i.e., $\varepsilon \ll 1$, we would like to follow the main ideas of the previous case for $\kappa = 0$, i.e., the nonmoving belt. However, the term proportional to κ generates secular terms in the equation for $p^{(1)}$. This indicates that the slow scale needed to get rid of the secular terms should be t. To leading order, the method of multiple scales would then give a hyperbolic reduced equation that does not describe the even slower relaxation towards a stationary solution on the scale $T = \epsilon t$. We need a perturbation method that yields a reduced equation with terms of different order in ϵ: the Chapman–Enskog method. As explained in [4] and [1], the Chapman–Enskog ansatz for the probability density is

$$p(\xi, \alpha, t; \epsilon) = \frac{1}{2\pi}\,\mathcal{P}(\xi, t; \epsilon) + \epsilon\, p^{(1)}(\xi, \alpha; \mathcal{P}) + \epsilon^2 p^{(2)}(\xi, \alpha; \mathcal{P}) + o(\epsilon^2). \tag{4.2}$$

The first term in this equation solves the leading order problem $\partial_\alpha^2 p = 0$. We have anticipated that after a transient in the fast scale $\tau = \epsilon t$, the slowly-varying density $\mathcal{P}$ becomes independent on α, as shown by the method of multiple scales. Of course, this ignores an initial layer that can be inferred from (3.4a): An additional term corresponding to $\sum_{j\in\mathbb{Z}\setminus\{0\}} e^{ij\beta - j^2 t/(2\epsilon)} C_j(r)$ in (3.4a) should be added to (4.2) to account for the effect of initial conditions, so that the probability density becomes

$$\begin{aligned} p(\xi, \alpha, t; \epsilon) = \frac{1}{2\pi}\,\mathcal{P}(\xi, t; \epsilon) + \sum_{j\in\mathbb{Z}\setminus\{0\}} \frac{e^{ij\alpha - j^2 t/(2\epsilon)}}{2\pi} \int_{-\pi}^{\pi} e^{-ija} p_0(\xi, a)\, da \\ + \epsilon\, p^{(1)}(\xi, \alpha; \mathcal{P}) + \epsilon^2 p^{(2)}(\xi, \alpha; \mathcal{P}) + o(\epsilon^2). \end{aligned} \tag{4.3}$$

The higher order terms $p^{(m)}$ depend on time only through their dependence on $\mathcal{P}$. Moreover, up to terms of order ϵ^2, we have

$$\partial_t \mathcal{P} = F^{(0)} + \varepsilon F^{(1)}. \tag{4.4}$$

$F^{(m)}$ are functionals of $\mathcal{P}$ to be determined so that the $p^{(m)}$ are bounded and 2π-periodic in α. Inserting (4.2) and (4.4) into (4.1), we find a hierarchy of problems. To ensure that $\mathcal{P}$ contains all the contributions from the homogeneous equations in the hierarchy, we have to impose the additional constraints

$$\int_{-\pi}^{\pi} p^{(m)}\, d\alpha = 0, \quad m = 1, 2, \ldots. \tag{4.5}$$

The following problem corresponds to the terms of order $\mathcal{O}(\varepsilon)$:

$$-\frac{1}{2}\,\partial_\alpha^2 p^{(1)} = -\,(s\cdot\nabla_\xi)\,\mathcal{P} - \kappa\partial_{\xi_1}\mathcal{P} - \partial_\alpha(c(\xi)\,(s\cdot\xi)\,\mathcal{P}) - F^{(0)},$$

together with (4.5). This problem has a normalized solution which is 2π-periodic in α provided that the average over one period of the right-hand side of the linear equation vanishes. This solvability condition yields $F^{(0)}$:

$$0 = \kappa\partial_{\xi_1}\mathcal{P} + F^{(0)}. \tag{4.6}$$

This condition means that the transport of $\mathcal{P}$ with the belt velocity κ in the ξ_1-direction occurs on the original time scale t. Furthermore, we get

$$p^{(1)} = -2\,[s\cdot(\nabla_\xi + c(\xi)\xi)\,\mathcal{P}],$$

which satisfies (4.5) for $m = 1$. Note that we have not added a term $\mathcal{A}(\xi, t)/(2\pi)$ to the right-hand side of this equation because of the condition (4.5) ensuring that all solutions of the homogeneous equation $\partial_\alpha^2\mathcal{A} = 0$ are included in $\mathcal{P}(\xi, t; \epsilon)$.

To determine the reduced Fokker–Planck equation in analogy to (3.5), we have to consider again the problem provided by terms of order $\mathcal{O}(\varepsilon^2)$

$$-\frac{1}{2}\,\partial_\alpha^2 p^{(2)} = -\left(s\cdot\nabla_\xi\right)p^{(1)} - \kappa\partial_{\xi_1}p^{(1)} - \partial_\alpha\left[c(\xi)\left(n\cdot\xi\right)p^{(1)}\right] - F^{(1)} + 2\left[s\cdot\left(\nabla_\xi + c(\xi)\xi\right)F^{(0)}\right],$$

together with (4.5). The solvability condition that the average of the right-hand side over one period in α should vanish yields $F^{(1)}$:

$$0 = \nabla_\xi\cdot\left(\nabla_\xi + c(\xi)\xi\right)\mathcal{P} - F^{(1)}. \tag{4.7}$$

Inserting the conditions (4.6) and (4.7) into (4.4) yields the reduced equation

$$\partial_t\mathcal{P} = \nabla_\xi\cdot\left(\varepsilon\nabla_\xi + \varepsilon c(\xi)\xi - \kappa e_1\right)\mathcal{P}. \tag{4.8}$$

This is the equation corresponding to (3.7); the difference lies in the transport term $\kappa\partial_{\xi_1}\mathcal{P}$. The stationary solution $\mathcal{P}_s(\xi)$ is characterized by

$$\nabla\cdot\left(\varepsilon\nabla + \varepsilon c(\xi)\xi - \kappa e_1\right)\mathcal{P}_s = 0$$

together with the normalization condition

$$\int_{\mathbb{R}^2}\mathcal{P}_s\,d\xi = 1.$$

The solution of this linear PDE is given by

$$\mathcal{P}_s(\xi) = ke^{-B(\xi)-\kappa\xi_1/\varepsilon}, \tag{4.9}$$

where $\nabla B(\xi) = c(\xi)\xi$ and k is the normalization constant. The associated SDE is

$$d\xi = -\epsilon c(\xi)\xi\,dt + \kappa e_1\,dt + \sqrt{2\epsilon}\,dW_t.$$

Remark 4.1. *In the case of a moving conveyor belt, the stationary distribution (4.9) depends on the noise, as $A = 1/\sqrt{\varepsilon}$. This contrasts with the case of the nonmoving belt, $\kappa = 0$, in which the stationary distribution is the same for deterministic ($A = 0$) or stochastic ($A > 0$) dynamics. Obviously, we obtain a stationary distribution independent of ε in the limit as $\varepsilon \to 0$ only if κ is proportional to $\varepsilon = 1/A^2$. This means we deal with the case of large A and small κ, and the turbulence noise happens to be of order $1/\sqrt{\kappa}$.*

Remark 4.2. *As in the case of the nonmoving belt, we consider the special case $c(\xi) = 1$, i.e., $b(r) = r$. Then, $B(\xi) = \xi_1^2/2 + \xi_2^2/2$ and we obtain the Ornstein–Uhlenbeck type process prescribed by*

$$d\xi = -\epsilon\xi\,dt + \kappa e_1\,dt + \sqrt{2\epsilon}\,dW_t \tag{4.10}$$

or, respectively,

$$\partial_t\mathcal{P} = \nabla\cdot\left(\varepsilon\nabla + \varepsilon\xi - \kappa e_1\right)\mathcal{P}.$$

Its stationary density distribution is Gaussian, centered at $\mu = (\kappa/\varepsilon, 0)$ with variance $\sigma^2 = 1$:

$$\mathcal{P}_s(\xi) = \frac{1}{2\pi}e^{-(\xi_1-\kappa/\varepsilon)^2/2-\xi_2^2/2}. \tag{4.11}$$

To investigate the relaxation to the stationary solution in more detail, we focus on the case $c(\xi) = 1$. To compute the density of the process explicitly, we assume that the initial distribution is a Dirac delta at some point $\mu_0 \in \mathbb{R}^2$. We make the following ansatz for the transient distribution

$$\mathcal{P}(\xi, t) = \frac{f(t)}{2\pi} e^{-(\xi-\mu(t)/\varepsilon)^2/(2\sigma(t))},$$

i.e., a Gaussian with moving center $\mu(t)$, variance $\sigma^2(t)$, and normalization constant $f(t)$. Plugging this ansatz into the reduced Fokker–Planck equation (4.8) and equating for all ξ_1, ξ_2 yields after some calculations

$$\begin{aligned} \frac{d\mu}{dt} &= \varepsilon\left(\kappa e_1 - \mu\right), \\ \frac{d\sigma}{dt} &= 2\varepsilon\left(1-\sigma\right), \\ \frac{df}{dt}\sigma + f\frac{d\sigma}{dt} &= 0. \end{aligned}$$

Together with the initial conditions $\mu(0) = \mu_0$, $\sigma(0) = 0$, and $f(0) = 1$, we obtain $f = 1/\sigma$ and the following motions of the mean and the standard deviation:

$$\begin{aligned} &\mu(t) = \kappa e_1(1 - e^{-\varepsilon t}) + \mu_0 e^{-\varepsilon t}, \\ &\sigma(t) = 1 - e^{-2\varepsilon t}. \end{aligned}$$

Compare this result with the explicit solution formulas for linear stochastic differential equations in [3].

Remark 4.3. *Note that the relaxation to the stationary solution, i.e., $\mu = \kappa e_1$ and $\sigma = 1$, happens on the slow time scale $T = \varepsilon t$. Furthermore, the decay rate for the standard deviation is twice the decay rate of the mean value.*

5. Convergence of the reduced Fokker–Planck equation. In the previous section we have derived the reduced Fokker–Planck equation (4.8)

$$\partial_t \mathcal{P} = \nabla \cdot \left(\varepsilon \nabla \mathcal{P} + \left(\varepsilon c \xi - \kappa e_1\right) \mathcal{P}\right)$$

in the case of dominating stochastic forcing $A^2 = 1/\varepsilon \gg 1$. The "relative velocity" κ of the lay-down process as well as the function $c = c(\xi)$ governing the deterministic fiber bending are still arbitrary. The stationary distribution $\mathcal{P}_s$ of (4.9) is of Gaussian type

$$\mathcal{P}_s(\xi) = k e^{-B(\xi) - \kappa \xi_1/\varepsilon}$$

with $\nabla B(\xi) = c(\xi)\xi$.

The convergence against this stationary solution can be proven by classical arguments; see, e.g., [2] for a recent discussion. Let us introduce the Kullback–Leibler relative entropy

$$S = \int \mathcal{P} \ln \frac{\mathcal{P}}{\mathcal{P}_s}. \tag{5.1}$$

Clearly, $S \geq 0$. The rate of dissipation of the entropy is given by

$$\partial_t S = \int \partial_t \mathcal{P} \ln \frac{\mathcal{P}}{\mathcal{P}_s} = \int \ln \frac{\mathcal{P}}{\mathcal{P}_s}\, \nabla \cdot \left[\varepsilon \nabla \mathcal{P} + (\varepsilon c \xi - \kappa e_1)\mathcal{P}\right]$$

and after integration by parts

$$\partial_t S = -\int \left[\nabla \ln \frac{\mathcal{P}}{\mathcal{P}_s}\right] \cdot \left[\varepsilon \nabla \mathcal{P} + (\varepsilon c \xi - \kappa e_1)\mathcal{P}\right].$$

Using the fact that $\varepsilon \nabla \mathcal{P}_s = -(\varepsilon c\xi - \kappa e_1)\mathcal{P}_s$, we get

$$\partial_t S = -\varepsilon \int \mathcal{P} \left(\nabla \ln \frac{\mathcal{P}}{\mathcal{P}_s}\right)^2 \leq 0.$$

Hence, the entropy is monotonically decaying in time and $S = 0$ if and only if $\mathcal{P} = \mathcal{P}_s$.

Applying the logarithmic Sobolev inequality [6], we obtain

$$\partial_t S \geq -2\varepsilon S \tag{5.2}$$

and hence a decay rate of $e^{-2\varepsilon t}$ for the entropy S. Using the Csiszar–Kullback inequality yields a decay rate of $e^{-\varepsilon t}$ for the $\mathcal{L}_1$-distance of $\mathcal{P}$ and $\mathcal{P}_s$.

6. Approximation quality of the Ornstein–Uhlenbeck process. In this section we investigate the process (2.1) with $c(\xi) = 1$ numerically and compare it with the limiting process for $A \to \infty$, i.e., (4.10):

$$d\xi = -\epsilon \xi \, dt + \kappa e_1 \, dt + \sqrt{2\epsilon}\, dW_t.$$

Its stationary probability density,

$$\mathcal{P}_s(\xi) = \frac{1}{2\pi} e^{-(\xi_1 - \kappa/\varepsilon)^2/2 - \xi_2^2/2},$$

is independent of ε for $\kappa A^2 = k$, $k \in \mathbb{R}$. To test how well $\mathcal{P}_s$ approximates the numerically obtained stationary probability distribution of the process (2.1), we compare both distributions for different values of A. Figure 4 shows the stationary marginal probability distributions for the components ξ_1 and ξ_2 when $k = 0.5$. The distributions are computed from 15000 Monte-Carlo simulations of the ξ-process (2.1); whereas the distribution functions for $A < 1$ are quite different from the marginals of $\mathcal{P}_s$ as they are qualitatively similar for $A = 1$ and show good agreement for $A > 2$. The $\mathcal{L}^\infty$- and $\mathcal{L}^2$-errors are less than 2% for $A > 2$ as illustrated in Figure 5. For $A > 2$ and $N = 15000$ Monte-Carlo simulations, the deviations of the stationary marginal probability distributions from the limiting marginals are within the range of the approximation error, i.e., of order $1/\sqrt{N} \sim 10^{-2}$. Consequently, the limit distribution is a good approximation of the true distributions—already for moderate values of A. However, we should note that the resulting "limit process" of our fiber model for $A \to \infty$, the Ornstein–Uhlenbeck process, is only continuous, not differentiable. Hence, its associated η-process $\eta(t) = \xi(t) - \kappa t e_1$, is not parameterized by arc-length and the lack of differentiability obviously affects the nonextensibility condition. In Figure 6 realizations of the Ornstein–Uhlenbeck (ξ-process of (4.10)) and its associated η-process are depicted and compared to our differentiable fiber process of section 2, assuming an initial value $\xi(0) = (0, 0)$, final time $T = 100$, and parameter values $\kappa = 0.1$, $A = 2.23$. Note that the same amount of fiber mass is laid down.

For the industrial application, it is important to know and control the mass distribution or other distributions of functionals of ξ. These distributions shed light into the structure of the fleece material and therefore may serve to assess its quality.

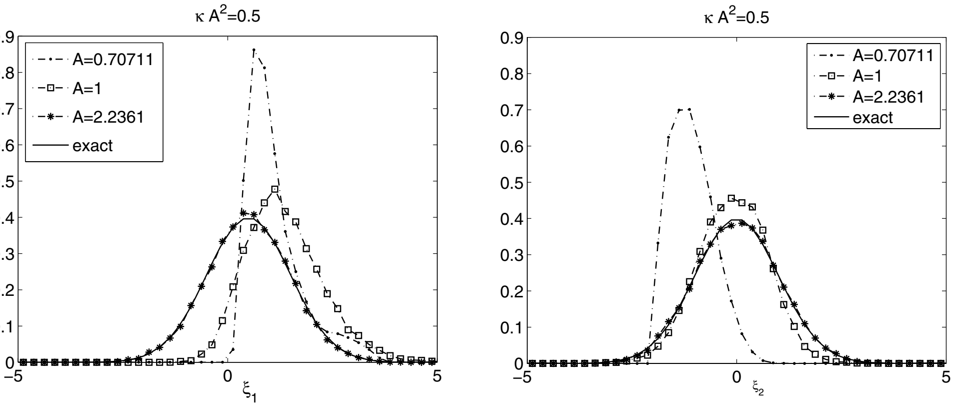


FIG. 4. *Stationary marginal distributions of ξ-components for $c = 1$, $\kappa A^2 = 0.5$, and several values of A.*

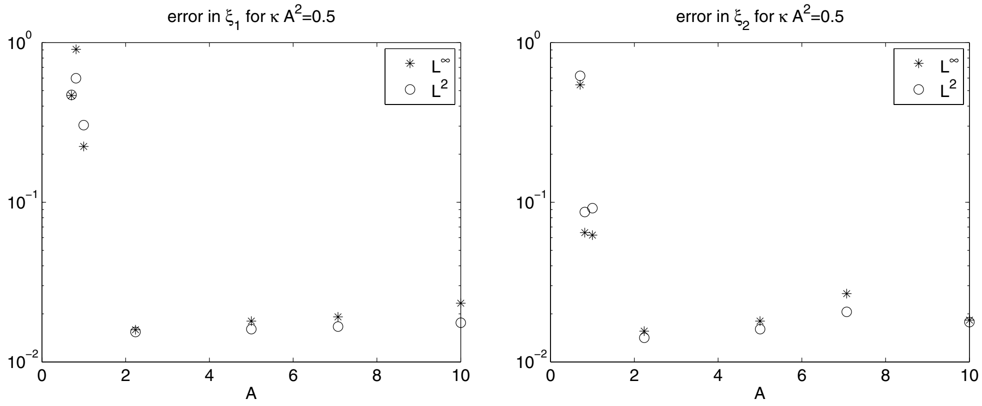


FIG. 5. *$\mathcal{L}^\infty$-error and $\mathcal{L}^2$-error between the stationary marginal distributions and the limiting ($A \to \infty$) stationary marginal distribution for different A.*

The fiber mass that lies in a prescribed spatial domain D can also be interpreted as the time the process stays in that domain. It is described by the distribution of the random variable

$$M = \int_{t_0}^{T} \chi_D(\eta(t))\, dt \tag{6.1}$$

for fixed T, $T > t_0$, with χ_D denoting the characteristic function of D. In the following we compare the distribution of (6.1) for the original fiber process given by (2.1) and the limit process (4.10). We evaluate the distribution of M numerically for the two processes and compare them using Monte-Carlo simulations for fixed $\kappa = 0.1$, $A = 2.23$. Figure 7 shows the probability distribution function (pdf) for the relative time that the respective ξ-processes, (2.1) and (4.10), spend in a square domain D. The square is centered at a point in the set $K = \{(0,0),(0,1),(1,0)\}$, its length may vary in the set $L = \{1, 0.5, 0.25\}$, initially at time $t_0 = 0$, $\xi(0) = (0,0)$, and the final time is $T = 100$. The respective means differ only by 1% which is within the order of the approximation error of the Monte-Carlo simulations. In contrast to this,

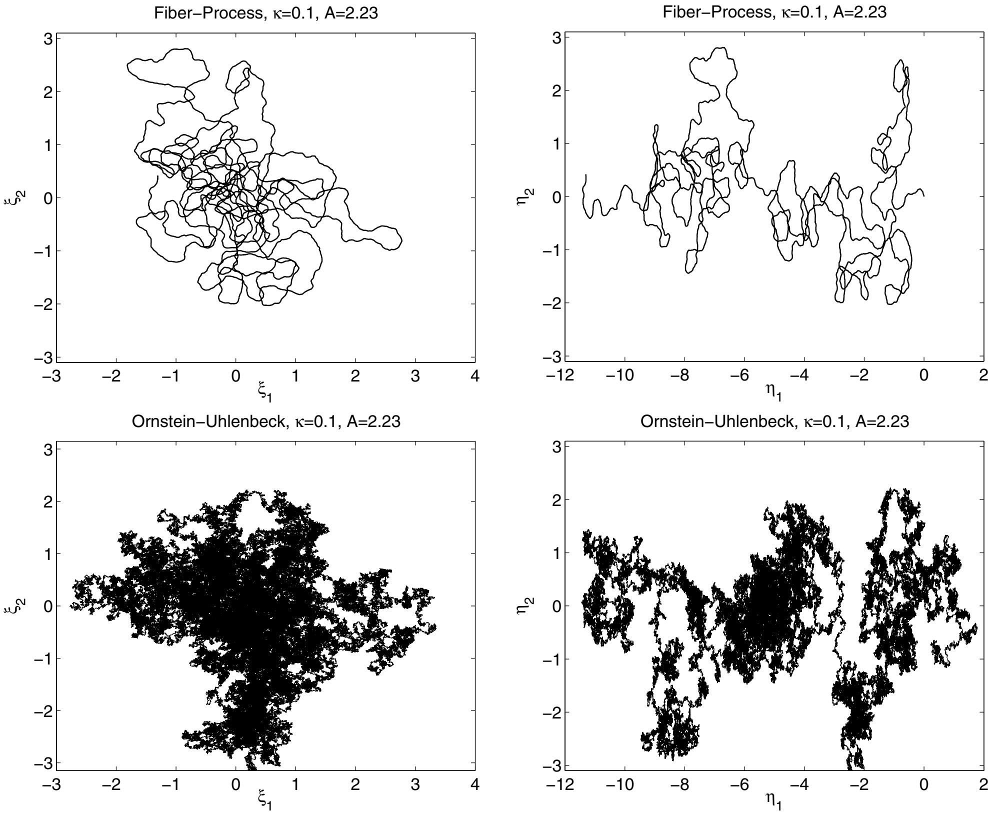


FIG. 6. *Differentiable fiber process (top) versus continuous Ornstein–Uhlenbeck limit process (bottom).*

the relative error of the standard deviations depends on the chosen size of the test domain: the smaller the domain, the higher the error—up to 14% for $L = 0.25$, but only 2% for $L = 1$.

Figure 8 compares the distributions of the mass of a single fiber laid down in a nonwoven web. This means we consider the distribution of (6.1) for the η-processes. We observe the same trend as for the ξ-processes for the relative time spent in a square D: a very good agreement for larger test domains and poor agreement for smaller domains. The symmetry axis for the η-processes is $\eta_2 = 0$. Hence, we consider domains with a certain distance d_{sym} from the center point to the symmetry axis: the larger d_{sym}, the lower the probability that mass lies in D. This tendency is amplified by the size of D: the smaller the test domain, the lower the probability. In contrast to this trend, the probability that mass is accumulated in small domains D is much higher for the Ornstein–Uhlenbeck process than for our fiber process. The reason is that a realization of the continuous Ornstein–Uhlenbeck can move more easily, whereas the differentiable fiber process stays longer in certain regions and therefore other regions are not covered.

Summarizing, the Ornstein–Uhlenbeck limit process approximates our fiber process well—not only as regards the joint probability distribution but also the mass distributions for test domains of size 1 which corresponds to the size of the throwing (lay-down) range of the fiber, but not for smaller domains.

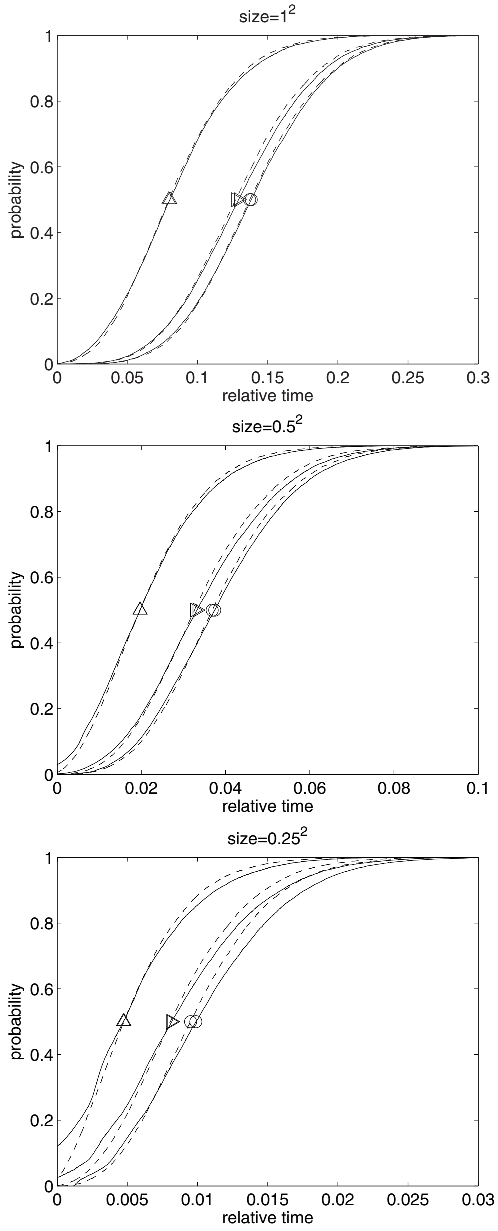


FIG. 7. *Pdf for the relative time that the fiber ξ-process (–) and the Ornstein–Uhlenbeck process (- -) spend in a square of size* $L^2 = \{1^2, 0.5^2, 0.25^2\}$ *(top to bottom), centered at* $K = \{(0,0),(0,1),(1,0)\}$ *(marked by* $\circ$, $\triangleright$, $\triangle$*).*

7. Conclusion. In this work we have presented an extended stochastic model for the fiber lay-down regime in a nonwoven production process that contains a moving conveyor belt. From the associated Fokker–Planck equation and using the method of multiple scales or the Chapman–Enskog technique, we have explicitly determined the limit processes and the stationary and transient joint probability distributions in the hydrodynamic limit, as $A \to \infty$. Quite generally and to leading order of these

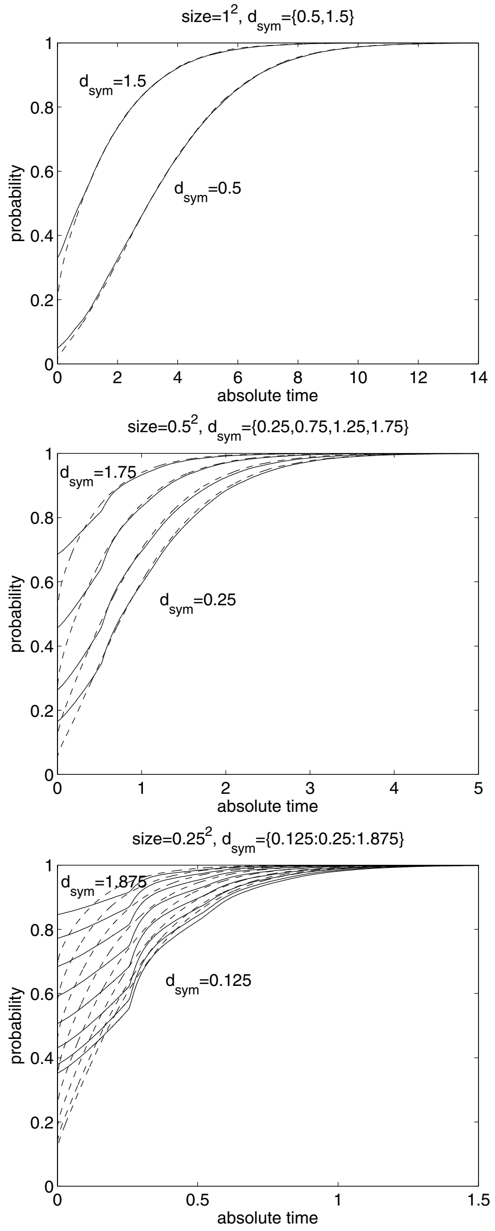


FIG. 8. *Pdf for the mass M of the fiber η-process (–) and the associated Ornstein–Uhlenbeck η-process (- -) laid in a square D, $|D| = L^2 = \{1^2, 0.5^2, 0.25^2\}$ (top to bottom), with different distances d_{sym} to the symmetry axis.*

perturbation methods, we have found that the limiting stationary distribution (as $A \to \infty$) approaches a Gaussian-type function. For the special choice $c = 1$ of the fiber coiling function, the limiting process is an Ornstein–Uhlenbeck process, and the mean of its stationary Gaussian distribution depends on the relation of "relative process velocity" and turbulence noise, κA^2. Already for moderate values of A, i.e., $A > 2$, this limiting distribution turns out to be a very good approximation according to our numerical simulations. Moreover, important distributions of functionals of the process, such as the mass distribution, are well approximated by the Ornstein–Uhlenbeck process for test squares D of the size of the typical throwing (lay-down) range of the fibers.

For the control and optimization of the production and quality of nonwoven materials, the parameters characterizing our model, c, A, κ, and samples sizes D, should be identified from FIDYST-simulations of the complete physical production process as well as from experimental data. If the ranges of these parameters are such that the limiting process studied in this work describes well the physical production, the fiber mass distribution in a fleece material could be determined from the superposition of many Ornstein–Uhlenbeck η-processes.

Acknowledgment. We thank Vincenzo Capasso for having brought us together to collaborate on this topic.

REFERENCES

[1] J. A. Acebrón, L. L. Bonilla, C. J. Pérez-Vicente, F. Ritort, and R. Spigler, *The Kuramoto model: A simple paradigm for synchronization phenomena*, Rev. Modern Phys., 77 (2005), pp. 137–185.

[2] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations*, Comm. Partial Differential Equations, 26 (2001), pp. 43–100.

[3] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Krieger, Malabar, FL, 1974.

[4] L. L. Bonilla, *Chapman-Enskog method and synchronization of globally coupled oscillators*, Phys. Rev. E (3), 62 (2000), pp. 4862–4868.

[5] T. Götz, A. Klar, N. Marheineke, and R. Wegener, *A stochastic model for the fiber lay-down process in the nonwoven production*, SIAM J. Appl. Math., 67 (2007), pp. 1704–1717.

[6] L. Gross, *Logarithmic Sobolev inequalities*, Amer. J. Math., 97 (1975), pp. 1061–1083.

[7] M. Grothaus and A. Klar, *Ergodicity and Rate of Convergence for a Non-Sectorial Fiber Lay-Down Process*, preprint, 2007.

[8] J. W. S. Hearle, M. A. I. Sultan, and S. Govender, *The form taken by threads laid on a moving belt, parts* i–iii, J. Textile Inst., 67 (1976), pp. 373–386.

[9] D. Hietel and N. Marheineke, *Mathematical modeling and numerical simulation of fiber dynamics*, in Proc. Appl. Math. Mech., Vol. 5, Wiley, 2005, pp. 667–670.

[10] L. Mahadevan and J. B. Keller, *Coiling of flexible ropes*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 1679–1694.

[11] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: General modeling framework*, SIAM J. Appl. Math., 66 (2006), pp. 1703–1726.

[12] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: Specific Taylor drag*, SIAM J. Appl. Math., 68 (2007), pp. 1–23.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 666–679

INVERSE BOUNDS OF TWO-COMPONENT COMPOSITES*

CHRISTIAN ENGSTRÖM†

Abstract. A method is presented for estimating microstructural parameters from permittivity measurements of two-component composites. This structural information is described by a particular positive measure in the Stieltjes integral representation of the effective permittivity. The dependence on the geometrical structure can be reduced to the problem of calculating the moments of the measure. We present a method that uses measurement data at a set of distinct frequencies or temperatures to calculate bounds on several moments. These inverse bounds are improved when the volume fraction is known or the material is isotropic. Composites with known geometrical structure illustrate the method.

Key words. effective permittivity, inverse bounds, Stieltjes series, Padé approximants

AMS subject classifications. 78A48, 41A20, 41A21, 30E05, 30E10

DOI. 10.1137/070683039

1. Introduction.

The bulk permittivity of composites is determined by the bulk properties of the components and the geometrical structure of the composite. In Bergman's representation of the effective permittivity the dependence on the geometry in a two-component composite is described by a particular positive measure [2, 15]. Various inverse algorithms for recovering this measure from experimental data have been developed [8, 9, 6, 25]. When the measure is recovered the volume fraction and the anisotropy of the material are given by the first two moments of the measure.

Based on Bergman's [2] representation, Milton [21] developed a general method for obtaining a hierarchy of bounds on the effective permittivity, given the first n moments. One disadvantage of the inverse algorithms above is that we lose the concept of bounds. The numerical methods cannot recover the measure when we have few or inaccurate measurements. Using the numerical approximations of the measure can then result in inaccurate values on the moments and, as a consequence, invalid bounds on effective properties.

Instead of seeking the measure, the measured bulk properties can be used directly to estimate the structural parameters. In other words, limits on the moments of the measure are derived directly. McPhedran, McKenzie, and Milton [18] and McPhedran and Milton [19] developed methods to estimate the volume fraction (the lowest-order moment) from measurement data. Cherkaeva and Golden [7] derived, in the case of measurements of lossy materials, explicit formulas for bounds on the volume fraction.

In two previous papers we proposed general methods for bounding the moments of the measure from measurements of real-valued [10] and complex-valued [11] bulk properties. Based on these inverse bounds, we develop a new algorithm that gives significantly tighter bounds on the lowest n moments.

Cross-property bounds incorporate information from measurements of one parameter to bound a related parameter. A general method for deriving these bounds is to apply Prager's method [23] on Bergman's and Milton's bounds [22]. Cross-property

*Received by the editors February 19, 2007; accepted for publication (in revised form) September 4, 2007; published electronically December 13, 2007. The author acknowledges the support of the German Research Foundation (RTG 1294).

http://www.siam.org/journals/siap/68-3/68303.html

†Department of Mathematics, Karlsruhe University, Karlsruhe 76131, Germany (christian.engstroem@math.uni-karlsruhe.de).

bounds can also be expressed in terms of bounds on the moments [10]. Thus, the inverse algorithm in this paper can be used to bound several other physical phenomena, such as electrical and thermal conductivity, magnetism, diffusion, and flow in porous media.

Before proceeding to this problem, we give a description of the Bergman–Milton theory [2, 4, 5, 20, 21, 22] and how the hierarchy of bounds can be used to derive a hierarchy of inverse bounds on the moments. The presented algorithm is illustrated by composites with known structure.

2. Representation of the effective permittivity. Consider a two-component material modeled by the permittivity $\boldsymbol{\epsilon}$, where the components are homogeneous and isotropic with permittivity ϵ_1 and ϵ_2. Assume that the electric field $\boldsymbol{E}$ and the electric flux density $\boldsymbol{D}$ satisfy the linear constitutive relation $\boldsymbol{D}(\boldsymbol{x}) = \boldsymbol{\epsilon}(\boldsymbol{x})\boldsymbol{E}(\boldsymbol{x})$. In many instances, the characteristic length of inhomogeneities in the d-dimensional composite material is small compared with the wavelength but much larger than the atomic scale. In this case an effective (bulk) permittivity $\boldsymbol{\epsilon}_{\mathrm{e}}$ is defined via

$$\langle \boldsymbol{D} \rangle = \langle \boldsymbol{\epsilon} \boldsymbol{E} \rangle = \boldsymbol{\epsilon}_{\mathrm{e}} \langle \boldsymbol{E} \rangle, \tag{2.1}$$

which relates the average, $\langle \cdot \rangle$, of the electric flux density $\langle \boldsymbol{D} \rangle$ to the average of the electric field $\langle \boldsymbol{E} \rangle$. The notation $\langle \cdot \rangle$ means spatial average over all of $\mathbb{R}^d$ or ensemble average. The averaged fields have no oscillations on the length scale of the microstructure, since they are smoothed out, but they retain slow macroscopic variations. A mathematical justification of the homogenization rule (2.1) can, for example, be found in [16, p. 15]. Let ϵ_{e} be one of the eigenvalues in the matrix $\boldsymbol{\epsilon}_{\mathrm{e}}$. In a two-component mixture the effective permittivity ϵ_{e} has the Stieltjes integral representation [2, 15]

$$\epsilon_{\mathrm{e}}(\epsilon_1, \epsilon_2) = \epsilon_2 - \epsilon_2 G(s), \tag{2.2}$$

where

$$G(s) = \int_0^1 \frac{\mathrm{d}m(y)}{s - y}, \qquad s = \frac{\epsilon_2}{\epsilon_2 - \epsilon_1} \notin [0, 1]. \tag{2.3}$$

The positive (Borel) measure m is a purely geometric quantity, which depends on the structure but not on the values of the components. If the geometrical structure is identical, the single integral (2.3) gives the effective permittivity, independent of the values of the components.

Let $s = -1/z$ in the representation (2.3). The integral representation of G is then transformed to

$$\hat{G}(z) = -\frac{1}{z} G\left(-\frac{1}{z}\right) = \int_0^1 \frac{\mathrm{d}m(y)}{1 + zy}, \tag{2.4}$$

which is the standard form of a Stieltjes integral representation [1, p. 229].

3. Real-valued bounds on the permittivity. Partial information concerning the microstructure, such as the volume fraction, can be used to derive exact bounds on the effective permittivity [2, 22]. We use the Stieltjes series expansion

$$\epsilon_{\mathrm{e}} = \epsilon_2 + \epsilon_2 z \hat{G}(z) = \epsilon_2 F(z), \qquad F(z) = \sum_{n=0}^{\infty} c_n z^n, \tag{3.1}$$

where $z = -1/s = (\epsilon_1 - \epsilon_2)/\epsilon_2$ is the contrast and the coefficients c_n are given by the moments of the measure

$$c_{n+1} = (-1)^n \int_0^1 y^n \, \mathrm{d}m(y). \tag{3.2}$$

The constants c_n depend on the microstructure but not on the values of the two components. If the structure is the same, the single series (3.1) gives the effective permittivity, independent of the value of the components. The zero-order moment c_1 is the volume fraction of the component ϵ_1, and c_2 depends on the anisotropy in the material. In the case of a d-dimensional statistically isotropic composite, the second moment is $-c_1(1-c_1)/d$ (see [2]).

Assume real-valued materials with $\epsilon_2 > \epsilon_1$. When $-1 < z < 0$, the denominator $1 + zy$ is bounded between zero and one. We find that the integral (2.4) can be estimated by

$$\int_0^1 \frac{\mathrm{d}m(y)}{1+zy} \geq \int_0^1 \mathrm{d}m(y) = c_1. \tag{3.3}$$

This implies that the following inequality is satisfied:

$$\frac{\epsilon_\mathrm{e}}{\epsilon_2} = 1 + z\tilde{G}(z) \leq 1 + c_1 z. \tag{3.4}$$

That is, the effective permittivity is bounded from above by the arithmetic mean $c_1\epsilon_1 + (1-c_1)\epsilon_2$. This estimate is satisfied as an equality if the measure m is a point mass concentrated at $y = 0$.

To derive lower bounds on the effective permittivity we use a representation of the inverse of the effective permittivity. Define the auxiliary function $z\tilde{H}(z)$ by

$$(1 + z\tilde{H}(z))(1 + z\hat{G}(z)) = 1 + z, \tag{3.5}$$

which is equivalent to

$$1 + z\tilde{H}(z) = \frac{\epsilon_1}{\epsilon_\mathrm{e}}. \tag{3.6}$$

The new function $z\tilde{H}(z)$ has the same analytic properties as the original function $z\hat{G}(z)$ [1, 5]. That is, the scaled inverse permittivity has the representation

$$\left(\frac{\epsilon_\mathrm{e}}{\epsilon_1}\right)^{-1} = 1 + z\tilde{H}(z) = \tilde{F}(z), \quad \tilde{F}(z) = \sum_{n=0}^{\infty} \tilde{c}_n z^n, \tag{3.7}$$

where $\tilde{H}$ is the Stieltjes integral

$$\tilde{H}(z) = \int_0^1 \frac{\mathrm{d}\tilde{m}(y)}{1+zy} \tag{3.8}$$

and the coefficients $\tilde{c}_n$ are given by the moments of the measure

$$\tilde{c}_{n+1} = (-1)^n \int_0^1 y^n \, \mathrm{d}\tilde{m}(y). \tag{3.9}$$

Expanding the product of the series for $1 + z\hat{G}(z)$ and for $1 + z\tilde{H}(z)$ in (3.5) and collecting terms with the same powers of z shows that the coefficients c_n and $\tilde{c}_n$ in the two series are related according to

$$\tilde{c}_0 = 1, \quad \tilde{c}_1 = 1 - c_1, \quad \tilde{c}_n = -\sum_{k=0}^{n-1} \tilde{c}_k c_{n-k}. \tag{3.10}$$

The coefficient c_1 is the volume fraction of component one and $\tilde{c}_1$ is the volume fraction of component two. For all $\epsilon_2 > \epsilon_1$, the integral (3.8) can be estimated by

$$\int_0^1 \frac{\mathrm{d}\tilde{m}(y)}{1 + zy} \geq \int_0^1 \mathrm{d}\tilde{m}(y) = \tilde{c}_1. \tag{3.11}$$

This implies that the effective permittivity is bounded from below by the harmonic mean,

$$\epsilon_\mathrm{e} \geq \frac{\epsilon}{1 + \tilde{c}_1 z} = \frac{\epsilon_1}{1 + \tilde{c}_1 z} = \left(\frac{c_1}{\epsilon_1} + \frac{\tilde{c}_1}{\epsilon_2} \right)^{-1}. \tag{3.12}$$

The estimate is satisfied as an equality if the measure $\tilde{m}$ is a point mass concentrated at $y = 0$.

The convexity of the function $(1 + zy)^{-1}$ can be used to derive the Hashin–Shtrikman bounds, which are finer estimations of the integrals [16, p. 219]. Finer estimations become more and more tricky, and a systematic method for obtaining bounds is preferable.

A general method for obtaining a hierarchy of bounds using the analytic properties of the effective permittivity was developed by Bergman [3, 4] and Milton [20, 21]. Alternatively, known lower and upper bounds of the Stieltjes functions in the form of continued fractions or Padé approximants can be used. These approximation methods lead to identical bounds [13, 1].

The $\epsilon_{p,q}$ Padé approximant to ϵ_e is defined by the equation

$$\epsilon^{\mathrm{eff}}(z)Q(z) - P(z) = \mathcal{O}(z^{p+q+1}), \tag{3.13}$$

where P and Q are polynomials of degree at most p and q, respectively [1]. This equation gives us an approximation of the effective permittivity by the rational function

$$\epsilon_{p,q} = \frac{P(z)}{Q(z)} = \frac{a_0 + \cdots + a_p z^p}{1 + b_1 z + \cdots + b_q z^q}. \tag{3.14}$$

When $\epsilon_2 > \epsilon_1$ and $N \geq 1$, the N-point upper bounds ϵ_N^U are obtained by forming the approximations

$$\epsilon_{2M+1}^\mathrm{U} = \epsilon_2 \epsilon_{M+1,M}(F), \quad \epsilon_{2M}^\mathrm{U} = \epsilon_2 \epsilon_{M,M}(F), \tag{3.15}$$

of the Stieltjes series (3.1). For example, the arithmetic mean (3.4) is obtained from the $\epsilon_{1,0}$ Padé approximant of (3.1),

$$\epsilon_1^\mathrm{U} = (\epsilon_2 + c_1 \epsilon_2 z) = (c_1 \epsilon_1 + \tilde{c}_2 \epsilon_2). \tag{3.16}$$

Lower bounds on ϵ_e are given from Padé approximations of the series (3.7). The N-point lower bounds ϵ_N^L, when $\epsilon_2 > \epsilon_1$ and $N \geq 1$, are obtained from

$$\epsilon_{2M+1}^\mathrm{L} = \epsilon_1 [\epsilon_{M+1,M}(\tilde{F})]^{-1}, \qquad \epsilon_{2M}^\mathrm{L} = \epsilon_1 [\epsilon_{M,M}(\tilde{F})]^{-1}. \tag{3.17}$$

For example, the harmonic mean (3.12) corresponds to the $\epsilon_{1,0}$ Padé approximant of the expansion (3.7)

$$\epsilon_1^{\mathrm{L}} = \frac{\epsilon_1}{1+\tilde{c}_1 z} = \left(\frac{c_1}{\epsilon_1} + \frac{\tilde{c}_1}{\epsilon_2}\right)^{-1}. \tag{3.18}$$

The $\epsilon_{1,1}$ Padé approximant of the expansion (3.7) gives the lower bound

$$\epsilon_2^{\mathrm{L}} = \epsilon_1[\tilde{c}_1 - \tilde{c}_2 z][\tilde{c}_1 - \tilde{c}_2 z + \tilde{c}_1^2 z]^{-1}, \tag{3.19}$$

where $\tilde{c}_2 = -c_2 - c_1\tilde{c}_1$. In the isotropic case $c_2 = -c_1\tilde{c}_1/d$, the bound (3.19) is equivalent to the lower Hashin–Shtrikman bound. For this reason, we call c_2 the anisotropy parameter.

3.1. Complex-valued bounds on the permittivity. Bergman [3, 4] and Milton [20, 21] extended the real-valued bounds above to the complex case. We write these bounds in terms of bounds on the moments c_n. The minimum $c_n^{\min}$ and the maximum $c_n^{\max}$ of c_n are functions of the lower-order parameters $c_1, c_2, \ldots, c_{n-1}$. The extreme values can be determined by varying the c_n parameter in the n-point bounds and using that the n-point bounds are forbidden to violate the $(n-1)$-point bounds [10]. The volume fraction is of course bounded between zero and one. The bounds ϵ_2^{L} and ϵ_1^{U} are equal when $c_2 = 0$, and ϵ_2^{L} and ϵ_1^{L} are equal when $c_2 = -c_1\tilde{c}_1$. This implies the inequality

$$-c_1\tilde{c}_1 \le c_2 \le 0. \tag{3.20}$$

In the same way, we get the inequality

$$\frac{c_2^2}{c_1} \le c_3 \le -c_2\left(1 + \frac{c_2}{\tilde{c}_1}\right). \tag{3.21}$$

In the general case, when the values of the components are complex, the real segment $l = \{c_n;\, c_n^{\min} \le c_n \le c_n^{\max}\}$ is for fixed values on $c_1, c_2, \ldots, c_{n-1}$ mapped by $\epsilon_n^{\mathrm{L}}(c_n)$ and $\epsilon_n^{\mathrm{U}}(c_n)$ on a circle or a line segment. The bounds on the moments c_n give a parameterization of the lens-shaped boundary. For example, we get complex-valued bounds from the lens-shaped region bounded by

$$\epsilon_2^{\mathrm{L}}(\tilde{c}_2; \epsilon_1, \epsilon_2, \tilde{c}_1), \quad \epsilon_2^{\mathrm{U}}(c_2; \epsilon_1, \epsilon_2, c_1) \tag{3.22}$$

with the structural parameter c_2 varying between the endpoints in (3.20), and $\tilde{c}_2$ according to (3.10) related to c_2 by $\tilde{c}_2 = -c_2 - c_1\tilde{c}_1$.

Alternatively, we can describe the bounds $\epsilon_n^{\mathrm{L}}(c_n)$ and $\epsilon_n^{\mathrm{U}}(c_n)$ in terms of the points through which the circles pass [3, 20, 21].

4. Inverse bounds from permittivity measurements. In some cases, the volume fraction c_1 is known and the d-dimensional material is usually assumed to be isotropic $c_2 = -c_1(1-c_1)/d$. Higher-order moments depend on the detailed geometrical structure and are in most cases unknown. This gives us at most two coefficients in the series expansion (3.1), but in many cases even the volume fraction is uncertain.

The moments c_n can be expressed in terms of integrals over correlation functions [24, p. 520]. The calculations of higher-order correlation functions are in general very demanding [24]. The complex bounds on the effective permittivity in section

3.1 were parameterized in terms of bounds on the moments (3.20)–(3.22). We use inverse bounds from permittivity measurements to improve these geometry-independent bounds on the moments c_n that characterize the structure. This method avoids the cumbersome calculations of the correlation functions.

Bounds on the lowest-order moment c_1 (the volume fraction) have been derived using various methods [18, 19, 7, 10, 11]. In [10, 11] the author proposed general methods for deriving bounds on the higher-order moment c_n.

The inverse algorithm presented in section 4.1 applies to both the real-valued bounds in [10] and to the complex-valued bounds in [11]. In the examples in sections 5 and 6 we use complex-valued permittivities and bound the volume fraction c_1, the anisotropy parameter c_2, and c_3.

The $\epsilon_{1,1}$ Padé approximant to the series (3.1) gives an upper bound ϵ_2^{U} that in the isotropic case corresponds to the upper Hashin–Shtrikman bound [22, p. 574]. When $\epsilon_{\mathrm{e}} \notin \{\epsilon_1, \epsilon_2\}$, the upper bound ϵ_2^{U} can be inverted giving a bound on c_1 [11, 7]. Explicitly, the volume fraction is in the complex case bounded from below by [11, 7]

$$c_1^{\mathrm{L}} = \Im(z)\frac{(\Im(\epsilon^{\mathrm{eff}}) - \Im(\epsilon_2))^2 + (\Re(\epsilon^{\mathrm{eff}}) - \Re(\epsilon_2))^2}{|z|^2(\Im(\epsilon^{\mathrm{eff}})\Re(\epsilon_2) - \Re(\epsilon^{\mathrm{eff}})\Im(\epsilon_2))}. \tag{4.1}$$

In the same way, the generalization of the lower Hashin–Shtrikman bound ϵ_2^{L} [22, p. 574] can be inverted. Explicitly, the volume fraction is bounded from above by [11, 7]

$$c_1^{\mathrm{U}} = 1 - \Im(z)\frac{(\Im(\epsilon^{\mathrm{eff}}) - \Im(\epsilon_1))^2 + (\Re(\epsilon^{\mathrm{eff}}) - \Re(\epsilon_1))^2}{|z|^2(\Re(\epsilon^{\mathrm{eff}})\Im(\epsilon_1) - \Im(\epsilon^{\mathrm{eff}})\Re(\epsilon_1))}. \tag{4.2}$$

If the volume fraction c_1 is known, we derive in the same way bounds on the anisotropy parameter c_2. The $\epsilon_{2,1}$ Padé approximant to the series (3.1) gives an upper bound ϵ_3^{U} that in the isotropic case corresponds to the upper Beran bound [22, p. 574]. When $\epsilon_{\mathrm{e}} \notin \{\epsilon_1^{\mathrm{L}}, \epsilon_1^{\mathrm{U}}\}$, the upper bound ϵ_3^{U} can be inverted, giving a bound on c_2, and when $\epsilon_{\mathrm{e}} \notin \{\epsilon_2^{\mathrm{L}}, \epsilon_2^{\mathrm{U}}\}$, the $\epsilon_{2,2}$ Padé approximant can be inverted, giving a bound on c_3. Explicit bounds on the second moment c_2 can be found in [12].

A disadvantage with the inverse bounds above, and the corresponding bounds in the real case, is that the parameters $c_1, c_2, \ldots, c_{n-1}$ need to be known to bound c_n. Below we develop an algorithm that bounds $c_1, c_2, \ldots, c_n$ without any knowledge of the geometrical structure. Moreover, the bounds on the lower-order parameters are significantly tighter than in [10, 11].

The bounds on the moments depend on the geometrical structure. In order to study the dependence on the measure, we consider real-valued materials with $\epsilon_2 \geq \epsilon_1$. From the relation $\epsilon_1 \leq \epsilon_{\mathrm{e}} \leq \epsilon_2$ we have

$$0 \leq G(s) \leq \frac{1}{s} \leq 1. \tag{4.3}$$

Let $s = 1 + \delta$, $\delta > 0$. From the inequality above we have

$$1 \geq G(1+\delta) = \int_0^1 \frac{1}{1-y+\delta}\,\mathrm{d}m(y) \geq \frac{m(\{1\})}{\delta}, \tag{4.4}$$

which implies that the measure m of the set $\{1\}$ is zero, since $\delta > 0$ is arbitrary. The moments of the measure

$$c_{n+1} = (-1)^n \int_0^1 y^n\,\mathrm{d}m(y) \tag{4.5}$$

then vanish in the limit $n \to \infty$, and the absolute value of the moments $c_1, c_2, \dots$ form a nonincreasing sequence $|c_1| \geq |c_2| \dots$. The convergence rate of the moments c_n to zero depend strongly on the support of the measure. If m has no support close to $y = 1$, the convergence is exponential. The bounds c_1^{L} and c_1^{U} come close together when the volume fraction c_1 is low or high, and the bounds on c_2 come close together for anisotropic materials. The most challenging structures are therefore when the moments are the arithmetic mean of their maximum and minimum values, for example, when the first two moments are $c_1 = 1/2$ and $c_2 = -c_1\tilde{c}_1/2$, which in two dimensions corresponds to an isotropic material.

4.1. The inverse algorithm. Assume that we are given M sets of measurement data $\{\epsilon_1, \epsilon_1, \epsilon_{\mathrm{e}}\}$ and calculate the lower bound on the volume fraction c_1^{L} and the upper bound on the volume fraction c_1^{U} for all the sets. The tightest bounds on the volume fraction c_1 that can be obtained directly from the inverse bounds (4.1) and (4.2) are

$$(c_1^{\mathrm{L}})_1 = \max c_1^{\mathrm{L}}, \quad (c_1^{\mathrm{U}})_1 = \min c_1^{\mathrm{U}}, \tag{4.6}$$

where the maximum and minimum are taken over all data sets.

In the second step, fix $c_1 \in [(c_1^{\mathrm{L}})_1, (c_1^{\mathrm{U}})_1]$ and calculate the inverse bounds c_2^{L} and c_2^{U} for all M sets of measurement data. It is required that the anisotropy parameter c_2 for a fixed value on the volume fraction c_1 satisfy

$$\max c_2^{\mathrm{L}} \leq \min c_2^{\mathrm{U}}, \tag{4.7}$$

where the maximum and minimum are taken over all M data sets. This condition gives improved restrictions on the possible volume fraction c_1 and bounds on the anisotropy parameter c_2. The attainable values of (c_1, c_2) consist of a bounded region in the (c_1, c_2)-plane, denoted by Ω_2. We define

$$(c_1^{\mathrm{L}})_2 = \min\{c_1; c_1 \in \Omega_2\}, \tag{4.8}$$

$$(c_1^{\mathrm{U}})_2 = \max\{c_1; c_1 \in \Omega_2\} \tag{4.9}$$

and

$$(c_2^{\mathrm{L}})_2 = \min\{c_2; c_2 \in \Omega_2\}, \tag{4.10}$$

$$(c_2^{\mathrm{U}})_2 = \max\{c_2; c_2 \in \Omega_2\}. \tag{4.11}$$

The c_1-independent bounds $(c_2^{\mathrm{L}})_2$ and $(c_2^{\mathrm{U}})_2$ in Figure 4.1 do not take into account that the bounds on c_2 depend on c_1, but they can be used to simplify the algorithm.

In a third step, fix $(c_1, c_2) \in \Omega_2$ and calculate the inverse bounds c_3^{L} and c_3^{U} for all M sets of measurement data. As above, we require that the parameter c_3 for a fixed value on c_1 and c_2 satisfy

$$\max c_3^{\mathrm{L}} \leq \min c_3^{\mathrm{U}}, \tag{4.12}$$

where the maximum and minimum are taken over all M data sets. This requirement gives further restrictions on the structural parameters c_1 and c_2. Moreover, we get restrictions on the possible values on the c_3. The attainable values of (c_1, c_2, c_3) consist of a bounded region in the (c_1, c_2, c_3)-space, denoted by Ω_3.

The same procedure can be used to an arbitrary order. In general, $c_1, \dots, c_{n-1}$-independent bounds are obtained from

$$(c_n^{\mathrm{L}})_m = \min\{c_n; c_n \in \Omega_m \subset \mathbb{R}^m\}, \tag{4.13}$$

$$(c_n^{\mathrm{U}})_m = \max\{c_n; c_n \in \Omega_m \subset \mathbb{R}^m\}, \tag{4.14}$$

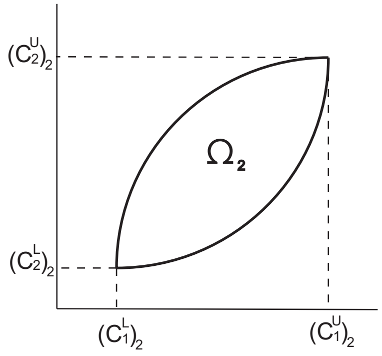


FIG. 4.1. *Region Ω_2 represents the attainable values of (c_1, c_2) after two steps in the algorithm.*

where m is the number of steps in the algorithm and $n = 1, 2, \ldots, m$. The tighter bounds Ω_m consist of a bounded region in $\mathbb{R}^m$ that for $m > 2$ is difficult to illustrate. Instead we use the bounds (4.13) and (4.14) in the presentation of bounds on higher-order structural parameters.

For the complex case, inverse bounds on higher-order moments c_n were derived in [11]. One disadvantage of the previous method [11] is that more steps in the algorithm do not improve the bounds on the lower-order moments. For example, the tightest possible bounds on the volume fraction with the previous method are equivalent to step one in the new method presented here. Every further step in the new algorithm improves the bounds on the volume fraction. In practice, the accuracy in the measurements limits the tightness of the bounds on the structural parameters c_n.

In many cases, partial information of the geometrical structure is available; for example, in the random case the composite is usually assumed to be isotropic, $c_2 = -c_1(1 - c_1)/d$. This knowledge can be used in the algorithm to derive tighter bounds on the volume fraction c_1 and on the higher-order parameters. The bounds on the structural parameters that were derived above can in principle be calculated analytically; however, the complexity in the formulas makes this difficult after a few steps of the algorithm. Below we present numerical calculations of the bounds when three steps of the algorithm are used.

5. Examples. The inverse algorithm in section 4.1 is illustrated by composites with known geometry. That is, we calculate bounds on the moments c_n and compare with the exact values. We give three examples of the method when no structural information is supposed to be known but information from measurements of the effective permittivity is available. We show that the size of the attainable values of (c_1, c_2), denoted by Ω_2, decreases for each step and present values on the bounds $(c_n^{\mathrm{L}})_m$ and $(c_n^{\mathrm{U}})_m$ when $n = 1, 2, 3$ and $m = 1, 2, 3$.

Three sets of measurements are used in all calculations in this section, but the number of sets is arbitrary. In all cases, we assume that component one is a frequency-independent material $\epsilon_1(\omega) = 3$ in the chosen range of frequencies. The second component is dispersive and is at the frequencies ω_0, ω_1, and ω_2 measured to be

$$\epsilon_2(\omega_0) = 4.1 + 4.5\mathrm{i}, \quad \epsilon_2(\omega_1) = 4.6 + 0.06\mathrm{i}, \quad \epsilon_2(\omega_2) = 3.7 + 0.04\mathrm{i}. \tag{5.1}$$

The values on the components were previously used in [11]. We use the same values in this paper to simplify the comparison between the methods. The algorithm presented

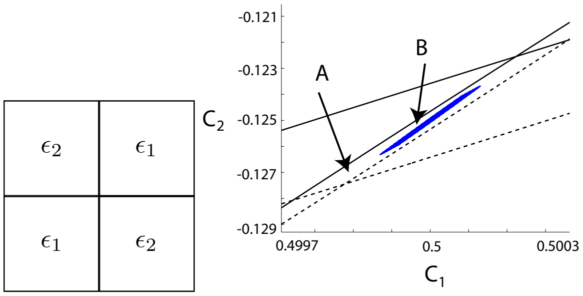


FIG. 5.1. *Left: The checkerboard structure, a two-dimensional and periodic geometry. Right: Bounds on the structural parameters c_1 and c_2 in the checkerboard case. The exact values on the parameters are $c_1 = 0.5$ and $c_2 = -0.125$. Region A, which is bounded by the two solid lines and the two dotted lines, represents the attainable values of (c_1, c_2) after two steps in the algorithm. Region B represents the possible values of (c_1, c_2) after three steps in the algorithm.*

TABLE 5.1

The table shows the bounds (4.13) and (4.14) using the checkerboard structure, $\epsilon_e = \sqrt{\epsilon_1\epsilon_2}$. The bounds are on the parameters c_1, c_2, and c_3 when one, two, and three steps in the algorithm are used.

c_1^L	c_1^U	c_2^L	c_2^U	c_3^L	c_3^U
0.4986	0.5014	-	-	-	-
0.4996	0.5004	-0.1281	-0.1219	-	-
0.4999	0.5001	-0.1263	-0.1237	0.0595	0.0655

in this paper gives very tight bounds from the measurement data (5.1). In section 6, we use measurements from a gold-magnesium oxide nanocomposite, which is a more challenging example.

5.1. The checkerboard. The two-dimensional checkerboard structure $\epsilon_e = \sqrt{\epsilon_1\epsilon_2}$ corresponds exactly to Bruggeman's formula at the percolation threshold $c_1 = 0.5$ [24]. From the Stieltjes inversion formula [1, 22] it follows that $\epsilon_e = \sqrt{\epsilon_1\epsilon_2}$ is obtained from the measure $\mathrm{d}m_C(y) = \frac{1}{\pi}\sqrt{(1-y)/y}\,\mathrm{d}y$. The moments (3.2) of m_C are

$$c_{n+1} = \begin{pmatrix} 1/2 \\ n+1 \end{pmatrix}, \tag{5.2}$$

where the first three moments of the measure are $c_1 = 0.5$, $c_2 = -0.125$, and $c_3 = 0.0625$. In this case, the moments converge very slowly to zero. The effective permittivities of the three sets (5.1) are in this case

$$\epsilon_e(\omega_0) = 3.91 + 1.727\mathrm{i}, \quad \epsilon_e(\omega_1) = 3.72 + 0.024\mathrm{i}, \quad \epsilon_e(\omega_2) = 3.33 + 0.018\mathrm{i}. \tag{5.3}$$

Figure 5.1 shows the possible values of (c_1, c_2) when the inverse algorithm above is used in two and three steps. The dashed lines $c_2^L(c_1)$ and the solid lines $c_2^U(c_1)$ are calculated from the sets (5.1) and (5.3) when $\epsilon_1 = 3$. The figure shows the tightest bounds on $c_2^L(c_1)$ and $c_2^U(c_1)$, which correspond to the frequencies ω_1 and ω_2. The bounds on (c_1, c_2) get tighter for every step in the algorithm, and additional sets of measurement data improve the bounds on the moments. Table 5.1 shows $(c_n^L)_m$ and $(c_n^U)_m$ when $n = 1, 2, 3$ and $m = 1, 2, 3$. The algorithm gives only bounds on c_1 in the first step, but the inequalities (3.20) and (3.21) always hold. In this case, we have the

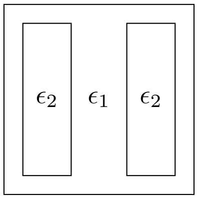


FIG. 5.2. *The geometry used to generate the result shown in Figure* 5.3 *and in Table* 5.2. *Two rods with length* 0.8 *and width* 0.25 *are located, at distance* 0.3 *apart, in a unit square. The volume fraction is then* $c_1 = 0.6$. *The applied field is oriented perpendicular to the rods.*

bounds $-0.25 \leq c_2 \leq 0$ and $0 \leq c_3 \leq 0.126$ after the first step. The bounds on the lower-order moments come closer together for each step in the algorithm.

5.2. Rational functions. A rational effective permittivity ϵ_{e} corresponds in the Stieltjes integral representation (2.2) to a sum of Dirac measures. The rational function ϵ_3^{U} (3.15) is obtained from a sum of two Dirac measures, that we formally write

$$\mathrm{d}m_{3\mathrm{U}}(y) = \frac{c_2^2}{c_3}\delta\left(y + \frac{c_3}{c_2}\right)\mathrm{d}y + \frac{c_1 c_3 - c_2^2}{c_3}\delta(y)\,\mathrm{d}y. \tag{5.4}$$

At the extreme point $c_3 = c_3^{\min} = c_2^2/c_1$ the measure (5.4) corresponds to ϵ_2^{U}, which is the Maxwell–Garnett formula [22]. The arithmetic mean ϵ_1^{U} is obtained from the measure $m_{3\mathrm{U}}$ when $c_2 \to c_2^{\max} = 0$. The moments of the measure (5.4) are

$$c_{n+1} = \frac{c_2^2}{c_3}\left(\frac{c_3}{c_2}\right)^n, \quad n = 1, 2, \ldots. \tag{5.5}$$

Hence, the moments converge exponentially to zero. We use the values $c_1 = 0.5$, $c_2 = -0.125$, and $c_3 = 0.0625$ on the lowest three moments, which equals the three lowest moments in the checkerboard case (5.2). The values on the components (5.1) imply that the effective permittivity $\epsilon_{\mathrm{e}} = \epsilon_3^{\mathrm{U}}$ in the three cases is

$$\epsilon_{\mathrm{e}}(\omega_0) = 3.87 + 1.70\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\omega_1) = 3.72 + 0.024\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\omega_2) = 3.33 + 0.018\mathrm{i}. \tag{5.6}$$

The inverse algorithm gives $0.498 \leq c_1 \leq 0.5001$ in the first step and, to numerical accuracy, the exact values on c_1, c_2, and c_3 after three steps in the algorithm.

5.3. An anisotropic example. Using the same material parameters as above, we also give an example in the anisotropic and periodic case, Figure 5.2. The effective permittivity and the moments $c_1 = 0.6$, $c_2 = -0.125$, and $c_3 = -0.1841$ were for this periodic structure numerically calculated in [10]. The numerical values on the permittivity at the frequencies ω_0, ω_1, and ω_2 are

$$\epsilon_{\mathrm{e}}(\omega_0) = 3.9426 + 0.9852\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\omega_1) = 3.5147 + 0.01554\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\omega_2) = 3.253 + 0.01306\mathrm{i}. \tag{5.7}$$

Figure 5.3 shows the attainable values of (c_1, c_2) after two and three steps in the inverse algorithm.

The dashed lines $c_2^{\mathrm{L}}(c_1)$ and the solid lines $c_2^{\mathrm{U}}(c_1)$ are calculated from the sets (5.1) and (5.7) when $\epsilon_1 = 3$. The figure shows the tightest bounds on $c_2^{\mathrm{L}}(c_1)$ and $c_2^{\mathrm{U}}(c_1)$, which correspond to the frequencies ω_1 and ω_2. Table 5.2 presents the bounds

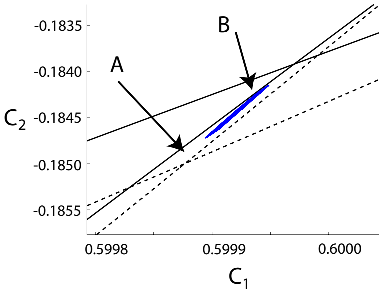


FIG. 5.3. *Bounds on the moments c_1 and c_2 using the anisotropic geometry in Figure* 5.2. *The numerical values on the parameters are $c_1 = 0.6$ and $c_2 = -0.1841$* [10]. *Region A, which is bounded by the two solid lines and the two dotted lines, represents the attainable values of* (c_1, c_2) *after two steps in the algorithm. Region B (shaded) represents the possible values of* (c_1, c_2) *after three steps in the algorithm.*

TABLE 5.2

The table shows the bounds (4.13) *and* (4.14) *using the geometry depicted in Figure* 5.2. *The bounds are on the parameters c_1, c_2, and c_3 when one, two, and three steps in the inverse algorithm are used.*

c_1^L	c_1^U	c_2^L	c_2^U	c_3^L	c_3^U
0.5984	0.6001	-	-	-	-
0.5999	0.6000	-0.1853	-0.1838	-	-
0.5999	0.6000	-0.1847	-0.1841	0.0950	0.0964

(4.13) and (4.14) on the three lowest moments after one, two, and three steps in the algorithm. The inequalities (3.20) and (3.21) give the bounds $-0.25 \leq c_2 \leq 0$ and $0 \leq c_3 \leq 0.149$ after the first step.

The geometry and the values on the components were previously used in [10, 11]. Here we obtain tight bounds on the lowest three moments without any previous knowledge about the structure.

6. Nanocomposites. The effective permittivity belongs to the wedge bounded by the rays $\epsilon_0^{\mathrm{L}} = t\epsilon_1$ and $\epsilon_0^{\mathrm{U}} = t\epsilon_2$, $0 \leq t \leq \infty$ (see [22]). In a metal/dielectric nanocomposite the angle between the rays is large compared to the dielectric case, which results in less tight bounds on the moments c_n. We give an example where the composite is composed of gold and magnesium oxide and measured at the optical wavelengths $\lambda_0 = 300\,\mathrm{nm}$, $\lambda_1 = 500\,\mathrm{nm}$, $\lambda_2 = 700\,\mathrm{nm}$, and $\lambda_3 = 900\,\mathrm{nm}$. The permittivity of gold is [17]

$$\epsilon_1(\lambda_0) = -1.23 + 5.78\mathrm{i}, \quad \epsilon_1(\lambda_1) = -2.27 + 3.81\mathrm{i} \tag{6.1}$$

and

$$\epsilon_1(\lambda_2) = -16.79 + 1.07\mathrm{i}, \quad \epsilon_1(\lambda_3) = -32.00 + 2.04\mathrm{i}. \tag{6.2}$$

The magnesium oxide is lossless at optical wavelengths: $\epsilon_2(\lambda_0) = 3.26$, $\epsilon_2(\lambda_1) = 3.05$, $\epsilon_2(\lambda_2) = 3.00$, and $\epsilon_2(\lambda_3) = 2.98$ [14]. We use the effective permittivity $\epsilon_{\mathrm{e}} = \epsilon_3^{\mathrm{U}}$ and the moments $c_1 = 0.5$, $c_2 = -0.125$, and $c_3 = 0.0625$. The gold-magnesium oxide composite has the effective permittivities

$$\epsilon_{\mathrm{e}}(\lambda_0) = 3.24 + 3.18\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\lambda_1) = 2.73 + 3.48\mathrm{i} \tag{6.3}$$

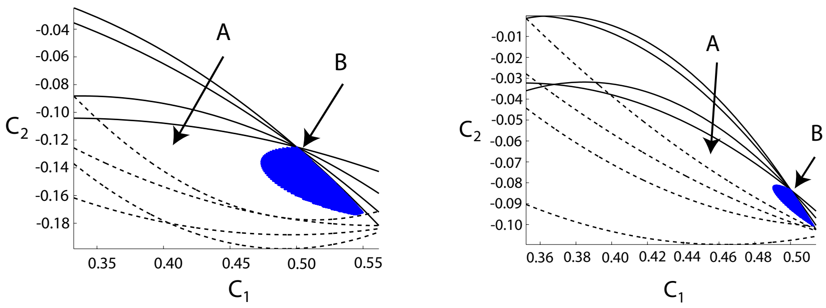


FIG. 6.1. *The nanocomposite: Bounds on the structural parameters c_1 and c_2 using $\epsilon_{\mathrm{e}} = \epsilon_3^{\mathrm{U}}$. Region A, which is bounded by two solid lines and one dotted line, represents the attainable values of (c_1, c_2) after two steps in the algorithm. Region B (shaded) represents the possible values of (c_1, c_2) after three steps in the algorithm. Left: The exact values on the parameters are $c_1 = 0.5$, $c_2 = -0.125$, and $c_3 = 0.0625$. Right: The exact values on the parameters are $c_1 = 0.5$, $c_2 = -0.0833$, and $c_3 = 0.06$.*

TABLE 6.1

The nanocomposite: The table shows the bounds (4.13) *and* (4.14) *using the effective permittivity $\epsilon_{\mathrm{e}} = \epsilon_3^{\mathrm{U}}$ and the moments $c_1 = 0.5$, $c_2 = -0.125$, and $c_3 = 0.0625$. The bounds are on the parameters c_1, c_2, and c_3 when one, two, and three steps in the algorithm are used.*

c_1^{L}	c_1^{U}	c_2^{L}	c_2^{U}	c_3^{L}	c_3^{U}
0.332	0.562	-	-	-	-
0.350	0.554	-0.178	-0.104	-	-
0.473	0.551	-0.174	-0.125	0.0625	0.1015

TABLE 6.2

The nanocomposite: The table shows the bounds (4.13) *and* (4.14) *using the effective permittivity $\epsilon_{\mathrm{e}} = \epsilon_3^{\mathrm{U}}$ and the moments $c_1 = 0.5$, $c_2 = -0.0833$, and $c_3 = 0.06$. The bounds are on the parameters c_1, c_2, and c_3 when one, two, and three steps in the algorithm are used.*

c_1^{L}	c_1^{U}	c_2^{L}	c_2^{U}	c_3^{L}	c_3^{U}
0.352	0.514	-	-	-	-
0.398	0.513	-0.101	-0.037	-	-
0.490	0.513	-0.101	-0.081	0.0593	0.0788

and

$$\epsilon_{\mathrm{e}}(\lambda_2) = 0.625 + 0.314\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\lambda_3) = -0.344 + 0.513\mathrm{i}. \tag{6.4}$$

Figure 6.1 shows the possible values of the lowest two moments c_1 and c_2 after two and three steps in the inverse algorithm. Table 6.1 presents the bounds (4.13) and (4.14) on the moments c_1, c_2, and c_3 after one, two, and three steps in the algorithm. The identical measure is used in section 5.2, where the inverse algorithm determines c_1, c_2, and c_3 from the three sets (5.1) and (5.6) when $\epsilon_1 = 3$ for the frequencies ω_0, ω_1, and ω_2.

The convergence rate of the moments depends strongly on the measure m. Assume that the first three moments are $c_1 = 0.5$, $c_2 = -c_1\tilde{c}_1/3 = -0.0833$, and $c_3 = 0.06$. The effective permittivity from the measure (5.4) is for the gold-magnesium oxide composite. Then

$$\epsilon_{\mathrm{e}}(\lambda_0) = 2.05 + 2.63\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\lambda_1) = 1.64 + 1.97\mathrm{i} \tag{6.5}$$

and

$$\epsilon_{\mathrm{e}}(\lambda_2) = -4.00 + 0.42\mathrm{i}, \quad \epsilon_{\mathrm{e}}(\lambda_3) = -9.92 + 0.79\mathrm{i}. \tag{6.6}$$

Figure 6.1 shows the possible values of the lowest two moments c_1 and c_2 after two and three steps in the inverse algorithm, and Table 6.2 presents the bounds (4.13) and (4.14) on the moments c_1, c_2, and c_3 after one, two, and three steps in the algorithm. If we assume that the material is known to be a three-dimensional isotropic composite, $c_2 = -c_1\tilde{c}_1/3$, the algorithm gives the tighter bounds $0.490 \leq c_1 \leq 0.500$ and $0.060 \leq c_3 \leq 0.063$.

7. Conclusions. We have presented a method to calculate structural parameters (moments) from measured bulk properties of two-component composites, based on the inverse bounds in [10, 11]. The method gives tight bounds on the volume fraction c_1 and also bounds on the higher-order structural parameters c_2 and c_3 after three steps in the algorithm. The tightness of the bounds is sensitive to the geometry and the contrast. The bounds are tighter for low-contrast materials and for anisotropic materials. The results can be improved, e.g., tighter bounds on low-order moments can be calculated, if at least one of the structural parameters is known or additional measurement data is available.

The presented method can be extended to handle inaccurate measurement data. Importantly, the method in this paper can be used together with arbitrary large error bars on the measurement data and will still produce correct bounds on the structural parameters. If the error bars are too large, however, the method can only reproduce the fundamental bounds on c_n. A first step in this direction was considered in [12] where an algorithm based on the inverse bounds in [11] was used.

REFERENCES

[1] G. A. Baker, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] D. J. Bergman, *The dielectric constant of a composite material—A problem in classical physics*, Phys. Rep., 43 (1978), pp. 377–407.

[3] D. J. Bergman, *Exactly solvable microscopic geometries and rigorous bounds for the complex dielectric constant of a two-component composite material*, Phys. Rev. Lett., 44 (1980), pp. 1285–1287.

[4] D. J. Bergman, *Rigorous bounds for the complex dielectric constant of a two component composite*, Ann. Phys., 138 (1982), pp. 78–114.

[5] D. J. Bergman, *Hierarchies of Stieltjes functions and their application to the calculation of bounds for the dielectric constant of a two-component composite medium*, SIAM J. Appl. Math., 53 (1993), pp. 915–930.

[6] E. Cherkaeva, *Inverse homogenization for evaluation of effective properties of a mixture*, Inverse Problems, 17 (2001), pp. 1203–1218.

[7] E. Cherkaeva and K. M. Golden, *Inverse bounds for microstructural parameters of composite media derived from complex permittivity measurements*, Waves in Random Media, 8 (1998), pp. 437–450.

[8] E. Cherkaeva and A. C. Tripp, *Inverse conductivity for inaccurate measurements*, Inverse Problems, 12 (1996), pp. 869–883.

[9] A. R. Day and M. F. Thorpe, *The spectral function of composites: The inverse problem*, J. Phys. Condens. Matter, 11 (1999), pp. 2551–2568.

[10] C. Engström, *Bounds on the effective tensor and the structural parameters for anisotropic two-phase composite material*, J. Phys. D Appl. Phys., 38 (2005), pp. 3695–3702.

[11] C. Engström, *Inverse bounds and bulk properties of complex-valued two-component composites*, SIAM J. Appl. Math., 67 (2006), pp. 194–213.

[12] C. Engström, *Structural information of nanocomposites from measured optical properties*, J. Phys. Condens. Matter, 19 (2007), paper 106212.

[13] D. A. Field, *Series of Stieltjes, Padé approximants and continued fractions*, J. Math. Phys., 17 (1976), pp. 843–884.

[14] J. I. Gittleman, B. Abeles, P. Zanzucchi, and Y. Arie, *Optical properties and selective solar absorption of composite material films*, Thin Solid Films, 45 (1977), pp. 9–18.

[15] K. Golden and G. Papanicolaou, *Bounds for effective parameters of heterogeneous media by analytic continuation*, Comm. Math. Phys., 90 (1983), pp. 473–491.

[16] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[17] P. B. Johnson and R. W. Christy, *Optical constants of the noble metals*, Phys. Rev. B, 6 (1972), pp. 4370–4379.

[18] R. C. McPhedran, D. R. McKenzie, and G. W. Milton, *Extraction of structural information from measured transport properties of composites*, Appl. Phys. A, 29 (1982), pp. 19–27.

[19] R. C. McPhedran and G. W. Milton, *Inverse transport problems for composite media*, Material Resources Society, Symposium Proceedings, 195 (1990), pp. 257–277.

[20] G. W. Milton, *Bounds on the complex dielectric constant of a composite material*, Appl. Phys. Lett., 37 (1980), pp. 300–302.

[21] G. W. Milton, *Bounds on the transport and optical properties of two-component composite material*, J. Appl. Phys., 52 (1981), pp. 5294–5304.

[22] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.

[23] S. Prager, *Improved variational bounds on some bulk properties of a two-phase random medium*, J. Chem. Phys., 50 (1969), pp. 4305–4312.

[24] S. Torquato, *Random Heterogeneous Materials: Microstructure and Microscopic Properties*, Springer-Verlag, Berlin, 2002.

[25] E. Tuncer, *Extracting the spectral density function of a binary composite without a priori assumption*, Phys. Rev. B, 71 (2005), paper 012101.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 680–693

TRAVELING WAVES IN A BIOREMEDIATION MODEL*

SHANGBING AI†

Abstract. We study a bioremediation model that arises in restoring ground water and soil contaminated with organic pollutants. It describes an in situ bioredimedation scenario in which a sorbing substrate of contaminated soil is degraded by indigenous microorganisms in the presence of an injected nonsorbing electron acceptor. The model relates to the coupling of the advection, dispersion, and biological reaction simultaneously for the substrate, electron acceptor, and the total biomass by two advection-reaction-diffusion equations and an ODE. We establish the existence of traveling waves for the model with wider classes of kinetic functions. Our result generalizes previous results for this model which were established only for multiplicative Monod kinetics. In addition, the proof of our result, which is based on a dynamical systems approach, is simpler.

Key words. biodegradation model, traveling waves, shooting argument

AMS subject classifications. 34B15, 34C37, 35K57, 92C45

DOI. 10.1137/070685865

1. Introduction.

In situ bioremediation makes use of microorganisms to transfer hazardous chemicals into nontoxic products in places where contaminants are a concern. Because it is less costly, safer, and faster than conventional treatment methods, this technology has found wide applications in the cleanup of groundwater and soils polluted with chlorinated solvents, fuel hydrocarbons, and explosives [8, 19, 20]. Through the injection of growth nutrients and electron acceptors into the contaminated site, the organic pollutants serving as substrates are broken down through the metabolism of the microorganisms. This is a complex process which involves physical, chemical, and biological reactions as well as interactions between microorganisms and the physical condition of the subsurface, such as fluid flow rates, media heterogeneity, and contaminant availability. The success of bioremediation is determined by key factors that control or limit the performance of the microorganisms. Mathematical modeling has been extensively used for understanding this process and identifying these key factors therefore to provide guidance in the improvement of in situ bioremediation technology [2, 3, 4, 7]. In recent years, analytical study of traveling wave solutions for various bioremediation models becomes of interest to researchers in this area [1, 9, 10, 12, 16]. Via analyzing the traveling waves, one can answer questions such as: how fast does the incoming nutrient front travel? what is the contaminant removal rate? which parameters are most significant in controlling the degradation rate? etc.

In this paper, we consider traveling waves for a basic one-dimensional bioremediation model that has been discussed in papers of Odencrantz [13, 14], Odencrantz, Valocchi, and Rittman [15], Valocchi, Odencrantz, and Rittman [22], and Oya and Valocchi [16]. The model relates to the coupling of the advection, dispersion, and biological reaction simultaneously for the substrate, electron acceptor, and the total biomass. It is assumed (cf. [1]) that (i) microbes are attached to the aquifer particles in the soil and thus do not move, (ii) the acceptor is nonsorbing and thus travels

*Received by the editors March 20, 2007; accepted for publication (in revised form) August 29, 2007; published electronically December 19, 2007.

http://www.siam.org/journals/siap/68-3/68586.html

†Department of Mathematical Sciences, University of Alabama in Huntsville, Huntsville, AL 35899 (ais@email.uah.edu).

through the soil at pore water velocity v, and (iii) the substrate is sorbing and travels at the retarded velocity $1/R$ with the retardation factor $R > 1$. Let $S(x,t)$, $A(x,t)$, and $M(x,t)$ be the concentrations of the substrate, acceptor, and microorganisms, respectively. Then the mass balance equations for S, A, and M are

$$\left\{\begin{array}{l} S_t = \frac{d}{R}S_{xx} - \frac{v}{R}S_x - \frac{1}{R}f(S,A,M), \\ A_t = dA_{xx} - vA_x - \gamma f(S,A,M), \\ M_t = -b(M - M_0) + \beta f(S,A,M), \end{array}\right. \tag{1.1}$$

where d is the hydrodynamic dispersion coefficient, γ is the coefficient equal to the mass of A utilized by the biomass per unit mass of the substrate degraded, b is the cell decay coefficient for the biomass M, M_0 is the neutral background biomass concentration, β is the cell yield coefficient for the electron donor, and the reaction function f describes the biodegradation rate, which has been taken to be the so-called multiplicative Monod kinetics

$$f = f(S,A,M) = qM\left(\frac{S}{K_S + S}\right)\left(\frac{A}{K_A + A}\right), \tag{1.2}$$

where q is the maximum specific rate of the substrate utilization, and K_S, K_A are the half-maximum rate concentrations of the substrate S and the acceptor A.

In the above model, the condition $R > 1$ physically means that the advective velocity of the substrate S is slower than that of the acceptor A so that there is a region overlap where the two concentrations mix. In this region, known as a biologically active zone (BAZ), microorganisms actively grow by consuming the nutrient and degrading the substrates and the three components travel together. It is Oya and Valocchi [16] who, via numerical simulations, first observed traveling waves with constant wave speeds that consist of monotonically increasing fronts in S, monotonically decreasing fronts in A, and pulses for M. They then analytically studied the traveling wave solutions of (1.1) of the form $(S(\xi), A(\xi), M(\xi))$ with a constant wave speed c ($\xi = x - ct$) that satisfy the boundary value problem

$$\left\{\begin{array}{l} dS_{\xi\xi} + (Rc - v)S_\xi = f, \\ dA_{\xi\xi} + (c - v)A_\xi = \gamma f, \\ cM_\xi = b(M - M_0) - \beta f, \end{array}\right. \tag{1.3}$$

and

$$(S,A,M)(-\infty) = (0, A_0, M_0), \quad (S,A,M)(\infty) = (S_0, 0, M_0), \tag{1.4}$$

and $S(\xi) > 0$, $A(\xi) > 0$, and $M(\xi) > 0$ for all $\xi \in \mathbb{R}$, where A_0 and S_0 are positive constants, representing the input of the nutrient concentration and the output of the pollutant concentration. Note that the background bacteria population M_0 is the unique equilibrium between cell growth and decay. The asymptotic conditions (1.4) can be explained as follows (cf. [1]): at $-\infty$, behind the BAZ, the substrate has been completely degraded, the acceptor level is equal to its injection level, and the microorganism population has returned to its equilibrium level; at ∞, ahead of the BAZ, the soil remains undisturbed and contaminated, and thus the substrate, acceptor, and microorganisms are all equal to their initial levels. By formally adding the appropriate multiples of the first two equations of (1.3) and integrating over $\mathbb{R}$, Oya and Valocchi obtained the following explicit formula for c:

$$c = \frac{v(A_0 + \gamma S_0)}{A_0 + \gamma R S_0}. \tag{1.5}$$

Using this formula they could predict the removal rate of the substrate and determine the important parameters that control this rate.

Subsequently, Murray and Xin [12] established rigorously the existence of solutions of (1.3)–(1.4) under the condition $R > 1$. They also obtained the following estimates for the biomass M:

$$M_0 < M(\xi) \leq M_0 + \frac{\beta(R-1)A_0S_0}{A_0 + \gamma S_0}. \tag{1.6}$$

Note that the upper bound for M is independent of the reaction function f. Recently, Beck, Doelman, and Kaper [1] reestablished the existence of solutions of (1.3)–(1.4) in the case that both half saturation constants K_S and K_A are sufficiently large. Using a change of coordinates, they reduced the problem into a singular perturbed one, for which they were able to apply the theory of the geometric singular perturbations. As they noted, their method provides further insight into the geometric structure of the traveling wave solutions, which are useful in studying the stability of these wave fronts.

The above three references [1, 12, 16] all considered the case where the reaction function f takes the form in (1.2). However, the bioreaction in bioremediation is not limited to the Monod equation. Instead, depending on the nature of the pollutants, the kinetic term in the model may take other forms. For example, when microbial growth is inhibited by the substrate, such as phenol degradation [11, 17, 21], the biodegradation function f can be described by the Haldane inhibition model

$$f = qM\left(\frac{S}{K_S + S + S^2/K_I}\right)\left(\frac{A}{K_A + A}\right),$$

where K_I is the inhibition constant. Biodegradation of certain organic compounds is also described by the Moser equation [6]

$$f = qM\left(\frac{S^n}{K_S + S^n}\right)\left(\frac{A}{K_A + A}\right),$$

where n is the order of the enzyme reaction. The first order kinetic function $f = qMSA$ was discussed in [14]. There are numerous other kinetic models that are available. Generally speaking, the selection of a kinetic model appropriate to the conditions of the site being modeled is difficult because the differences among the models are not fully understood.

In this paper, we extend the works in [1, 12, 16] to a large class of kinetic models, including those mentioned above. We prove that there exist traveling waves for all of these kinetic models. This indicates biologically that the system (1.1) can be applied to model the bioremediation of broader categories of organic pollutants. Our main result is as follows.

Theorem 1.1. *Assume that $R > 1$ and that f satisfies*

$$\begin{cases} f(S, A, M) = \varphi_1(S)\varphi_2(A)\varphi_3(M), \\ f \text{ is } C^2 \text{ in a neighborhood of } (S_0, 0, M_0), \\ \varphi_i \text{ is locally Lipschitz on } (0, \infty) \quad (i = 1, 2, 3), \\ \varphi_i > 0 \quad \text{on } (0, \infty) \quad (i = 1, 2, 3), \\ \varphi_1(0) = \varphi_2(0) = 0. \end{cases} \tag{1.7}$$

Then (1.3)–(1.4) *admits a traveling wave solution* (S, A, M, c) *with* c *given in* (1.5). *Moreover,* M *satisfies* (1.6) *and, for all* $\xi \in \mathbb{R}$,

$$\left\{\begin{array}{ll} 0 < S(\xi) < S_0, & 0 < S'(\xi) < \frac{v(R-1)A_0S_0}{d(A_0+\gamma RS_0)}, \\ 0 < A(\xi) < A_0, & -\frac{\gamma v(R-1)A_0S_0}{d(A_0+\gamma RS_0)} < A'(\xi) < 0. \end{array}\right. \tag{1.8}$$

We prove Theorem 1.1 in the next section by a different approach from those in [1, 12]. The approach in [12] is to apply the Leary–Schauder degree theory to a regularized system with Dirichlet boundary conditions on a large but finite interval and then to pass to the infinite line and remove the regularization. This method involves many rather delicate a priori estimates that apparently depend on the specific form of f given in (1.2). Our approach is a dynamical systems approach; namely, we directly study flows of (1.3) in the phase space with an aid of the stable manifold theorem and then apply a simple shooting argument. This approach is commonly used for planar systems and, generally speaking, is difficult for higher dimensional systems. For the system (1.3), it turns out that this approach is simpler and yields a much shorter proof. An outline of the proof is presented after the proof of Lemma 2.2. We conclude this paper with a short summary.

Before ending this section we give the following remark. The problem (1.3)–(1.4) does not have any positive solutions if $0 < R \leq 1$. To see this, we assume on the contrary that (1.3)–(1.4) has a positive solution (S, A, M). Note that $c - v \geq 0$ from (1.5). Using the second equation of (1.3) and $A(-\infty) = 0$, we obtain

$$A'(\xi) = \frac{\gamma}{d}\int_{-\infty}^{\xi} e^{-\frac{c-v}{d}(\xi-\eta)} f(S(\eta), A(\eta), M(\eta))\, d\eta > 0 \qquad \forall \xi \in \mathbb{R},$$

which yields that $A(\infty) = 0$ is impossible.

2. Proof of Theorem 1.1. Throughout this section, we assume that $R > 1$. Let c be defined in (1.5). We have that

$$b_1 := Rc - v = \frac{v(R-1)A_0}{A_0+\gamma RS_0} > 0, \quad b_2 := -(c-v) = \frac{\gamma v(R-1)S_0}{A_0+\gamma RS_0} = \frac{\gamma S_0}{A_0}b_1 > 0.$$

Employing the rescalings

$$\begin{array}{l} \tilde{b}_1 = \dfrac{b_1}{d}, \quad \alpha = \dfrac{1}{S_0 d}, \quad \tilde{b}_2 = \dfrac{b_2}{d}, \quad \tilde{\gamma} = \dfrac{\gamma}{A_0 d}, \quad b_3 = \dfrac{b}{c}, \quad \tilde{\beta} = \dfrac{\beta}{M_0 c}, \\ \tilde{S} = \dfrac{S}{S_0}, \quad \tilde{A} = \dfrac{A}{A_0}, \quad \tilde{M} = \dfrac{M}{M_0}, \quad \tilde{f}(\tilde{S}, \tilde{A}, \tilde{M}) = f(S, A, M), \quad z = -\xi, \end{array}$$

we reduce (1.3) and (1.4) into (after dropping the tildes)

$$\left\{\begin{array}{l} S'' - b_1S' = \alpha f, \\ A'' + b_2A' = \gamma f, \\ M' = -b_3(M-1) + \beta f, \end{array}\right. \tag{2.1}$$

and

$$(S, A, M)(-\infty) = (1, 0, 1), \quad (S, A, M)(\infty) = (0, 1, 1), \tag{2.2}$$

where "prime" is d/dz. Therefore, in order to prove Theorem 1.1, it suffices to establish the following theorem.

THEOREM 2.1. *Assume that b_1, b_2, b_3, α, β, and γ are positive constants satisfying $\alpha b_2 = \gamma b_1$ and that f satisfies* (1.7) *with $S_0 = A_0 = M_0 = 1$. Then* (2.1)–(2.2) *admits a solution* (S, A, M) *such that, for* $z \in \mathbb{R}$,

$$(2.3)\qquad \begin{cases} 0 < S(z) < 1, \quad -b_1 < S'(z) < 0, \\ 0 < A(z) < 1, \quad 0 < A'(z) < b_2, \\ 1 < M(z) < 1 + \beta b_1/\alpha. \end{cases}$$

We remark that if the diffusion coefficients in the first two equations of (1.3) are distinct, then with slight modifications of the above rescalings we can still reduce (1.3) into the system (2.1), yielding that Theorem 1.1 also holds in this more general case. Noting that all bounds in (2.3) are independent of f, this observation plays a key role in the last part of the proof of Theorem 2.1.

In order to prove Theorem 1.1, we need the following lemma.

LEMMA 2.2. (i) *If* (S, A, M) *is a solution of* (2.1)–(2.2), *then* (S, T, A, M) *is a solution of*

$$(2.4)\qquad \begin{cases} S' = T, \\ T' = \alpha f + b_1 T, \\ A' = -(\gamma b_1/\alpha)(S-1) + (\gamma/\alpha)T - b_2 A, \\ M' = \beta f - b_3(M-1), \end{cases}$$

and

$$(2.5)\qquad (S, T, A, M)(-\infty) = (1, 0, 0, 1), \quad (S, T, A, M)(\infty) = (0, 0, 1, 1).$$

Conversely, if (S, T, A, M) *is a solution of* (2.4)–(2.5), *then* (S, A, M) *is a solution of* (2.1)–(2.2).

(ii) *The line* $T = b_1(S-1)$, $A = 0$, *and* $M = 1$ *in the* (S, T, A, M)*-phase space lies on the unstable manifold* W^u *of* (2.4) *at* $(1, 0, 0, 1)$.

(iii) *If* (S, T, A, M) *is a solution of* (2.4) *that satisfies the boundary condition at* $-\infty$ *in* (2.5) *with its maximal existence interval* $(-\infty, \omega)$, *then, for* $z \in (-\infty, \omega)$,

$$(2.6)\qquad \begin{cases} T(z) - b_1[S(z) - 1] = \alpha \int_{-\infty}^{z} f(S(\eta), A(\eta), M(\eta))\, d\eta, \\ A'(z) = \gamma \int_{-\infty}^{z} e^{-b_2(z-\eta)} f(S(\eta), A(\eta), M(\eta))\, d\eta, \\ A(z) = (\gamma/b_2) \int_{-\infty}^{z} [1 - e^{-b_2(z-\eta)}] f(S(\eta), A(\eta), M(\eta))\, d\eta, \\ M(z) = 1 + \beta \int_{-\infty}^{z} e^{-b_3(z-\eta)} f(S(\eta), A(\eta), M(\eta))\, d\eta. \end{cases}$$

If, in addition, $S(z) > 0$, $T(z) < 0$ *for all* $z \in (-\infty, \omega)$ *and* $A(z) > 0$ *for all sufficiently negative* z, *then* $\omega = \infty$, $(S, T, A, M)(\infty) = (0, 0, 1, 1)$ *and, for* $z \in \mathbb{R}$,

$$(2.7)\qquad \begin{cases} 0 < S(z) < 1, \quad b_1[S(z) - 1] < T(z) < 0, \\ 0 < A(z) < \frac{\gamma b_1}{\alpha b_2}[1 - S(z)], \quad 0 < A'(z) < \frac{\gamma b_1}{\alpha}[1 - S(z)], \\ 1 < M(z) < 1 + \frac{\beta b_1}{\alpha}[1 - S(z)], \end{cases}$$

and, in particular, (2.3) *holds.*

Proof. (i) and (ii) can be verified directly. Equation (2.6) follows directly from (2.4) and the boundary condition at $-\infty$ in (2.5). We now show (2.7). First, the assumptions in this part imply $f(S(z), A(z), M(z)) > 0$ for all sufficiently negative z which together with the second, third, and fourth equations in (2.6) yields that $A(z) > 0$, $A'(z) > 0$, $M(z) > 1$, and $f(S(z), A(z), M(z)) > 0$ for all $z \in (-\infty, \omega)$. Consequently, the first equation in (2.6) yields the first inequality for T in (2.7) and $\int_{-\infty}^{z} f(S(\eta), A(\eta), M(\eta))\, d\eta < (b_1/\alpha)[1 - S(z)]$ which together with the second, third, and fourth equations in (2.6) yields the upper bounds for A, A', and M in (2.7). This shows (2.7). Finally, using these estimates and the fact that $(0,0,1,1)$ is an equilibrium point of (2.4), we conclude that $\omega = \infty$ and $(S,T,A,M)(\infty) = (0,0,1,1)$. $\square$

Now we outline the proof of Theorem 2.1 given later. From Lemma 2.2(i) it suffices to show that the problem (2.4)–(2.5) has a solution. We start with analyzing the local dynamics of (2.4) at $(1,0,0,1)$. Simple algebra yields, at $(1,0,1)$, $D_S f = \varphi_1'(1)\varphi_2(0)\varphi_3(1) = 0$, $D_M f = \varphi_1(1)\varphi_2(0)\varphi_3'(1) = 0$, and

$$\sigma := D_A f = \varphi_1(1)\varphi_2'(0)\varphi_3(1) \geq 0, \tag{2.8}$$

and so the coefficient matrix for the linearized system of (2.4) at $(1,0,0,1)$ is

$$E = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & b_1 & \alpha\sigma & 0 \\ -\gamma b_1/\alpha & \gamma/\alpha & -b_2 & 0 \\ 0 & 0 & \beta\sigma & -b_3 \end{pmatrix},$$

and the eigenvalues of E are

$$\begin{cases} \lambda_1 := b_1, \\ \lambda_2 := \frac{1}{2}\left(-b_2 + \sqrt{b_2^2 + 4\gamma\sigma}\right), \\ \lambda_3 := \frac{1}{2}\left(-b_2 - \sqrt{b_2^2 + 4\gamma\sigma}\right), \\ \lambda_4 = -b_3. \end{cases} \tag{2.9}$$

It follows that $\lambda_3 < 0$, $\lambda_4 < 0$, $\lambda_1 > 0$, and $\lambda_2 > 0$ if $\sigma > 0$ and $\lambda_2 = 0$ if $\sigma = 0$. Hence, (2.4) has a two-dimensional local unstable manifold W^u_{loc} at $(1,0,0,1)$ if $\sigma > 0$ and a one-dimensional unstable manifold and one-dimensional center manifold if $\sigma = 0$. Note that, from (2.8), $\sigma = 0$ if and only if $\varphi_2'(0) = 0$. We proceed the proof of Theorem 2.1 in three cases based on $0 < \lambda_2 < \lambda_1$, or $\lambda_1 < \lambda_2$, or either $\lambda_2 = 0$ or $\lambda_1 = \lambda_2$.

Case 1. We consider $0 < \lambda_2 < \lambda_1$. Then there is a unique one-dimensional strongly unstable manifold on W^u_{loc} that is tangent to the eigenvector p_1 at $(1,0,0,1)$ and indeed lies on the line $T = b_1(S-1)$, $A = 0$, and $M = 1$ from Lemma 2.2(ii); the rest flows on W^u_{loc} are all tangent to the eigenvector p_2 at $(1,0,0,1)$. We take a continuous "circular arc" Γ_1 on W^u_{loc} whose projection on the (S,T)-plane is displayed in the left figure in Figure 1, and then show that (i) the components $S(z)$, $T(z)$, and $A(z)$ of every solution starting on Γ_1 at $z = 0$ satisfy $S(z) < 1$, $T(z) < 0$, and $A(z) > 0$ for all sufficiently negative z; (ii) for each solution starting near one end of Γ_1 with $T(0) > 0$, there exists a $z_0 \in \mathbb{R}$ such that $T(z_0) = 0$, $T'(z_0) > 0$, $T(z) < 0$ for $z < z_0$ and $S(z) > 0$ for $z \leq z_0$ (i.e., $T = 0$ occurs before $S = 0$ does); (iii) for any solution starting near the other end of Γ_1 with $T(0) < 0$, there is a $\bar{z}_0$ such that $S(\bar{z}_0) = 0$, $S(z) > 0$ for $z < \bar{z}_0$, and $T(z) < 0$ for $z \leq \bar{z}_0$ (i.e., $S = 0$ occurs before $T = 0$ does). Using the fact that $S = 0$ and $T = 0$ cannot occur at the same time, we conclude by a shooting argument that there is a point on Γ_1 such that the solution of (2.4) starting at this point satisfies $S > 0$ and $T < 0$ on $(-\infty, \omega)$, which gives a desired solution by Lemma 2.2(iii).

Case 2. We consider $\lambda_1 < \lambda_2$. Then the unique one-dimensional strongly unstable manifold on W^u_{loc} is tangent to the eigenvector p_2 at $(1,0,0,1)$, and the rest flows on W^u_{loc} are all tangent to the eigenvector p_1 at $(1,0,0,1)$. We take a continuous "circular

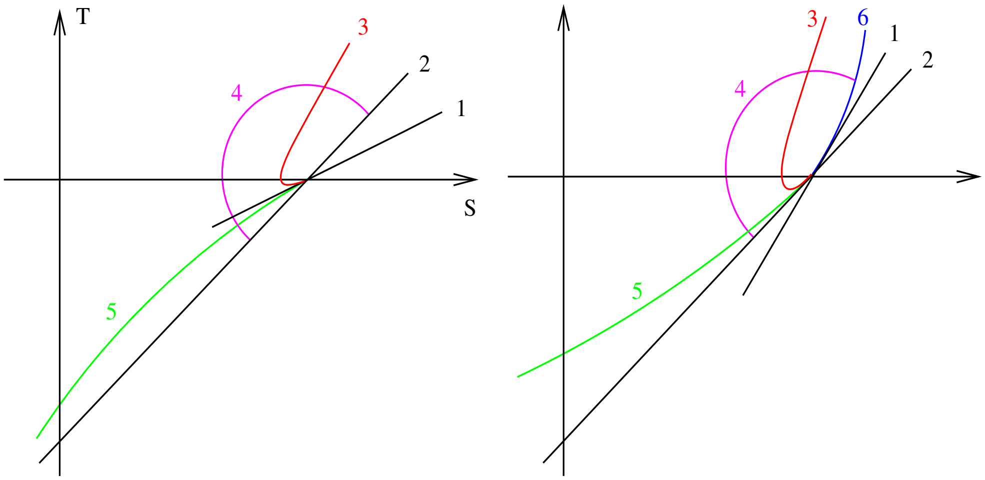


FIG. 1. *In the left and right figures, the curves sketched are the projections of the curves in the (S,T,A,W)-phase space into the (S,T)-plane corresponding to $0<\lambda_2<\lambda_1$ and $\lambda_1<\lambda_2$, respectively: curve* 1 *is the projection of the eigenvector p_2, curve* 2 *is the line $T=b_1(S-1)$, which is parallel to the projection of the eigenvector p_1, curve* 4 *is the projection of Γ_1 (resp., Γ_2) which is taken on W^u_{loc}, curves* 3 *and* 5 *are the projections of two solution curves of* (2.4) *passing through Γ_1 (resp., Γ_2) at $z=0$, and curve* 6 *in the right figure is the projection of the strongly unstable manifold which is tangent to p_2.*

arc" Γ_2 on W^u_{loc} whose projection on the (S,T)-plane is displayed in the right figure in Figure 1, and then show that the statements (i), (ii), and (iii) in Case 1 hold after all Γ_1 there are replaced by Γ_2. Then using the same shooting argument as used in Case 1 shows that there is a point on Γ_2 such that the solution of (2.4) starting at this point gives a desired solution in this case.

Case 3. We consider either $\lambda_2=0$ or $\lambda_1=\lambda_2$. We first construct approximating systems to (2.4) with each positive integer n

$$\left\{\begin{array}{l} S'=T,\\ T'=b_1T+\alpha f_n,\\ A'=-(\gamma b_1/\alpha)(S-1)+(\gamma/\alpha)T-b_2A,\\ M'=-b_3(M-1)+\beta f_n, \end{array}\right. \tag{2.10}$$

where $f_n(S,A,M)=\varphi_1(S)[\varphi_2(A)+\frac{1}{n}A]\varphi_3(M)$, which clearly satisfies (1.7) with $S_0=A_0=M_0=1$. Let $\sigma_n:=\varphi_1(1)[\varphi_2'(0)+\frac{1}{n}]\varphi_3(1)=\sigma+\frac{1}{n}\varphi_1(1)\varphi_3(1)$, $\lambda_{1,n}:=b_1$, and $\lambda_{2,n}:=\frac{1}{2}\left(-b_2+\sqrt{b_2^2+4\gamma\sigma_n}\right)$. It follows that $\lambda_{1,n}<\lambda_{2,n}$ if $\lambda_1=\lambda_2$ and $n\geq 1$, and $0<\lambda_{2,n}<\lambda_{1,n}$ if $\lambda_2=0$ and $n\geq n_0$ for some $n_0>0$. Note that $\lambda_{1,n}$ and $\lambda_{2,n}$ are two positive eigenvalues of the linearized system of (2.10) at $(1,0,0,1)$. Then applying the results from Cases 1 and 2 to (2.10) for each $n\geq n_0$ yields a sequence of solutions (S_n,T_n,A_n,M_n) of (2.10) which have the same estimates given in (2.3) that are independent of n. Then, applying the Arzela–Ascoli theorem (cf. [5]) on the interval $[-k,k]$ for each positive integer k and then using a diagonal selection process yields a subsequence $(S_{n_k},T_{n_k},A_{n_k},M_{n_k})$ that converges uniformly on any compact subset of $\mathbb{R}$, whose limit function gives a desired solution of (2.4) in this case.

Proof of Theorem 2.1. As discussed previously, we first assume that λ_1 and λ_2 are both positive and $\lambda_1\neq\lambda_2$. We then find that

$$p_1 = \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix}, \qquad p_2 = \begin{pmatrix} 1 \\ \lambda_2 \\ \frac{\lambda_2(\lambda_2-\lambda_1)}{\alpha\sigma} \\ \frac{\beta\lambda_2(\lambda_2-\lambda_1)}{\alpha(\lambda_2+b_3)} \end{pmatrix}$$

are the eigenvectors of E associated with λ_1 and λ_2, respectively. It follows from the stable manifold theorem that (2.4) has a two-dimensional local unstable manifold W^u_{loc} near $(1,0,0,1)$ which is tangent to the plane spanned by p_1 and p_2. More precisely, we make change of variables

$$\begin{pmatrix} S-1 \\ T \\ A \\ M-1 \end{pmatrix} = P \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} := [p_1, p_2, p_3, p_4] \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix},$$

where p_3 and p_4 are (generalized) eigenvectors of E associated with λ_3 and λ_4, respectively, transform (2.4) into

$$x' = (P^{-1}EP)x + G(x), \qquad P^{-1}EP := \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}, \tag{2.11}$$

where $Q_1 = \left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)$ and $G(x) = O(|x|^2)$, and correspondingly transform W^u_{loc} into a two-dimensional local unstable manifold $\widetilde{W}^u_{loc}$ of (2.11) at $x = 0$ which is tangent to the (x_1, x_2)-plane. Consequently, $\widetilde{W}^u_{loc}$ can be written as the graph of

$$x_3 = h_1(x_1, x_2) = O(x_1^2 + x_2^2), \qquad x_4 = h_2(x_1, x_2) = O(x_1^2 + x_2^2),$$

where x_1 and x_2 are sufficiently small, and the x_1 and x_2 components of the flows of (2.11) on $\widetilde{W}^u_{loc}$ satisfy a planar system

$$x_1' = \lambda_1 x_1 + N_1(x_1, x_2), \qquad x_2' = \lambda_2 x_2 + \widetilde{N}_2(x_1, x_2)x_2, \tag{2.12}$$

where $N_1(x_1, x_2) = O(x_1^2 + x_2^2)$ and $\widetilde{N}_2(x_1, x_2) = O(\sqrt{x_1^2 + x_2^2})$. We note that the factor x_2 in the nonlinear term $\widetilde{N}_2(x_1, x_2)x_2$ is due to the fact that x_1-axis in the x space corresponds to the line $T = \lambda_1(S-1)$, $A = 0$, and $M = 1$ in the (S, T, A, M) space (indeed, this fact also implies that $h_1(x_1, 0) = 0$ and $h_2(x_1, 0) = 0$ so that we can write $h_1(x_1, x_2) = \tilde{h}_1(x_1, x_2)x_2$ and $h_2(x_1, x_2) = \tilde{h}_2(x_1, x_2)x_2$, correspondingly). To proceed further, we need to distinguish two cases based on whether $0 < \lambda_2 < \lambda_1$ or $\lambda_1 < \lambda_2$.

Case 1. Assume that $0 < \lambda_2 < \lambda_1$. We first fix a $\delta > 0$ so small that the set $x_1^2 + x_2^2 \leq \delta^2$ is negatively invariant for (2.12), the unstable manifold of (2.4)

$$W^u_{loc}(\delta) := \left\{ \begin{pmatrix} S \\ T \\ A \\ M \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} + P \begin{pmatrix} x_1 \\ x_2 \\ h_1(x_1, x_2) \\ h_2(x_1, x_2) \end{pmatrix} : x_1^2 + x_2^2 \leq \delta^2 \right\} \tag{2.13}$$

is negatively invariant and, furthermore, $S > 0$ and $M > 0$ if $(S, T, A, M) \in W^u_{loc}(\delta)$. Hence, if $(x_1(z), x_2(z))$ is a solution of (2.12) with $x_1^2(0) + x_2^2(0) \leq \delta^2$, then, for all $z \in (-\infty, 0]$, $x_1^2(z) + x_2^2(z) \leq \delta^2$, and

$$\begin{pmatrix} S(z)-1 \\ T(z) \\ A(z) \\ M(z)-1 \end{pmatrix} = x_1(z)p_1 + x_2(z)p_2 + O\Big(x_1^2(z) + x_2^2(z)\Big) \tag{2.14}$$

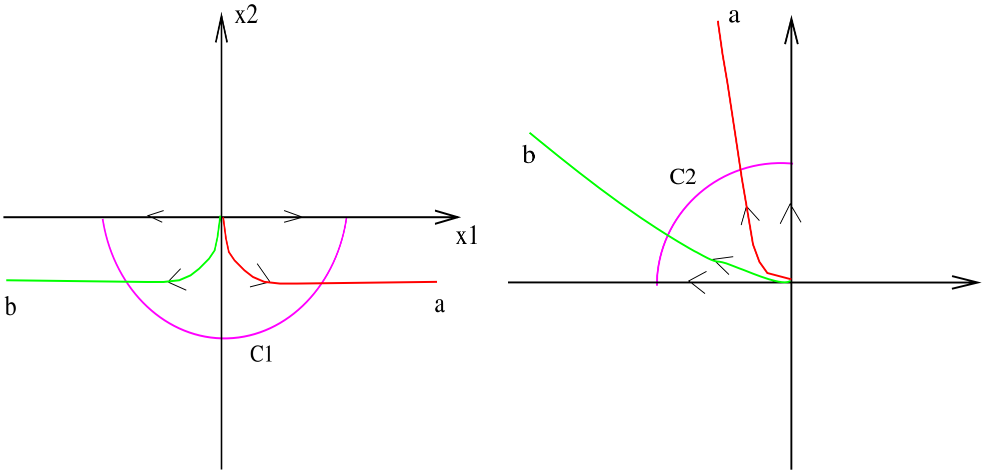


FIG. 2. *In the left figure, curve C_1 and four arrowed solution curves of system* (2.12) *are sketched in the (x_1, x_2)-plane: C_1 corresponds to curve* 4 *in the left figure of Figure* 1*, curve* a *to curve* 3*, curve* b *to curve* 5*, x_1-axis to curve* 2*, and x_2-axis to curve* 1*. In the right figure, curve C_2 and four arrowed solution curves of system* (2.17) *are sketched in the $(\tilde{x}_1, x_2)$-plane: C_2 corresponds to curve* 4 *in the right figure of Figure* 1*, curve* a *to curve* 3*, curve* b *to curve* 5*, $\tilde{x}_1$-axis to curve* 2*, and x_2-axis to curve* 6*.*

is a solution of (2.4) lying on $W^u_{loc}(\delta)$ for $z \in (-\infty, 0]$.

We now consider the solutions of (2.12) with initial values $(x_{10}(\theta), x_{20}(\theta)) = (\delta\cos\theta, \delta\sin\theta)$ with $\theta \in [-\pi, 0]$ on the lower half of the circle C_1: $x_{10}^2 + x_{20}^2 = \delta^2$; see the left figure in Figure 2. One can verify easily that the image of C_1 under the linear mapping P^{-1} defines the continuous curve Γ_1 as described in the outline of the proof of Theorem 2.1. Note that, in this case, the x_1-axis is the strongly one-dimensional unstable manifold of (2.12) (correspondingly, the line $T = b_1(S-1)$, $A = 0$, and $M = 1$ is the strongly one-dimensional unstable manifold of (2.4)). Therefore, for $\theta \in (-\pi, 0)$, we obtain from (2.12) that, as $z \to -\infty$,

$$\begin{pmatrix} x_1(z,\theta) \\ x_2(z,\theta) \end{pmatrix} = \nu_\theta e^{\lambda_2 z}\left[\begin{pmatrix} 0 \\ 1 \end{pmatrix} + O(e^{\lambda_2 z})\right] \tag{2.15}$$

for some $\nu_\theta \neq 0$. It then follows from (2.14) that, as $z \to -\infty$,

$$\begin{pmatrix} S_\theta(z) - 1 \\ T_\theta(z) \\ A_\theta(z) \\ M_\theta(z) - 1 \end{pmatrix} = \nu_\theta e^{\lambda_2 z}\left[p_2 + O(e^{\lambda_2 z})\right]. \tag{2.16}$$

Note that, for $\theta \in (-\pi, 0)$, $x_{20}(\theta) < 0$. It follows from (2.12) that, for $z < 0$,

$$x_2(z,\theta) = x_{20}(\theta) e^{\int_0^z [\lambda_2 + \widetilde{N}_2(x_1(\eta,\theta), x_2(\eta,\theta))]\, d\eta} < 0,$$

which yields from (2.15) that $\nu_\theta < 0$. Hence, from the signs of the components of p_2 and (2.16) we get $S_\theta(z) - 1 < 0$, $T_\theta(z) < 0$, and $A_\theta(z) > 0$ for all sufficiently negative z. Since $S_\theta > 0$ and $M_\theta > 0$ on $(-\infty, 0]$, (2.6) in Lemma 2.2(iii) gives $A_\theta > 0$, $M_\theta > 1$, and $f(S_\theta, A_\theta, M_\theta) > 0$ on $(-\infty, 0]$. Thus, we conclude that (a) $T_\theta = 0$ occurs at most once in $(-\infty, 0]$ since at such a point $T'_\theta = \alpha f(S_\theta, A_\theta, M_\theta) > 0$; (b) if $T_\theta(0) < 0$, then $T_\theta(z) < 0$ for all $z \in (-\infty, 0]$.

Next, note that, for $\theta = 0$, $(x_{10}(0), x_{20}(0)) = (\delta, 0)$ and (2.14) yield that $S_\theta(0) > 1$ and $T_\theta(0) = \lambda_1(S_\theta(0) - 1) > 0$. Then, the continuity of S_θ and T_θ in both z and θ implies $S_\theta(z) - 1 > 0$ and $T_\theta(z) > 0$ for z sufficiently close to 0 and $\theta \in (-\pi, 0)$ sufficiently close to 0. This together with the conclusions (a) and (b) yields that, for $\theta \in (-\pi, 0)$ sufficiently close to 0, T_θ changes sign exactly once in $(-\infty, 0)$. Note that, for any $\theta \in [-\pi, 0]$, $S^2_\theta(z) + T^2_\theta(z) > 0$ for any z in its domain (for if this is not true, the uniqueness theorem would yield $S_\theta \equiv 0$ and $T_\theta \equiv 0$, contradicting that $S^2_\theta(0) + T^2_\theta(0) > 0$). Therefore, letting

$$\Theta_1 = \left\{ \theta \in (-\pi, 0) : \exists z_\theta \in \mathbb{R} \text{ such that } \left\{ \begin{array}{l} T_\theta(z_\theta) = 0, \\ T_\theta < 0 \text{ on } (-\infty, z_\theta), \\ S_\theta > 0 \text{ on } (-\infty, z_\theta] \end{array} \right. \right\},$$

we see that Θ_1 contains all $\theta \in (-\pi, 0)$ sufficiently close to 0. Furthermore, for any $\theta \in \Theta_1$, by virtue of (2.6) and $S_\theta > 0$ on $(-\infty, z_\theta]$ we have $A_\theta > 0$, $M_\theta > 1$, $f(S_\theta, A_\theta, M_\theta) > 0$ on $(-\infty, z_\theta]$, and, consequently, $T'_\theta = \alpha f(S_\theta, A_\theta, M_\theta) > 0$ at $z = z_\theta$, which together with the conclusion (a) above and the continuity of $(S_\theta, T_\theta, A_\theta, M_\theta)$ in z and θ yields that Θ_1 is open (relative to the interval $(-\pi, 0)$).

Now, we define

$$\Theta_2 := \left\{ \theta \in (-\pi, 0) : \exists \bar{z}_\theta \in \mathbb{R} \text{ such that } \left\{ \begin{array}{l} S_\theta(\bar{z}_\theta) = 0, \\ S_\theta > 0 \text{ on } (-\infty, \bar{z}_\theta), \\ T_\theta < 0 \text{ on } (-\infty, \bar{z}_\theta] \end{array} \right. \right\},$$

and claim that, if $\theta \in (-\pi, 0)$ sufficiently close to $-\pi$, then $\theta \in \Theta_2$. This is because, for $\theta = -\pi$, $(S_{-\pi}, T_{-\pi})$ lies on the line $T = b_1(S - 1)$ with $T_{-\pi} < 0$, and, as $\theta \to -\pi$, the fact that $T_\theta(0) \to T_{-\pi}(0) < 0$ and the conclusion (b) above yields that $T_\theta < 0$ on $(-\infty, 0]$; furthermore, since there is a finite $\bar{z}_{-\pi} > 0$ such that $S_{-\pi}(\bar{z}_{-\pi}) = 0$, $S_{-\pi} > 0$ on $[0, \bar{z}_{-\pi})$, and $T_{-\pi} < 0$ on $[0, \bar{z}_{-\pi}]$, the continuity of S_θ and T_θ in z and θ yields that there is a $\bar{z}_\theta$ such that $S_\theta(\bar{z}_\theta) = 0$, $S_\theta > 0$ on $[0, \bar{z}_\theta)$, and $T_\theta < 0$ on $[0, \bar{z}_\theta]$. This shows the above claim. Since $T'_\theta(\bar{z}_\theta) < 0$ for any $\theta \in \Theta_2$, the same reasoning above yields that Θ_2 is open.

Since Θ_1 and Θ_2 are open and disjointed, the connectedness of the interval $(-\pi, 0)$ yields that there is a $\theta = \theta^* \in (-\pi, 0) \setminus (\Theta_1 \cup \Theta_2)$ so that $S_{\theta^*} > 0$ and $T_{\theta^*} < 0$ on its domain $(-\infty, \omega)$ (here we again used the fact that S_{θ^*} and T_{θ^*} cannot equal zero at the same z). Then, applying Lemma 2.2(iii) yields that $(S_{\theta^*}, T_{\theta^*}, A_{\theta^*}, M_{\theta^*})$ gives a desired solution of (2.4).

Case 2. Assume that $\lambda_1 < \lambda_2$. In this case, (2.12) has a unique strongly unstable manifold $x_1 = h(x_2)$ that is tangent to the x_2-axis (cf. [18]). We straighten this manifold by introducing $\tilde{x}_1 = x_1 - h(x_2)$ and use the identity (due to the local invariance of this manifold) $\lambda_1 h(x_2) + N_1(h(x_2), x_2) = h'(x_2)[\lambda_2 x_2 + \tilde{N}_2(h(x_2), x_2)x_2]$ to write (2.12) as, in terms of the new variables $(\tilde{x}_1, x_2)$,

$$(2.17) \qquad \left\{ \begin{array}{l} \tilde{x}'_1 = \lambda_1 \tilde{x}_1 + \tilde{N}_1(\tilde{x}_1, x_2)\tilde{x}_1, \\ x'_2 = \lambda_2 x_2 + \tilde{N}_2(\tilde{x}_1 + h(x_2), x_2)x_2, \end{array} \right.$$

where

$$\begin{aligned} \tilde{N}(\tilde{x}_1, x_2)\tilde{x}_1 = {} & N_1(\tilde{x}_1 + h(x_2), x_2) - N_1(h(x_2), x_2) \\ & - h'(x_2)[\tilde{N}_2(h(x_2), x_2) - \tilde{N}_2(\tilde{x}_1 + h(x_2), x_2)]x_2. \end{aligned}$$

As in Case 1, we fix a $\delta > 0$ so small that the set $\tilde{x}_1^2 + x_2^2 \le \delta^2$ is negatively invariant for (2.17), the unstable manifold $W^u_{loc}(\delta)$ of (2.4) is negatively invariant, and $S > 0$ and $M > 0$ if $(S, T, A, M) \in W^u_{loc}(\delta)$, where $W^u_{loc}(\delta)$ is defined in the same way as in (2.13) except that each x_1 in $(x_1, x_2, h_1(x_1, x_2), h_2(x_1, x_2))^\top$ is replaced by $\tilde{x}_1 + h(x_2)$ and that $x_1^2 + x_2^2 \le \delta^2$ is replaced by $\tilde{x}_1^2 + x_2^2 \le \delta$. Hence, if $(\tilde{x}_1, x_2)$ is a solution of (2.17) with $\tilde{x}_1^2(0) + x_2^2(0) \le \delta^2$, then, for all $z \in (-\infty, 0]$, $\tilde{x}_1^2(z) + x_2^2(z) \le \delta^2$, and

$$\begin{pmatrix} S(z) - 1 \\ T(z) \\ A(z) \\ M(z) - 1 \end{pmatrix} = \tilde{x}_1(z) p_1 + x_2(z) p_2 + O(\tilde{x}_1^2(z) + x_2^2(z)) \tag{2.18}$$

is a solution of (2.4) lying on $W^u_{loc}(\delta)$ for $z \in (-\infty, 0]$. We note that, in the $(\tilde{x}_1, x_2, x_3, x_4)$-coordinate system, $\tilde{x}_1$-axis corresponds to the line $T = b_1(S - 1)$, $A = 0$, and $M = 1$, and the curve defined by $\{(\tilde{x}_1 + h(x_2), x_2, h_1(\tilde{x}_1 + h(x_2), x_2), h_2(\tilde{x}_1 + h(x_2), x_2)) : \tilde{x}_1 = 0, |x_2| \le \delta\}$ corresponds to the strongly unstable manifold of (2.4) which is tangent to p_2.

Therefore, we consider the initial values $(\tilde{x}_{10}(\theta), x_{20}(\theta))$ on the quarter of the circle C_2: $\tilde{x}_1^2 + x_2^2 = \delta^2$ lying in the second quadrant with $\tilde{x}_{10}(\theta) = \delta \cos\theta < 0$ and $x_{20}(\theta) = \delta \sin\theta > 0$ for $\theta \in [\pi/2, \pi]$; see the right figure in Figure 2. The image of C_2 under the mapping $\hat{P}$ defines the curve Γ_2 as described in the outline of the proof of Theorem 2.1, where $\hat{P}$ is the composition function defined by $\hat{P} = P^{-1} \circ \tilde{P}^{-1}$, and here $\tilde{P}$ is the nonlinear mapping $\tilde{P}(\tilde{x}_1, x_2, x_3, x_4) = (\tilde{x}_1 + h(x_2), x_2, x_3, x_4)$. It follows from (2.17) that, for $\theta \in (\pi/2, \pi)$, there is a $\mu_\theta \ne 0$ such that, as $z \to -\infty$,

$$\begin{pmatrix} \tilde{x}_1(z, \theta) \\ x_2(z, \theta) \end{pmatrix} = \mu_\theta e^{\lambda_1 z} \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} + O(e^{\lambda_1 z}) \right], \tag{2.19}$$

and, from (2.18),

$$\begin{pmatrix} S_\theta(z) - 1 \\ T_\theta(z) \\ A_\theta(z) \\ M_\theta(z) - 1 \end{pmatrix} = \mu_\theta e^{\lambda_1 z} \left[\begin{pmatrix} 1 \\ \lambda_1 \\ 0 \\ 0 \end{pmatrix} + O(e^{\lambda_1 z}) \right]. \tag{2.20}$$

Since $\theta \in (\pi/2, \pi)$, $x_{10}(\theta) = \delta \cos\theta < 0$, it follows that, for $z \in (-\infty, 0]$,

$$\tilde{x}_1(z, \theta) = x_{10}(\theta) e^{\int_0^z [\lambda_1 + \tilde{N}_1(\tilde{x}(\eta, \theta), x_2(\eta, \theta))] \, d\eta} < 0.$$

This implies from (2.19) that $\mu_\theta < 0$, and then (2.20) yields $0 < S_\theta(z) < 1$, $T_\theta(z) < 0$, and $M_\theta(z) > 0$ for all sufficiently negative z. However, since the third component of p_1 is zero, we are unable to determine the sign of $A_\theta(z)$ near $z = -\infty$ from (2.20). In order to see this sign, noting that, as $z \to -\infty$,

$$\begin{aligned} A_\theta'' + b_2 A_\theta' - \gamma\sigma A_\theta &= -\gamma \Big\{ [\varphi_1(S_\theta) - \varphi_1(1)] \varphi_2(A_\theta) \varphi_3(1) \\ &\quad + \varphi_1(1)[\varphi_2(A_\theta) - \varphi_2'(0) A_\theta] \varphi_3(M_\theta) \\ &\quad + \varphi_1(1) \varphi_2'(0) A_\theta [\varphi_3(M_\theta) - \varphi_3(1)] \Big\} \\ &= O(e^{\lambda_1 z}) A_\theta, \end{aligned}$$

we obtain by the variation of constant formula that there is a $\tilde{\mu}_\theta \neq 0$ such that, as $z \to -\infty$,

$$\begin{pmatrix} A_\theta(z) \\ A'_\theta(z) \end{pmatrix} = \tilde{\mu}_\theta e^{\lambda_2 z} \left[\begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix} + O(e^{\lambda_1 z}) \right]. \tag{2.21}$$

We claim that $\tilde{\mu}_\theta > 0$. If not, then (2.21) implies that $A_\theta(z) < 0$ for all sufficiently negative z. This together with (2.6) implies that $A_\theta < 0$, $M_\theta < 1$, and $f(S_\theta, A_\theta, M_\theta) < 0$ for all $z \in (-\infty, 0]$. (Note that, if necessary, we redefine $\varphi_3(M)$ for $M < 1/2$ such that $\varphi_3(M) = \varphi_3(1/2)$.) Then, using the first equation in (2.6) yields $T_\theta(0) - b_1[S_\theta(0) - 1] < 0$, which contradicts our choice of θ. Therefore, we must have $\tilde{\mu}_\theta > 0$ and so $A_\theta(z) > 0$ for all sufficiently negative z, as expected. Consequently, we have that $f(S_\theta, A_\theta, M_\theta) > 0$ on $(-\infty, 0]$, and we then conclude as in Case 1 that (a) $T_\theta = 0$ occurs at most once in $(-\infty, 0)$ since at such a point $T'_\theta = \alpha f(S_\theta, A_\theta, M_\theta) > 0$; (b) if $T_\theta(0) < 0$, then $T_\theta(z) < 0$ for all $z \in (-\infty, 0]$.

Next, we define the sets Θ_1 and Θ_2 in the same way as those in Case 1 except that $\theta \in (-\pi, 0)$ is replaced by $\theta \in (\pi/2, \pi)$ in each of those definitions. In a similar manner we show that Θ_1 contains all such $\theta \in (\pi/2, \pi)$ that are sufficiently close to $\pi/2$, and Θ_2 contains all such $\theta \in (\pi/2, \pi)$ that are sufficiently close to π. Consequently, the same shooting argument used in Case 1 yields that there exists at least one $\theta^{**} \in (\pi/2, \pi) \setminus (\Theta_1 \cup \Theta_2)$ such that the solution $(S_\theta, T_\theta, A_\theta, M_\theta)$ of (2.4) with $\theta = \theta^{**}$ gives a desired solution.

It remains to consider Case 3.

Case 3. We assume that either $\lambda_1 = \lambda_2$ or $\lambda_2 = 0$. As discussed in the outline of the proof of Theorem 2.1, we consider the approximating systems (2.10) with $n \geq n_0$ to which we are able to apply the results obtained in Cases 1 and 2 to obtain a sequence of solutions (S_n, T_n, A_n, M_n) of (2.10) that satisfy (2.5) and the estimates in (2.3). Since these estimates do not depend on n, using the equations in (2.10) we see that T'_n and M'_n are uniformly bounded on $\mathbb{R}$. Then, applying the Arzela–Ascoli theorem on $[-k, k]$ for each positive integer k and then using a diagonal selection process yield that there is a subsequence $(S_{n_k}, T_{n_k}, A_{n_k}, M_{n_k})$ that converges uniformly on any compact subset of $\mathbb{R}$, with its limit denoted by (S, T, A, M). Clearly, (S, T, A, M) is a solution of (2.4) on $\mathbb{R}$ and satisfies (2.3) in which the strict inequality signs "<" (resp., ">") are replaced by "$\leq$" (resp., "'$\geq$"). Consequently, since $S' = T \leq 0$ and $A' \geq 0$ on $\mathbb{R}$, the limits $(S(\pm\infty), A(\pm\infty))$ exist. Since M and $f(S, A, M)$ are bounded on $\mathbb{R}$, it follows from the last equation of (2.4) that M must satisfy the last equation in (2.6) which together with $f(S(z), A(z), M(z)) \to 0$ as $z \to \pm\infty$ yields that $M(\pm\infty) = 1$. By means of the first two equations in (2.1), we see that S'', A'' are bounded on $\mathbb{R}$ which together with $S' \leq 0$ and $A' \geq 0$ yields $T(\pm\infty) = A'(\pm\infty) = 0$. Therefore, $(S(\pm\infty), T(\pm\infty), A(\pm\infty), M(\pm\infty))$ must be equilibria of (2.4). Since all equilibria of (2.4) are given by $(S^*, 0, 0, 1)$ and $(0, 0, A^*, 1)$ for $A^* \geq 0$ and $S^* \geq 0$, it follows that $A(-\infty) = 0$, $S(\infty) = 0$, and then sending $z \to \pm\infty$ in the third equation of (2.4) yields $A(\infty) = 1$ and $S(-\infty) = 1$. Finally, we need to show (2.3). We first show that $A > 0$ on $\mathbb{R}$. For if not, since $A' \geq 0$, there is $z_0 \in \mathbb{R}$ such that $A = 0$ on $(-\infty, z_0]$ and $A > 0$ on (z_0, ∞), and then by the local uniqueness theorem it follows that $S' = b_1[S(z) - 1]$, $A \equiv 0$, and $M \equiv 1$ on $\mathbb{R}$, which is impossible. Thus $A > 0$ on $\mathbb{R}$. Next, we show that $S > 0$ on $\mathbb{R}$. For if not, since $S' \leq 0$, there is a $z_1 \in \mathbb{R}$ such that $S = 0$ on $[z_1, \infty)$ and $S > 0$ on $(-\infty, z_1)$, and again the local uniqueness theorem implies that $S \equiv 0$, $A' + b_2 A = 0$, and $M \equiv 1$ on $\mathbb{R}$, which is also impossible. This ensures $S > 0$ on $\mathbb{R}$. Hence, by virtue of $M \geq 1$, we have $f = f(S, A, M) > 0$ on $\mathbb{R}$.

Consequently, applying Lemma 2.2(iii) yields (2.7) and thereby (2.3). This completes the proof of Theorem 2.1. □

3. Conclusions. In this paper we have studied a basic bioremediation model that characterizes the essentials of a biodegradation process. The model describes the interactions between a dissolved contaminant, an injected nutrient, and single microbial species. We have established that, in a biologically active zone, the dissolved solute concentration and the advancing nutrient concentration move together as traveling fronts, while the bacteria concentration travels along as a traveling pulse. This confirms some earlier numerical observations. Compared to the existing results in the literature, the main improvement of our result lies in the application of a much broader range of kinetics which are more biologically realistic than the multiplicative Monod kinetics when modeling different kinds of pollutants. The estimates we obtained for traveling waves and their wave speeds provide qualitative and quantitative information on the concentrations of contaminant, nutrient, and bacteria as well as the removal rates of pollutants, and help identify key parameters in the model. We have employed a new dynamical systems approach in the proof of our main result, which produces a shorter and simpler proof. A future work will be investigating the uniqueness of the traveling waves in the sense that, for each wave speed, there is a unique traveling wave solution of the model.

Acknowledgment. The author thanks the referees for their valuable suggestions on this manuscript.

REFERENCES

[1] M. Beck, A. Doelman, and T. J. Kaper, *A geometric construction of traveling waves in a bioremediation model*, J. Nonlinear Sci., 16 (2006), pp. 329–349.

[2] R. Borden and P. Bedient, *Transport of dissolved hydrocarbons influenced by oxygen-limited biotransformation*, Water Resour. Res., 22 (1986), pp. 1973–1982.

[3] C. Y. Chiang, C. N. Dawson, and M. F. Wheeler, *Modeling of in situ biorestoration of organic compounds in groundwater*, Transp. Porous Media, 6 (1991), pp. 667–702.

[4] J. A. Christ, M. N. Goltz, and J. Huang, *Development and application of an analytical model to aid design and implementation of in situ remediation technologies*, J. Contaminant Hydrol., 37 (1999), pp. 295–317.

[5] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[6] B. M. Dolgonosov and T. N. Gubernatorova, *A nonlinear model of contaminant transformations in an aquatic environment*, Water Resources, 32 (2005), pp. 291–304.

[7] Y. F. Huang, G. H. Huang, G. Q. Wang, Q. G. Lin, and A. Chakma, *An integrated numerical and physical modeling system for an enhanced in situ bioremediation process*, Environ. Pollut., 144 (2006), pp. 872–885.

[8] D. Hunkeler, P. Hohener, and J. Zeyer, *Engineered and subsequent intrinsic in situ bioremediation of a diesel fuel contaminated aquifer*, J. Contaminant Hydrol., 59 (2002), pp. 231–245.

[9] H. Keijzer, M. I. J. van Dijke, and S. E. A. T. M. van der Zee, *Analytical approximation to characterize the performance of in situ aquifer bioremediation*, Adv. in Water Res., 23 (1999), pp. 217–228.

[10] H. Keijzer, R. J. Schotting, and S. E. A. T. M. van der Zee, *Semi-analytical traveling wave solution of one-dimensional aquifer bioremediation*, Comm. Appl. Nonlinear Anal., 7 (2000), pp. 1–20.

[11] A. Livingston and H. Chase, *Modeling phenol biodegradation in a fluidized-bed bioreactor*, AIChE J. 35 (1989), pp. 1980–1992.

[12] R. Murray and J. Xin, *Existence of traveling waves in a biodegradation model for organic contaminants*, SIAM J. Math. Anal., 30 (1999), pp. 72–94.

[13] J. E. Odencrantz, *Modeling the Biodegradation Kinetics of Dissolved Organics Contaminants in a Heterogeneous Two-Dimensional Aqufier*, Ph.D. dissertation, University of Illinois, Urbana, IL, 1992.

[14] J. E. Odencrantz, *Comparison of minimum-rate and multiplicative Monod biodegradation kinetic models applied to in situ bioremediation*, in Proceedings of the Fifth International Conference on Solving Groundwater Problems with Models, Dallas, TX, sponsored by the National Water Well Association and the International Groundwater Modeling Center, 1992, pp. 479–496.

[15] J. E. Odencrantz, A. J. Valocchi, and B. Rittman, *Modeling the interaction of sorption and biodegradation on transport in ground water in situ bioremediation systems*, in Proceedings of the 1993 Ground Water Modeling Conference, E. Poeter, S. Ashlock, and J. Proud, eds., Int. Ground Water Model. Cent., Golden, CO, 1993, pp. 2-5:2-12.

[16] S. Oya and A. J. Valocchi, *Characterization of traveling waves and analytical estimation of pollutant removal in one-dimensional subsurface bioremediation modeling*, Water Resour. Res., 33 (1997), pp. 1117–1127.

[17] S. Seker, H. Beyenal, B. Salih, and A. Tanyolac, *Multi-substrate growth kinetics of Pseudomonas putida for phenol removal*, Appl. Microbiol. Biotechnol., 47 (1997), pp. 610–614.

[18] L. P. Shilnikov, A. L. Shilnikov, D. V. Turaev, and L. O. Chua, *Methods of Qualitative Theory in Nonlinear Dynamics. Part* I, World Scientific, River Edge, NJ, 1998.

[19] Z. Snellinx, A. Nepovim, S. Taghavi, J. Vangronsveld, T. Vanek, and D. Van der Lelie, *Biological remediation of explosives and related nitroaromatic compounds*, Environ. Sci. Pollut. Res. Int., 9 (2002), pp. 48–61.

[20] M. Takeuchi, K. Nanba, H. Iwamoto, H. Nirei, T. Kusuda, O. Kazaoka, M. Owaki, and K. Furuya, *In situ bioremediation of a cis-dichloroethylene-contaminated aquifer utilizing methane-rich groundwater from an uncontaminated aquifer*, Water Research, 39 (2005), pp. 2438–2444.

[21] W.-T. Tang and L.-S. Fan, *Steady state phenol degradation in a draft-tube gas-liquid-solid fluidized-bed bioreactor*, AIChE J., 33 (1987), pp. 239–249.

[22] A. J. Valocchi, J. E. Odencrantz, and B. Rittman, *Computational studies of the transport of reactive solutes: Interaction between adsorption and biotransformation*, in Advances in Hydrosciences, Vol. I, Proceedings of the First International Symposium on Hydroscience and Engineering, S. Y. Wang, ed., Washington, DC, 1993, pp. 1845–1852.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 694–719

CUCKER–SMALE FLOCKING UNDER HIERARCHICAL LEADERSHIP*

JACKIE (JIANHONG) SHEN†

Abstract. A mathematical theory on flocking serves the foundation for several ubiquitous multiagent phenomena in biology, ecology, sensor networks, and economics, as well as social behavior like language emergence and evolution. Directly inspired by the recent fundamental works of Cucker and Smale on the construction and analysis of a generic flocking model, we study the emergent behavior of Cucker–Smale flocking under *hierarchical leadership*. The rates of convergence towards asymptotically coherent group patterns in different scenarios are established. The consistent convergence towards coherent patterns may well reveal the advantages and necessities of having leaders and leadership in a complex (biological, technological, economic, or social) system with sufficient intelligence.

Key words. bio-flocking, leaders, leadership, Cucker–Smale model, dynamic graphs, graph Laplacian, Fiedler number, convergence, perturbation, free will

AMS subject classifications. 92D50, 92D40, 91D30, 91C20

DOI. 10.1137/060673254

1. Introduction and motivations.

1.1. General background on flocking. Flocking, a universal phenomenon of multiagent interactions, has gained increasing interest from various research communities in biology, ecology, robotics and control theory, and sensor networks, as well as sociology and economics.

(i) (biology and ecology) The emergent behavior of bird flocks, fish schools, wolf packs, elephant herds, or bacteria aggregations, for example, has long been a major research topic in population and behavioral biology and ecology [4, 7, 8, 11, 12, 17, 23, 24].

(ii) (robotics and control) The coordination and cooperation among multiple mobile agents (robots or sensors) have been playing central roles in sensor networking, with broad applications in military, environmental control, and various field tasks [14, 25].

(iii) (economy and languages) Emergent economic behavior, such as a common belief in a price system in a complex market environment, is also intrinsically connected to flocking. The emergence of a common language in primitive societies is yet another example of a coherent collective behavior emerging within a complex system [8, 9].

The present work can largely be categorized into the biology realm, and has been directly inspired by the recent mathematical works of Cucker and Smale [7, 8], as the title suggests. Mathematical abstraction and rigorous analysis are more the focus herein than actual biological or physical realizability or feasibility. As in physics, the study of idealized models can often shed light on various observed patterns in the real world, *if* such models can indeed capture the very essence.

*Received by the editors October 24, 2006; accepted for publication (in revised form) September 21, 2007; published electronically December 19, 2007. This work was partially supported by (USA) NSF grant DMS-0604510 (2006–2009).

http://www.siam.org/journals/siap/68-3/67325.html

†Wall Street, New York, NY 10166 (jackieneoshen@gmail.com).

In biology and physics, the main goal of flocking study is to be able to interpret, model, analyze, predict, and simulate various flocking or multiagent aggregating behavior. Most works have been focusing on modeling and simulation [15, 27]. See, for example, the several important models investigated by Flierl et al. [12] (and their stochastic formulation). The more recent paper of Parrish, Viscido, and Grünbaum [18] also provides a comprehensive comparison among some major existing models and their governing variables (in the context of fish schooling). Quantitative analysis (as in [7, 8, 14]) on the asymptotic rates of emergence and convergence, on the other hand, has been relatively rare.

Mathematical efforts are gradually gaining strength in this multidisciplinary area. In the continuum limit, for example, there have been several recent efforts made by Bertozzi's group [23, 24], in which global swarming (i.e., with densely populated agents) patterns are modeled and analyzed via suitable spatiotemporal differential equations. Discrete-to-continuum limits of interacting particle systems have also been investigated by the same group [2, 11] recently. Consistent and generic mathematical analysis has been very much in an early stage for many biological aggregation phenomena. In the current paper, following the recent remarkable works of Cucker and Smale [7, 8] on flocking analysis, we attempt to further extend such research.

1.2. Cucker–Smale flocking model. Given a flock of k agents (birds, fish, wolves, etc.) labeled $i = 1, 2, \dots, k$, the Cucker–Smale flocking model is specified by the *nonlinear* autonomous dynamic system:

$$\begin{cases} \dot{x}_i(t) = v_i, \\ \dot{v}_i(t) = \sum_{j \in \mathcal{L}(i)} a_{ij}(x)(v_j - v_i), \qquad i = 1:k,\ t > 0, \end{cases} \tag{1.1}$$

where $x_i(t)$ and $v_i(t)$ are 3D (three-dimensional, which is nonessential) position and velocity vectors at time t, $x = (x_1, \dots, x_k) \in (\mathbb{R}^3)^k$, and $\mathcal{L}(i) \subseteq \{1, \dots, k\}$ denotes the subgroup of agents that directly influence agent i. Furthermore, the *connectivity coefficients* $a_{ij}(x)$ take the form

$$a_{ij}(x) = w(|x_i - x_j|^2) \qquad \text{for some nonnegative weight profile } w(y).$$

In the current paper, by *Cucker–Smale flocking model*, we require as in [7, 8] that the interaction weight function $w(y)$ take the form

$$w(y) = \frac{H}{(1+y)^\beta} \qquad \text{or} \qquad w(y) \geq \frac{H}{(1+y)^\beta}, \tag{1.2}$$

where H and β are two positive system parameters. One shall see that the two ($=$ vs. $\geq$) make no difference in the analysis hereafter as long as $w(y)$ is bounded and sufficiently smooth (also see [7]). We also must point out that this model has been put in a more general and abstract setting in the subsequent work of Cucker and Smale [8].

The look of system (1.1) is not entirely new. For example, the 2D model studied by Vicsek et al. [27] is very similar in that v_i's share the same magnitude (or speed), while their heading directions θ_i's satisfy a similar set of equations.

It is the particular choice of the connectivity coefficients in (1.2) that has made the Cucker–Smale model mathematically more attractive. Vicsek et al.'s model (in discrete time) [27] can be considered as taking the following cut-off weight function:

$$w(y) = w_r(y) = 1_{y \leq r^2}(y), \qquad \mathcal{L}(i) \equiv \{1, \dots, k\} \quad \forall i.$$

That is, two distinct agents, x_i and x_j, interact if and only if they are within a distance of $r > 0$, which is assigned a priori and fixed throughout. Moreover, the nonzero weights are uniformly 1's. The lack of long-range interactions has made the model very difficult to analyze. For example, the remarkable efforts of Jadbabaie, Lin, and Morse [14] on emergence analysis avoided the actual dynamic dependence of a_{ij} on the configuration x. Instead, they focused on an altered setting that involves switching controls. The convergence results obtained for this approach, however, rely on the infinite time-sequence of states of the system.

The main results of Cucker and Smale [7] can be summarized as follows: when $\beta < 1/2$, the flock converge to some translating rigid structure (moving at a constant velocity) *unconditionally*, i.e., regardless the initial configuration; and when $\beta \geq 1/2$, the initial velocities and positions have to satisfy certain compatible conditions so that the entire flock can converge asymptotically.

In summary, in the modeling and analysis of Cucker and Smale [7, 8], not only are the conditions for pattern emergence easily verifiable (i.e., by checking the initial conditions), but the role of long-range interaction is also clearly quantified. A smaller β signifies more intense long-range interactions among agents, while a bigger β leads to much weaker ones. It has been shown that the critical exponent $\beta_c = 1/2$ is sharp and necessary. Previously, the connection between global pattern emergence and individual action rules has often only been observed experimentally or addressed empirically. (Vicsek et al. [27], for example, experimentally observed phase transition induced by population density ρ and random fluctuation η. A higher density corresponds to more interaction among agents, or loosely, smaller β in the Cucker–Smale model.)

1.3. Motivations and main results of current work. In the current work, we investigate the emergent behavior of Cucker–Smale flocking under hierarchical leadership (HL), which will be defined in detail in the next section.

Roughly, an HL flock is one whose members can be ordered in such a way that lower-rank agents are led and only led by some agents of higher ranks. As explained in more detail in section 2, for HL flocks, it is often either nontrivial or impossible to define a "fixed" inner product so that the Fiedler number of the associated (graph) Laplacian can be exploited, which is the key to the original work of Cucker and Smale [7] and its subsequent generalization [8]. The current work thus takes a somewhat different approach in order to fully benefit from the characteristic structures of HL.

As far as applications are concerned, there are two types of HL: passive and active.

(A) (passive/transient leadership)

(A.1) (disturbed bird flocks) In nature, certain types of leadership emerge in a transient and dynamic fashion and are often prompted by a specific environment. For a disturbed bird flock at rest, for example, the bird that first senses the approach of an unexpected pedestrian or predator often takes flight first, warns others, and first gains full speed, and consequently flies ahead of the entire flock and serves as a virtual leader.

(A.2) (driving in a traffic) During rush hours, each individual driver mainly maneuvers according to the moving patterns of several cars right ahead in the visual field. Thus a chain of leadership naturally arises and extends linearly along the traffic. The leadership here is also prompted by the environment rather than being intrinsic among the stranger drivers.

(B) (active/intrinsic leadership)

(B.1) (governmental/military hierarchies) Such hierarchical leadership is inherent in various social groups or structures, and often leads to more efficient management. Examples include the chain of President–Governor–Mayor in the governmental system, and the military chain of command from the Commander-in-Chief all the way down to the soldiers.

(B.2) (social animals) For some social animals such as monkeys, wolves, or elephants [4], the group or social status of each member is clearly recognized by others, is stably maintained, and guides the action of each individual in the hierarchies. (See also the recent work of Couzin et al. [4] for nonhierarchical but "effective" leadership.)

Our main results are the three theorems summarized below. All HL flocks are assumed to have Cucker–Smale connectivity introduced in the preceding subsection.

(i) (section 3) For an HL $(k+1)$-flock marching at a sufficiently small *discrete* time step h, under a similar classification scheme according to whether $\beta < \beta_c$, $= \beta_c$, or $> \beta_c$, as in Cucker and Smale [7, 8], the velocities of the flock converge at a rate of $O(\rho_h^n n^{k-1})$, where the factor $\rho_h \in (0,1)$ depends only on h, system parameters, and the initial configuration of the flock. The critical exponent is given by $\beta_c = 1/(2k)$, instead of $\beta_c = 1/2$ in the original work of Cucker and Smale [7]. (For a 2-flock (with $k = 1$) they are the same. For $k > 1$, the β_c herein could be overrestrictive and due to the deficiency of the particular methodology adopted.)

(ii) (section 4) For an HL flock under continuous-time dynamics, when $\beta < 1/2$, there exists some $B > 0$, such that the velocities of the flock converge at an exponential rate of $O(e^{-Bt})$. The constant B depends only on the system parameters and the initial configuration of the flock. (From the simple calculation on an HL 2-flock, $\beta_c = 1/2$ is sharp in order to achieve *unconditional* convergence.)

(iii) (section 5) For an HL $(k+1)$-flock $[0, 1, \ldots, k]$ of which the overall leader agent 0 takes a free-will acceleration $\dot{v}_0 = f(t)$ (thus the system is no longer autonomous), as long as the overall leader behaves moderately so that $f(t) = O((1+t)^{-\mu})$ for some $\mu > k$, the velocities of the flock will still converge at a rate of $O((1+t)^{-(\mu-k)})$ when $\beta < 1/2$. (By (ii) where $f \equiv 0$, $\beta_c = 1/2$ is again sharp for unconditional convergence.)

We also mention that Jadbabaie, Lin, and Morse [14] also studied (under discrete time and working with Vicsek et al.'s orientation model [27]) the effect of a *single* leader moving at a *fixed constant* velocity. As mentioned above, due to the difficulty in dealing with configuration-dependent dynamics, the authors switched to the study of an altered *control* problem (under the assumption of intermittent joint connectivity).

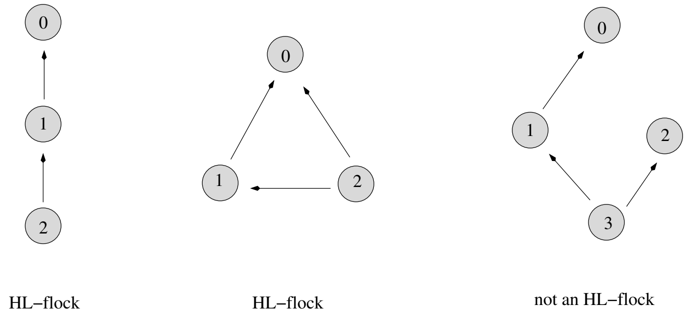
In addition to the three main sections mentioned above, definitions and further detailed background will be introduced in section 2. The conclusion is drawn in section 6.

2. HL flocks and definability of compatible inner products.

2.1. Flocks under hierarchical leadership (HL flocks).

DEFINITION 2.1 (an HL flock). *A $(k+1)$-flock is said to be under* hierarchical leadership *if the agents (birds, fish, wolves, etc.) can be labeled* $[0, 1, \ldots, k]$, *such that*

(i) $a_{ij} = a_{\text{agent } i \text{ led by } j} \neq 0$ *implies that* $j < i$; *and*



HL−flock HL−flock not an HL−flock

Fig. 2.1. *Two examples of HL flocks and one example of a non-HL flock. The arrow $i \to j$ means that agent i is led by agent j, or, equivalently, $a_{ij} > 0$. Visually, it means that i looks up to j.*

(ii) *if we define the* leader set *of each agent i by*

$$\mathcal{L}(i) = \{j \mid a_{ij} > 0\},$$

then for any $i > 0$, $\mathcal{L}(i) \neq \varnothing$ (nonempty).
If so, the flock is called an HL flock.

Notice that the second condition requires that, except for agent 0, all the others must be subject to some leadership. On the other hand, the first condition implies that $\mathcal{L}(0) = \varnothing$. Thus agent 0 is the overall leader (*direct or indirect*) for the entire flock. Figure 2.1 depicts the connectivity structure of two HL flocks and one non-HL flock.

Proposition 2.2 (connectivity matrix of an HL flock). *A $(k+1)$-flock is an HL flock if and only if after some ordered labeling $[0, 1, \dots, k]$ the connectivity matrix $K = (a_{ij})_{0 \le i,j \le k}$ is lower triangular and, for any row $i > 0$, there exists at least one positive off-diagonal element a_{ij}.*

Subject to convenience, in what follows a generic HL flock shall be denoted by either $[0, 1, \dots, k]$ or $[1, \dots, k]$. Following a setting similar to Cucker and Smale [7] or Chung [3], define the graph Laplacian matrix by

$$L = D - K, \qquad D = \operatorname{diag}(d_0, \dots, d_k), \qquad d_i = \sum_j a_{ij}. \tag{2.1}$$

Similarly, define the two (nonorthogonally) complementing subspaces of $\mathbb{R}^{k+1}$:

$$\Delta = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{k+1} \right\} \quad \text{and} \quad \mathbb{R}^k = \left\{ \begin{pmatrix} 0 \\ x_1 \\ \vdots \\ x_k \end{pmatrix} \mid x_i's \in \mathbb{R} \right\}.$$

Then it is easy to see that

$$\Delta = \operatorname{Ker}(L), \qquad \mathbb{R}^k = \operatorname{Range}(L) \text{ is } L\text{-invariant}.$$

Notice that the kernel assertion is directly guaranteed by the second condition of an HL flock, without which the kernel could be larger.

From now on, as in Cucker and Smale [7, 8], we shall consider only the restriction of the Laplacian on the reduced space $\mathbb{R}^k$. Then it becomes nonsingular and shall still be denoted by L for convenience. *We also must point out that when applied to actual flocking, the reduced Laplacian L is applied to $\mathbb{R}^{3k}$ (instead of $\mathbb{R}^k$) via the three spatial dimensions individually.*

2.2. Definability of compatible inner products. The general framework of Cucker and Smale [7] relies upon the Fiedler number of the Laplacian operator L, i.e., the smallest positive eigenvalue in the reduced space. In particular, it assumes the existence of a fixed inner product $\langle \cdot, \cdot \rangle$ such that

$$\langle Lv, v \rangle \geq \xi \langle v, v \rangle \qquad \text{for any } v \in \mathbb{R}^k. \tag{2.2}$$

Then an a priori lower bound on $\xi = \xi(x)$ constitutes the core to the convergence results established by Cucker and Smale [7, 8]. Below we show, however, that such inner products could fail to exist for nonsymmetric systems like HL flocks.

THEOREM 2.3. *Consider the special HL $(k+1)$-flock $[0, 1, \ldots, k]$ such that $\mathcal{L}(i) = \{i-1\}$ for $i > 0$, and an instant when $a_{i,i-1} \equiv a$ for some fixed $a > 0$ and any $i > 0$. Then the smallest eigenvalue is $\xi = a$, but there exists no inner product $\langle \cdot, \cdot \rangle$ in the reduced space $\mathbb{R}^k$ such that*

$$\langle Lv, v \rangle \geq a \langle v, v \rangle, \qquad v \in \mathbb{R}^k.$$

Proof. It is easy to see that the (reduced) Laplacian L is given by

$$L = L_a = \begin{bmatrix} a & 0 & \dots & 0 & 0 \\ -a & a & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -a & a \end{bmatrix}_{k \times k}.$$

In particular, $L_a = aL_1$, and it suffices to prove the case when $a = 1$. If such an inner product did exist, one would have

$$\langle L_1 v, v \rangle \geq \langle v, v \rangle \quad \text{or} \quad \langle Jv, v \rangle \geq 0,$$

where $J = L_1 - Id$. Notice that $Jv = (0, -z_1, \ldots, -z_{k-1})^T$ for $v = (z_1, \ldots, z_k)^T$.

Let $\{e_1, \ldots, e_k\}$ denote the canonical basis of $\mathbb{R}^k$, and define

$$G = (g_{ij}) = (\langle e_i, e_j \rangle)_{k \times k}$$

to be the associated Grammian matrix of the inner product. Then G must be positive definite. For any $v = (z_1, \ldots, z_k)^T$, one has

$$\langle v, Jv \rangle = v^T G \cdot Jv = (z_1, \ldots, z_k)(g_{ij})(0, -z_1, \ldots, -z_{k-1})^T.$$

Consider a special vector of the form $w = w_t = (0, \ldots, 0, 1, t)^T \in \mathbb{R}^k$. Then

$$\langle w_t, Jw_t \rangle = (0, \ldots, 0, 1, t)(g_{ij})(0, \ldots, 0, 1)^T = g_{k-1,k} + g_{k,k} t.$$

Notice that $g_{k,k} = \langle e_k, e_k \rangle > 0$. Then for any

$$t < -\frac{|g_{k-1,k}|}{g_{k,k}},$$

one must have $\langle w_t, Jw_t \rangle < 0$, which is contradictory. ∎

Even when such compatible inner products do exist, for a general nonsymmetric flock, they often depend on the configuration of the flock, and are thus time-dependent. This causes much inconvenience or a potential impasse for the Cucker–Smale approach in [7, 8]. The efforts in the current work follow a different approach by exploiting the specific structures of HL flocks.

3. Discrete-time emergence. Recall that in continuous time, the Cucker–Smale flocking model is given by

$$\begin{cases} \dot{x} = v, \\ \dot{v} = -L_x v, \qquad t > 0, \end{cases} \tag{3.1}$$

where the reduced Laplacian $L = L_x$ is defined as in (2.1) and both x and v are considered in the reduced (quotient) space. For a $(k+1)$-flock, both of them belong to $\mathbb{R}^{3k}$.

Fix a discrete time step $h > 0$. Define

$$x[n] = x(nh), \qquad v[n] = v(nh), \quad \text{and} \quad L_n = L_{x[n]}.$$

(Note: the parenthesis-bracket correspondence follows the convention in digital signal processing [22].) Then the continuous-time system (3.1) is discretized to

$$\begin{cases} x[n+1] = x[n] + hv[n], \\ v[n+1] = S[n]v[n], \qquad n = 0, 1, \ldots, \end{cases} \tag{3.2}$$

where $S[n] = S^h[n] = Id - hL_n$.

For an HL $(k+1)$-flock $[0, 1, \ldots, k]$, recall that the reduced Laplacian is given by

$$L_n = \begin{bmatrix} d_1[n] & 0 & \ldots & 0 & 0 \\ -a_{21}[n] & d_2[n] & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -a_{k1}[n] & -a_{k2}[n] & \ldots & -a_{k,k-1}[n] & d_k[n] \end{bmatrix}_{k \times k}. \tag{3.3}$$

For $i > 0$, since the leader set $\mathcal{L}(i) \neq \varnothing$, we have

$$d_i[n] = \sum_{j=1}^{k} a_{ij}[n] = \sum_{j \in \mathcal{L}[i]} a_{ij}[n] > 0. \tag{3.4}$$

Under the Cucker–Smale model, one has for any $j \in \mathcal{L}(i)$

$$a_{ij}[n] = \frac{H}{\left(1 + |\tilde{x}_j[n] - \tilde{x}_i[n]|^2 / 2\right)^{\beta}}, \tag{3.5}$$

where $\tilde{x}_i$ denotes the original 3D position vector of agent i (and the factor $1/2$ is for convenience). In the reduced quotient space, one has $x_i = \tilde{x}_i - \tilde{x}_0 \in \mathbb{R}^3$ since the original configuration vector $\tilde{x} \in \mathbb{R}^{3(k+1)}$ and the reduced representation $x \in \mathbb{R}^{3k}$ are connected via

$$\tilde{x} = \begin{bmatrix} \tilde{x}_0 \\ \tilde{x}_1 \\ \vdots \\ \tilde{x}_k \end{bmatrix} = \begin{bmatrix} \tilde{x}_0 \\ \tilde{x}_0 \\ \vdots \\ \tilde{x}_0 \end{bmatrix} + \begin{bmatrix} 0 \\ \tilde{x}_1 - \tilde{x}_0 \\ \vdots \\ \tilde{x}_k - \tilde{x}_0 \end{bmatrix} = \begin{bmatrix} \tilde{x}_0 \\ \tilde{x}_0 \\ \vdots \\ \tilde{x}_0 \end{bmatrix} + \begin{bmatrix} 0 \\ x \end{bmatrix}.$$

As a result, for any pair $i, j > 0$,

$$|\tilde{x}_i - \tilde{x}_j|^2 = |x_i - x_j|^2 \leq 2(|x_i|^2 + |x_j|^2) \leq 2|x|^2.$$

In combination with (3.4) and (3.5), this implies that under the Cucker–Smale connectivity,

$$d_i[n] \geq \frac{H}{(1+|x[n]|^2)^{\beta}}, \qquad i > 0. \tag{3.6}$$

Assume, as in Cucker and Smale [7], that under suitable initial conditions (according to whether $\beta <, =$, or $> \beta_c = 1/(2k)$), one has the uniform bound on the reduced position vector:

$$|x[n]|^2 \leq B_h \qquad \text{for } n = 0, 1, \ldots, \tag{3.7}$$

where B_h is a constant bound depending only on h, the system parameters H and β, as well as the initial configuration. (The existence of B_h is a crucial ingredient of the proof and will be further addressed immediately after this main line.) Then one has, for any $n \geq 0$ and $i > 0$,

$$d_i[n] \geq d_* = \frac{H}{(1+B_h)^{\beta}}. \tag{3.8}$$

Proposition 3.1 (uniform elementwise bound on S). *For $0 < h < \frac{1}{2kH}$, $S_{ij}[n] \geq 0$ for any i, j, and*

$$\max_{i,j} S_{ij}[n] \leq 1 - hd_* := \rho_h, \qquad n = 0, 1, \ldots. \tag{3.9}$$

Proof. By definition,

$$S[n] = Id - hL_n \begin{bmatrix} 1-hd_1[n] & 0 & \ldots & 0 & 0 \\ ha_{21}[n] & 1-hd_2[n] & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ ha_{k1}[n] & ha_{k2}[n] & \ldots & ha_{k,k-1}[n] & 1-hd_k[n] \end{bmatrix}_{k\times k}.$$

Under the condition on h, for the off-diagonals $i > j$, we have

$$S_{ij}[n] = ha_{ij} \leq hH < \frac{1}{2k} \leq \frac{1}{2}.$$

For the diagonals, since $a_{ij} \leq H$, we have $d_i \leq (k-1)H$, and

$$S_{ii}[n] = 1 - hd_i \geq 1 - h(k-1)H > 1 - \frac{1}{2} = \frac{1}{2}.$$

Therefore,

$$\max_{ij} S_{ij}[n] = \max_i S_{ii}[n] = 1 - h \min_i d_i \leq 1 - hd_*,$$

which completes the proof. □

Next, our goal is to be able to control the growth rate of the matrix iteration:

$$S[n]S[n-1]\cdots S[0] \qquad \text{as } n \to \infty.$$

Normally, such asymptotic behavior is investigated via the so-called joint spectral radius (e.g., Rota and Strang [19], Daubechies and Lagarias [10], or Shen [20, 21]),

$$\lim_{n\to\infty} \|S[n]S[n-1]\cdots S[0]\|^{\frac{1}{n}},$$

which is often too complex to be feasible since the matrices evolve and generally do not commute. The approach below resembles the Lebesgue dominant convergence theorem in analysis [16].

Definition 3.2 (domination). *A matrix $B = (b_{ij})$ is said to be dominated by another matrix $C = (c_{ij})$ of equal dimensions if*

$$|b_{ij}| \leq c_{ij} \qquad \text{for any } i, j.$$

If so, we write $B \prec C$.

Proposition 3.3. *There exists some constant α, such that whenever $B \prec C$,*

$$\|B\| \leq \alpha \|C\|.$$

Here α depends only on the type of matrix norm adopted.

Proof. All norms in a finite-dimensional Banach space are equivalent. Therefore, it suffices to establish the inequality under any special matrix norm. Consider the Fröbenius norm:

$$\|B\|^2 = \text{trace}(BB^T) = \sum_{i,j} b_{ij}^2 \leq \sum_{i,j} c_{ij}^2 = \text{trace}(CC^T) = \|C\|^2,$$

with $\alpha = 1$ (the superscript T here denotes the transpose). The general constant α resurfaces when another norm is used instead. □

Proposition 3.4. *Suppose $B_i \prec C_i$ for $i = 0, \ldots, n$. Then*

$$B_n B_{n-1} \cdots B_0 \prec C_n C_{n-1} \cdots C_0.$$

The proof is trivial. Next we define a "complete" lower triangular matrix $T = (t_{ij})_{k\times k}$ by

$$t_{ij} = \begin{cases} 1, & i \geq j; \\ 0 & \text{otherwise.} \end{cases}$$

Then the elementwise bound established in Proposition 3.1 directly implies the following.

Corollary 3.5. *Let $\rho_h = 1 - hd_*$ as in Proposition* 3.1*. Then*

$$S[n] \prec \rho_h T, \quad \text{and} \quad S[n-1] \cdots S[0] \prec \rho_h^n T^n, \qquad n = 0, 1, \ldots.$$

Lemma 3.6. *Let $T = (t_{ij})_{k\times k}$ be defined as above. Then $\|T^n\| = O(n^{k-1})$.*

Proof. Denote by J the $k \times k$ lower triangular matrix whose nonzero elements are all 1's and are *only* distributed right below the diagonal, e.g., the 3×3 case,

$$J = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Then it is easy to see that

$$T = I + J + \cdots + J^{k-1}.$$

Since $J^k = J^{k+1} = \cdots = 0_{k\times k}$, one can also write

$$T = \sum_{m=0}^{\infty} J^m.$$

More generally, for any t with $|t| < 1$, one can define

$$T(t) = \sum_{m=0}^{\infty} t^m J^m = (I - tJ)^{-1}.$$

Then

$$T(t)^n = (I - tJ)^{-n} = \sum_{m=0}^{\infty} \binom{-n}{m} (-t)^m J^m = \sum_{m=0}^{k-1} \binom{n+m-1}{m} t^m J^m.$$

Letting $t \to 1$, we have

$$T^n = \lim_{t \to 1} T(t)^n = \sum_{m=0}^{k-1} \binom{n+m-1}{m} J^m \prec O(n^{k-1})T.$$

The proof is then complete via Proposition 3.3. □

Combining all the preceding results in this section, we have arrived at the following conclusion.

THEOREM 3.7. *In the discrete-time Cucker–Smale model* (3.2) *for an HL* $(k+1)$*-flock, for any sufficiently small marching step h (as in* (3.7)*, Proposition* 3.1*, and Cucker and Smale* [7, 8]*), there exists some* $\rho_h \in (0, 1)$ *under the conditions similar to* [7, 8] *based upon whether* $\beta <, =,$ *or* $> \beta_c = 1/(2k)$*, such that*

$$S[n] \cdots S[0] \prec O(\rho_h^n n^{k-1})T.$$

In particular, one has

$$|v[n]| \leq O(\rho_h^n n^{k-1})|v[0]|, \qquad n \to \infty.$$

The order constant in $O(\cdot)$ *depends only on the size k of the flock.*

We point out that the polynomial growth rate $O(n^{k-1})$ (coming from T^n in Lemma 3.6) is characteristic of triangular HL flocks. A "full" system would make the approach here infeasible since

$$\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{k \times k}^n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdots\cdots \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} = k^{n-1} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}.$$

The exponential growth rate k^n would thus overpower ρ_h^n and lead to an exponential blowup.

Finally, we further address the important issue raised earlier in the proof concerning the boundedness condition in (3.7): $|x[n]|^2 \leq B_h$ for all n. The existence of the convergence factor ρ_h has crucially depended on such a bound B_h. On the other hand, the very existence of B_h, as we intend to show now, depends on ρ_h. This *entanglement* is characteristic of the nonlinear Cucker–Smale flocking model (as well as in Vicsek et al. [27] and Jadbabaie, Lin, and Morse [14]) and makes this type of model difficult to analyze. In the rest of the section, we introduce the brilliant approach of Cucker and Smale in unraveling such entanglement, which then genuinely completes the proof.

Lemma 3.8. *For any given integer $k \geq 1$ and $t \in [0, 1)$, one has*

$$\sum_{m=0}^{\infty} t^m m^{k-1} \leq (k-1)!(1-t)^{-k}. \tag{3.10}$$

Proof. Notice that the equality holds when $k = 1$. Generally, for any $t \in [0, 1)$,

$$\begin{aligned}(k-1)!(1-t)^{-k} &= (k-1)! \sum_{m=0}^{\infty} \binom{-k}{m} (-t)^m \\ &= (k-1)! \sum_{m=0}^{\infty} \binom{m+k-1}{k-1} t^m \\ &= \sum_{m=0}^{\infty} (m+k-1) \cdots (m+1) t^m \\ &\geq \sum_{m=0}^{\infty} m^{k-1} t^m,\end{aligned}$$

which completes the proof. □

We now apply the self-bounding technique developed by Cucker and Smale in [7, 8] to establish the bound $\left|x[n]\right|^2 \leq B_h$ that is crucially needed in the proof of Theorem 3.7. It also explains the origin of the critical exponent $\beta_c = 1/(2k)$ and its role.

We thus return to the step in (3.7). This time, instead of assuming a priori that $\left|x[n]\right|^2 \leq B_h$ for *all* $n \geq 0$, we proceed as follows. Fix any discrete time mark N, and define

$$|x|_* = \max_{0 \leq n \leq N} |x|[n], \qquad N_* \in \operatorname{argmax}_{0 \leq n \leq N} |x|[n], \tag{3.11}$$

and similarly define

$$d_* = \frac{H}{(1 + |x|_*^2)^{\beta}}. \tag{3.12}$$

Thus $|x|_*$ could be considered as a "localized" version of B_h, restricted in any designated *finite* time segment $[0, N]$.

Then all the earlier analysis and results hold up to the bounding formula on $|v|[n]$ in Theorem 3.7, as long as one restricts n to be within $[0, N]$. In particular for $\rho_h = 1 - hd_*$,

$$|v|[n] \leq A \rho_h^n n^{k-1}, \qquad n = 0, \dots, N,$$

where the constant A depends only on k but on neither n nor N.

Therefore, by the first equation of HL flocking in (3.2), for any $n \in [0, N]$,

$$\begin{aligned}|x|[n] &\leq |x|[0] + \sum_{m=0}^{n-1} |x[m+1] - x[m]| = |x|[0] + h \sum_{m=0}^{n-1} |v[m]| \\ &\leq |x|[0] + Ah \sum_{m=0}^{n-1} \rho_h^m m^{k-1} \leq |x|[0] + Ah \sum_{m=0}^{\infty} \rho_h^m m^{k-1} \\ &\leq |x|[0] + (k-1)! Ah (1 - \rho_h)^{-k}.\end{aligned}$$

In particular, for $n = N_*$,

$$|x|_* = |x|[N_*] \le |x|[0] + (k-1)!Ah(1-\rho_h)^{-k}.$$

Now that

$$(1-\rho_h)^{-k} = h^{-k}d_*^{-k} = (hH)^{-k}(1+|x|_*^2)^{\beta k},$$

one has the Cucker–Smale type of self-bounding inequality for the unknown $|x|_*$:

$$|x|_* \le |x|[0] + (k-1)!Ah(hH)^{-k}(1+|x|_*^2)^{\beta k}.$$

Define $Z = (1+|x|_*^2)^{1/2}$. Then

$$Z \le 1 + |x|_* \le c + bZ^{2\beta k}, \tag{3.13}$$

with $c = 1 + |x|[0]$ and $b = (k-1)!Ah(hH)^{-k}$.

The rest of the analysis then goes exactly as in Cucker and Smale [7, 8]. Define

$$F(z) = z - bz^s - c, \qquad \text{with } s = 2\beta k \quad \text{and} \quad z > 0.$$

Then when $s < 1$, the nonlinear function $F(z)$ has a unique zero z_* after which F stays positive. Since $F(Z) \le 0$, one thus must have $Z \le z_*$, or

$$|x|_* \le Z \le z_*.$$

Now that z_* depends only on c and b, which are independent of the preassigned time mark N, we have obtained the uniform bound

$$|x|[N] \le |x|[N_*] = |x|_* \le z_* \qquad \forall\ N = 0, 1, \ldots.$$

Thus $B_h = z_*^2$ is the uniform bound needed in the proof of Theorem 3.7. This is the case when $\beta \le \beta_c = 1/(2k)$.

The other two cases, when $\beta = \beta_c$ and $\beta > \beta_c$ (corresponding to $s = 1$ and $s > 1$ for $F(z)$), can be analyzed exactly in the same manner as in Cucker and Smale [7, 8], and will be omitted here. In particular, in both cases, there will be *sufficient*-type conditions on the initial configurations in order for the bound B_h to exist. In the third case, $\beta > \beta_c$, there will also be a more stringent upper bound on the time marching size h. We refer the reader to Cucker and Smale for the detailed analysis on $F(z)$ in these two cases. This completes the proof of Theorem 3.7.

In the next section, we investigate the emergent behavior of the continuous-time HL flocking using quite different methods. There, the results hint that the unconditional convergence range $\beta \in [0, 1/(2k))$ just established might still be extendable onto $[0, 1/2)$, as in Cucker and Smale [7]. Thus the critical exponent $\beta_c = 1/(2k)$ might be further improved if other alternative approaches are to be investigated in the future.

4. Continuous-time emergence.

Let $[1, \ldots, k]$ be an HL k-flock in that order, connected via the Cucker–Smale strength with parameters β and H as in (1.2). In this section, we establish the emergence behavior for the entire flock when $\beta < 1/2$, via the methods of induction and perturbation. The associated intuition is as follows. If the subflock $[1, \ldots, i-1]$ almost reaches convergence, it shall look like a rigid one-body to agent i. Then $[1, \ldots, i-1, i]$ is not far from a simpler two-agent flock. Our goal is to develop rigorous mathematical analysis to quantify and support this perspective. (In this section, we shall work with $[1, \ldots, k]$ instead of $[0, 1, \ldots, k]$ due to the lack of advantage of introducing index 0.)

4.1. The property of positivity. The general properties to be established in this subsection are characteristic of the Cucker–Smale flocking model. They could be useful for any future works on the model, on top of their roles in the proof of the main results of this section.

Let $x_i, v_i \in \mathbb{R}^3$ denote the 3D position and velocity vectors of agent i. Recall that the Cucker–Smale flocking model is given by

$$\begin{cases} \dot{x}_i = v_i, \\ \dot{v}_i = -(L_x v)_i = \sum_{j \in \mathcal{L}(i)} a_{ij}(x)(v_j - v_i) \end{cases} \tag{4.1}$$

for $t > 0$, $i = 1, \ldots, k$, and $x = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^{3k}$. The Cucker–Smale connectivity strength is specified by

$$a_{ij}(x) = \frac{H}{(1 + |x_j - x_i|^2)^\beta}, \qquad j \in \mathcal{L}(i).$$

(As mentioned earlier in the introduction, changing "=" to "≥" does not affect the subsequent analysis as long as the $a_{ij}(x)$'s are bounded and sufficiently smooth.) Given a solution $(x(t), v(t))$ to the continuous Cucker–Smale model (4.1), we write for convenience

$$a_{ij}(t) = a_{ij}(x(t)) \quad \text{and} \quad L_t = L_{x(t)}.$$

Let $\eta = (\eta_1, \eta_2, \ldots, \eta_k)^T \in \mathbb{R}^k$ be k scalars, and consider the following system of ordinary differential equations:

$$\dot{\eta} = -L_t \eta, \qquad t > 0, \qquad \text{given } \eta^0 = \eta\big|_{t=0}. \tag{4.2}$$

Componentwise, we have

$$\dot{\eta}_i = \sum_{j \in \mathcal{L}(i)} a_{ij}(t)(\eta_j - \eta_i), \qquad i = 1, \ldots, k. \tag{4.3}$$

THEOREM 4.1 (positivity). *Suppose $\eta_i^0 \geq 0$ for $i = 1, \ldots, k$. Then for all $t > 0$ and i, $\eta_i(t) \geq 0$.*

Proof. For any agent i in the flock, define

$$\begin{aligned} &\mathcal{L}^0(i) = \{i\}, \\ &\mathcal{L}^m(i) = \mathcal{L}(\mathcal{L}^{m-1}(i)), \qquad \text{all } m\text{th level leaders of } i, \quad \text{and} \\ &[\mathcal{L}](i) = \mathcal{L}^0(i) \cup \mathcal{L}^1(i) \cup \mathcal{L}^2(i) \cdots, \qquad \text{all leaders of } i\text{, direct or indirect.} \end{aligned} \tag{4.4}$$

Then it is easy to see that system (4.3) applied to $[\mathcal{L}](i)$ is always self-contained, i.e., $(\eta_j \mid j \in [\mathcal{L}](i))$ is not influenced by any variables in $(\eta_j \mid j \notin [\mathcal{L}](i))$ (but certainly not vice versa).

For convenience, we shall call the restriction of system (4.2) or (4.3) on the subflock $[\mathcal{L}](i)$ the $[\mathcal{L}](i)$-*system.* Then it suffices to establish the theorem for each $[\mathcal{L}](i)$ system. In Figure 4.1, we have sketched an example of the hierarchies of leaders of a given agent i.

Suppose otherwise that the theorem were false on an $[\mathcal{L}](i)$-system for some particular agent i. There would exist some $\bar{j} \in [\mathcal{L}](i)$ and $\bar{t} > 0$ such that $\eta_{\bar{j}}(\bar{t}) < 0$. Define

$$t_* = \inf\{t > 0 \mid \text{there exists some } j \in [\mathcal{L}](i), \text{ such that } \eta_j(t) < 0\}.$$

Then $0 \leq t_* \leq \bar{t} < \infty$, and we claim additionally the following.

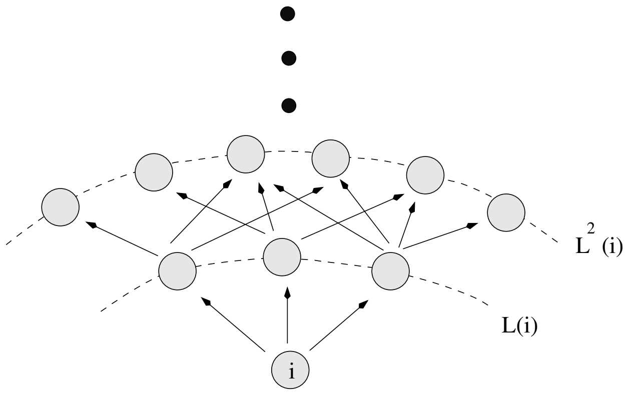


FIG. 4.1. *The leaders of an agent i at different levels:* $\mathcal{L}^0(i) = \{i\}, \mathcal{L}(i), \mathcal{L}^2(i), \ldots$.

(i) For any $j \in [\mathcal{L}](i)$, $\eta_j(t_*) \geq 0$.

(ii) There must exist some $\hat{j} \in [\mathcal{L}](i)$ and a sequence of moments (t_n) such that $t_n > t_*$, $t_n \to t_*$ as $n \to \infty$, and $\eta_{\hat{j}}(t_n) < 0$.

(iii) There must exist some $j_* \in [\mathcal{L}](\hat{j})$, such that $\eta_{j_*}(t_*) > 0$.

(i) and (ii) result directly from the definition of t_*. Suppose otherwise that (iii) were false. Then for any $j \in [\mathcal{L}](\hat{j})$, one must have $\eta_j(t_*) = 0$ by (i). Consider the $[\mathcal{L}](\hat{j})$-system after t_*:

$$\dot{\eta}_j = \sum_{l \in \mathcal{L}(j)} a_{jl}(t)(\eta_l - \eta_j), \qquad j \in [\mathcal{L}](\hat{j}), \quad t > t_*.$$

Since this is a homogeneous system with zero initial conditions at $t = t_*$, by the uniqueness theorem of ODEs (e.g., [13]), the solution to the $[\mathcal{L}](\hat{j})$-system must be identically zero: $\eta_j(t) \equiv 0$ for any $t > t_*$ and $j \in [\mathcal{L}](\hat{j})$. Now that $\hat{j} \in [\mathcal{L}](\hat{j})$, one must have $\eta_{\hat{j}}(t) \equiv 0$ for all $t > t_*$, which contradicts property (ii). Thus (iii) holds.

Define

$$\hat{m} = \min\{m \geq 0 \mid \text{there exists some } j_* \in \mathcal{L}^m(\hat{j}), \text{ such that } \eta_{j_*}(t_*) > 0\}.$$

Properties (ii) and (iii) imply that $0 < \hat{m} < \infty$. Then by iteratively differentiating the $[\mathcal{L}](\hat{j})$-system, one can easily establish

$$\eta_{\hat{j}}(t_*) = \eta'_{\hat{j}}(t_*) = \cdots = \eta_{\hat{j}}^{(\hat{m}-1)}(t_*) = 0, \qquad \eta_{\hat{j}}^{(\hat{m})}(t_*) > 0,$$

which contradicts property (ii). Thus the theorem must hold and the proof is complete. □

The most important consequence is the following bounding capability.

THEOREM 4.2 (boundedness of velocities under evolution). *The Cucker–Smale model* (4.1) *has the following closedness properties.*

(i) *Suppose Ω is a convex compact domain in $\mathbb{R}^3$, and for any agent i, initially $v_i(t = 0) \in \Omega$. Then for any $t > 0$ and i, $v_i(t) \in \Omega$.*

(ii) *In particular, let $D_0 = \max_i |v_i(t = 0)|$. Then $|v_i(t)| \leq D_0$ for all $t > 0$ and i.*

Proof. Since the closed ball $B_{D_0}(0)$ in $\mathbb{R}^3$ is convex and compact, (ii) is implied by (i). It suffices to establish (i).

Given any unit vector $n \in S^2$ and any vector $a \in \mathbb{R}^3$, we first claim that if

$$n \cdot (v_i - a)\big|_{t=0} \geq 0 \qquad \forall\, i,$$

then $n \cdot (v_i(t) - a) \geq 0$ remains valid for all $t > 0$ and i. To proceed, define $\eta_i = n \cdot (v_i - a)$.

$$\begin{aligned}\dot{\eta} &= n \cdot \dot{v}_i \\ &= n \cdot \left(\sum_{j \in \mathcal{L}(i)} a_{ij}(t)(v_j - v_i) \right) \\ &= n \cdot \left(\sum_{j \in \mathcal{L}(i)} a_{ij}(t) \left[(v_j - a) - (v_i - a) \right] \right) \\ &= \sum_{j \in \mathcal{L}(i)} a_{ij}(t)(\eta_j - \eta_i) \\ &= -(L_t \eta)_i.\end{aligned}$$

Then by the preceding theorem, the claim is indeed valid: $\eta_i(t) \geq 0$ for all $t > 0$ and i.

For any compact convex domain Ω, let $p : S^2 \to \mathbb{R}^3$ be its support function, so that for any unit direction $n \in S^2$, $a = p(n)$ has the property that $a \in \partial\Omega$ and the closed flat half-space

$$\pi_{a,-n} = \{x \in \mathbb{R}^3 \mid (-n) \cdot (x - a) \geq 0\}$$

contains Ω. When the domain is convex but not strictly convex, $p(n)$ could be a set of points, which, however, does not influence the argument herein (since the above half-spaces anchored at different points of $p(n)$ would be the same). Furthermore, we have

$$\Omega = \bigcap_{n \in S^2} \pi_{p(n),-n}.$$

Since each half-space has just been shown invariant under the Cucker–Smale evolution, we conclude that Ω must be invariant as well under the evolution, which completes the proof. □

4.2. Perturbation and induction. We now first prepare a lemma. Together with the boundedness property just established above, it facilitates the later analysis on the emergent behavior of HL flocks.

LEMMA 4.3. *Suppose $x(t), v(t) \in \mathbb{R}^3$ (which could be considered as $x_2 - x_1$ and $v_2 - v_1$ for a 2-flock), and satisfy the perturbed 2-flock system* parametrized *by some $T > 0$:*

$$\begin{cases} \dot{x} = v(t), \\ \dot{v} = -a_T(x,t)v(t) + \varepsilon_T(t), \qquad t \geq 0. \end{cases} \tag{4.5}$$

Assume in addition that the following conditions hold.

(i) $a_T(x,t) \geq \frac{H}{(1+|x|^2)^\beta}$, *with $\beta < 1/2$.*

(ii) $\varepsilon_T \in \mathbb{R}^3$, *and*

$$|\varepsilon_T(t)| \leq a e^{-b(t+T)^\eta} \qquad \textit{for some } \eta \in (0,1]. \tag{4.6}$$

(iii) *$|v(t)| \le D_0$ for all $t \ge 0$, and $|x_0| \le R_0 + D_0 T$.*

Here H, β, a, b, η, D_0, and R_0 are given constants independent of T. Let $(x^T(t), v^T(t))$ denote the dependency on T. Then

$$|v^T(T)| \le Ae^{-BT^{(1-2\beta)\wedge\eta^-}}, \tag{4.7}$$

where $\eta^- = \eta - \delta$ for any small but positive δ when $\eta < 1$, and $\eta^- = 1$ when $\eta = 1$, and A and B are two constants depending only upon H, β, a, b, η^-, D_0, and R_0 (but not T). The notation $a \wedge b$ represents $\min(a, b)$.

Before proceeding to the proof of the lemma, we first make two comments on the conditions.

(1) The all-time bound $|v(t)| \le D_0$ seems very stringent but is now natural by Theorem 4.2 in the preceding subsection.

(2) As outlined in the beginning of the current section, the lemma will be applied during the induction process going from the subflock $[1, \dots, i-1]$ to $[1, \dots, i]$. To agent i, the perturbation factor $\varepsilon_T(t)$ comes from the exponentially small deviation of the leading subflock $[1, \dots, i-1]$ from reaching exact consensus.

Proof. From the equation for v, we have

$$\langle v, \dot{v} \rangle = -a_T(x,t)\langle v, v \rangle + \langle v, \varepsilon_T(t) \rangle, \qquad \text{or}$$
$$|v| \cdot |v|_t = \frac{1}{2}\left(|v|^2\right)_t = -a_T|v|^2 + \langle v, \varepsilon_T(t) \rangle.$$

Assuming that v is never identically zero on any nonempty open time interval (note that the opposite scenario trivializes the lemma on any such interval, and the following argument would need only a minor modification), one has

$$\begin{aligned} |v|_t &\le -a_T|v| + |\varepsilon_T| \\ &\le -\frac{H}{(1+|x|^2)^\beta}|v| + ae^{-b(t+T)^\eta} \end{aligned}$$

by conditions (i) and (ii). By $\dot{x} = v$ and (iii),

$$\begin{aligned} |x| &\le |x_0| + \int_0^t |v|(\tau)d\tau \\ &\le R_0 + D_0T + D_0t = R_0 + D_0(t+T). \end{aligned}$$

As a result,

$$|v|_t \le -\frac{H}{(1+(R_0+D_0(t+T))^2)^\beta}|v| + ae^{-b(t+T)^\eta}.$$

Then by Gronwall-type integration,

$$\begin{aligned} |v(t)| &\le |v_0|e^{-\int_0^t \frac{H}{(1+(R_0+D_0(\tau+T))^2)^\beta}d\tau} + a\int_0^t e^{-b(\tau+T)^\eta} \cdot e^{-\int_\tau^t \frac{H}{(1+(R_0+D_0(s+T))^2)^\beta}ds}d\tau \\ &\le D_0 \cdot e^{-\frac{Ht}{(1+(R_0+D_0(t+T))^2)^\beta}} + \frac{a}{b\eta}T^{1-\eta}e^{-bT^\eta}. \end{aligned}$$

We denote $v(t)$ by $v^T(t)$ to indicate its dependency on T. Then

$$\begin{aligned} |v^T(T)| &\le D_0 \cdot e^{-\frac{H \cdot T}{(1+(R_0+2D_0T)^2)^\beta}} + \frac{\tilde{a}(a,b,\eta^-)}{b\eta^-}e^{-bT^{\eta^-}} \\ &\le D_0e^{-\tilde{H}(H,R_0,D_0,\beta)T^{1-2\beta}} + C(a,b,\eta^-)e^{-bT^{\eta^-}} \qquad \text{(when } T \ge 1\text{)} \\ &\le Ae^{-BT^{(1-2\beta)\wedge\eta^-}}, \end{aligned}$$

where the two constants A and B are independent of T. Also notice that when $\eta = 1$, the monomial factor $T^{1-\eta} = 1$ and the lowering from η to η^- is unnecessary in the first line. Finally, since $|v^T(t)| \le D_0$ by the given conditions, by suitably increasing A, the condition $T \ge 1$ in the last second line can actually be removed. This completes the proof. ◻

We are now ready to state and prove the main theorem.

THEOREM 4.4 (convergence of an HL flock). *Let* $[1, 2, \dots, k]$ *be a Cucker–Smale flock under hierarchical leadership with* $\beta < 1/2$. *Then for some* $B > 0$, *which depends only on the initial configuration and all the system parameters, one has*

$$\max_{1\le i,j\le k} |v_i(t) - v_j(t)| = O(e^{-Bt}), \qquad t > 0. \tag{4.8}$$

Proof. We prove the theorem by induction on the subflocks, from $[1, \dots, l-1]$ to $[1, \dots, l]$ (see Figure 4.2).

First we show that the theorem holds for a 2-flock $[1, 2]$. By definition, the leader set $\mathcal{L}(2)$ is nonempty and has to be $\mathcal{L}(2) = \{1\}$, i.e., $a_{21} > 0$. Let $x = x_2 - x_1$ and $v = v_2 - v_1$. Then

$$\begin{cases} \dot{x} = v, \\ \dot{v} = \dot{v}_2 - \dot{v}_1 = \dot{v}_2 = a_{21}(v_1 - v_2) = -a_{21}v. \end{cases}$$

Here $a_{21} = a_{21}(x) = \frac{H}{(1+|x|^2)^\beta}$, with $\beta < 1/2$. Then Cucker and Smale's analysis in [7] still applies directly, and $|v(t)| = O(e^{-Bt})$ for some $B > 0$.

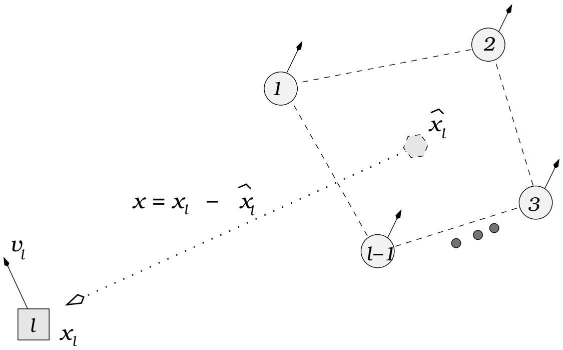


FIG. 4.2. *The induction process from* $[1, \dots, l-1]$ *to* $[1, \dots, l-1, l]$ *reduces the* l*-flock system to a perturbed* 2*-flock system.*

Assuming that the theorem holds for the subflock $[1, \dots, l-1]$, we now intend to show that it must be true for $[1, \dots, l-1, l]$ as well for $l > 2$. As a result, the main focus shall be the agent l.

By induction, there exists some $b > 0$, such that

$$\max_{i,j\in\{1,\dots,l-1\}} |v_i(t) - v_j(t)| = O(e^{-bt}), \qquad t \to \infty. \tag{4.9}$$

Define the average velocity (of the direct leaders of agent l) by

$$\hat{v}_l(t) = \frac{1}{d_l} \sum_{i\in\mathcal{L}(l)} v_i(t), \qquad \text{with } d_l = \#\mathcal{L}(l).$$

Then for any $j \in \mathcal{L}(l)$,

$$|v_j(t) - \hat{v}_l(t)| \le \frac{1}{d_l} \sum_{i \in \mathcal{L}(l)} |v_j - v_i| = O(e^{-bt}) \tag{4.10}$$

by the induction assumption. Similarly, define

$$\hat{x}_l(t) = \frac{1}{d_l} \sum_{i \in \mathcal{L}(l)} x_i(t) \quad \text{and} \quad x(t) = x_l(t) - \hat{x}_l(t), \quad v(t) = v_l(t) - \hat{v}_l(t).$$

Then $\dot{x} = v$, and

$$\begin{aligned} \dot{v} = \dot{v}_l - \frac{d\hat{v}_l}{dt} &= \sum_{j \in \mathcal{L}(l)} a_{lj} \cdot (v_j - v_l) - \frac{d\hat{v}_l}{dt} \\ &= \sum_{j \in \mathcal{L}(l)} a_{lj} \cdot (\hat{v}_l - v_l) + \underbrace{\sum_{j \in \mathcal{L}(l)} a_{lj} \cdot (v_j - \hat{v}_l) - \frac{d\hat{v}_l}{dt}}_{\varepsilon(t)}. \end{aligned} \tag{4.11}$$

Since each $\dot{v}_i$ $(i \in \mathcal{L}(l))$ is the linear combination of some $(v_j - v_i)$'s with $j \in \mathcal{L}(i) \subseteq \{1, \dots, l-1\}$, by (4.9), one must have

$$\frac{d\hat{v}_l}{dt} = O(e^{-bt}).$$

Similarly, due to (4.10) and the boundedness of the a_{lj}'s, one has

$$\left| \sum_{j \in \mathcal{L}(l)} a_{lj} \cdot (v_j - \hat{v}_l) \right| = O(e^{-bt}).$$

In combination, we conclude that

$$|\varepsilon(t)| \le c e^{-bt}, \qquad t > 0, \quad \text{for some } c > 0. \tag{4.12}$$

On the other hand, define

$$a = \sum_{j \in \mathcal{L}(l)} a_{lj} = \sum_{j \in \mathcal{L}(l)} \frac{H}{(1 + |x_j - x_l|^2)^\beta}. \tag{4.13}$$

Then (4.11) simply becomes

$$\dot{v} = -av + \varepsilon. \tag{4.14}$$

Define $g(s) = \frac{H}{(1+s)^\beta}$ with $s \ge 0$. Then $g(s)$ is convex, and

$$\frac{1}{d_l} \sum_{j \in \mathcal{L}(l)} g(s_j) \ge g\left(\frac{1}{d_l} \sum_{j \in \mathcal{L}(l)} s_j \right).$$

As a result, when $s_j = |x_j - x_l|$,

$$\sum_{j \in \mathcal{L}(l)} \frac{H}{(1 + |x_j - x_l|^2)^\beta} \ge d_l \frac{H}{\left(1 + \frac{1}{d_l} \sum_{j \in \mathcal{L}(l)} |x_j - x_l|^2\right)^\beta}. \tag{4.15}$$

By the least-squares principle,

$$\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|x_l-x_j|^2=|\overbrace{x_l-\hat{x}_l}^{x}|^2+\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|x_j-\hat{x}_l|^2, \tag{4.16}$$

since $\hat{x}_l$ is the center or mean of $\{x_j \mid j\in\mathcal{L}(l)\}$. By the induction assumption on the consensus of $[1,\ldots,l-1]\supseteq\mathcal{L}(l)$, it is easy to establish that there exists some $M>0$, such that

$$\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|x_j-\hat{x}_l|^2\le M-1. \tag{4.17}$$

(Simply notice that the induction assumption guarantees the exponentially small deviations of the v_j's from $\hat{v}_l$.) Combining (4.13)–(4.17), we have

$$a=a(x,t)\ge\frac{d_lH}{(M+|x|^2)^\beta}\ge\frac{\tilde{H}}{(1+|x|^2)^\beta}, \tag{4.18}$$

where the updated constant $\tilde{H}=\tilde{H}(H,d_l,M,\beta)$. (Notice that the notation $a(x,t)$ summarizes all the influence from $\{x_j \mid j\in\mathcal{L}(l)\}$ into the t-variable.)

The combination of (4.12), (4.14), and (4.18) leads to the reduced system

$$\begin{cases}\dot{x}=v,\\ \dot{v}=-a(x,t)v+\varepsilon(t),\end{cases} \tag{4.19}$$

with $a(x,t)\ge\frac{\tilde{H}}{(1+|x|)^\beta}$ and $|\varepsilon(t)|\le ce^{-bt}$. In order to apply Lemma 4.3, further define

$$D_0=2\max_{1\le i\le k}|v_i(t=0)| \quad\text{and}\quad R_0=2\max_{1\le i\le k}|x_i(t=0)|. \tag{4.20}$$

Then by Theorem 4.2, we have

$$|v_i(t)|\le\frac{D_0}{2}\qquad\forall\ i\text{ and }t>0.$$

Consequently,

$$|v(t)|\le\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|v_j-v_l|\le\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}D_0=D_0. \tag{4.21}$$

Similarly, for any $T>0$,

$$\begin{aligned}|x(T)-x(0)|&\le|x_l(T)-x_l(0)|+\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|x_j(T)-x_j(0)|\\ &\le\frac{D_0}{2}T+\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}\frac{D_0}{2}T=D_0T.\end{aligned}$$

As a result,

$$\begin{aligned}|x(T)|&\le|x(0)|+D_0T\\ &\le|x_l(0)|+\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}|x_j(0)|+D_0T\\ &\le\frac{R_0}{2}+\frac{1}{d_l}\sum_{j\in\mathcal{L}(l)}\frac{R_0}{2}+D_0T=R_0+D_0T.\end{aligned} \tag{4.22}$$

To conclude, for any $T > 0$, if we define

$$x^T(t) = x(t+T), \quad v^T(t) = v(t+T), \quad a_T(x^T, t) = a(x^T, t+T), \quad \text{and} \quad \varepsilon_T(t) = \varepsilon(t+T),$$

then

$$\begin{cases} \dot{x}^T = v^T, \\ \dot{v}^T = -a_T(x^T, t)v^T + \varepsilon_T(t), \qquad t > 0, \end{cases}$$

and all three conditions of Lemma 4.3 are satisfied (with $\eta = 1$). Therefore, there must exist two positive constants $\tilde{A}$ and $\tilde{B}$, such that for any $T > 0$,

$$|v(2T)| = |v^T(T)| \le \tilde{A}e^{-\tilde{B}T^{(1-2\beta)\wedge 1}} = \tilde{A}e^{-\tilde{B}T^{1-2\beta}}.$$

Since T is arbitrary, we therefore must have, after adjusting the constants,

$$|v(t)| \le \hat{A}e^{-\hat{B}t^{1-2\beta}}, \qquad t > 0, \qquad \text{for some constants } \hat{A} \text{ and } \hat{B}.$$

Moreover, since $\beta < 1/2$ by assumption, one then must have

$$\int_0^\infty |v(t)|dt < \infty,$$

which in return implies that there exists some constant $M > 0$, such that

$$|x(t)| \le M, \qquad t > 0.$$

Then by repeating the similar calculation in the proof of Lemma 4.3, assisted by this new constant bound $|x| \le M$ instead of $|x| \le R_0 + D_0(t+T)$ there, one arrives at

$$|v(t)| \le A'e^{-B't} \qquad (\text{since } \eta = 1)$$

for two positive constants A' and B' independent of t. Combined with the induction base (4.9), we thus conclude that the theorem must hold true for the subflock $[1, \dots, l-1, l]$ with the exponent coefficient $B = B' \wedge b$. This completes the proof. □

5. HL flocking under a free-will leader. In this section, partially inspired by the preceding perturbation methods, we investigate a more realistic scenario in which the ultimate leader agent 0 (in an HL flock $[0, 1, \dots, k]$) can have a *free-will acceleration*, instead of merely flying with a constant velocity.

The following phenomenon is not uncommon near lakes, grasslands, or any open spaces where a flock of birds often visits. When the flock is initially approached by an unexpected pedestrian or a predator from a corner on the outer rim, the bird which takes off first (and alerts others subsequently) generally takes a curvy flying path before it reaches a stable flying pattern with an almost constant velocity. Such a bird gains the full speed fast, flies ahead of the entire flock, and serves as a virtual overall leader.

For an HL flock $[0, 1, \dots, k]$, in addition to the Cucker–Smale system

$$(5.1) \qquad \begin{cases} \dot{x}_i = v_i(t), \\ \dot{v}_i = \sum_{j \in \mathcal{L}(i)} a_{ij}(x)(v_j(t) - v_i(t)), \qquad i > 0, \end{cases}$$

we now also impose for the ultimate leader agent 0

$$\begin{cases} \dot{x}_0 = v_0, \\ \dot{v}_0 = f(t), \qquad t > 0, \end{cases} \tag{5.2}$$

coupled with a given set of initial conditions. For convenience, we shall call $f(t)$ the *free-will acceleration* of the leader. In combination, the new system is no longer autonomous.

The main goal of this section is to establish the following theorem.

THEOREM 5.1. *Suppose an HL $(k+1)$-flock $[0, \dots, k]$ with a free-will leader satisfies both* (5.1) *and* (5.2)*, with the Cucker–Smale connectivity strength of $\beta < 1/2$. In addition, assume that the leader's free-will acceleration satisfies*

$$|f(t)| = O((1+t)^{-\mu}), \qquad \textit{with some exponent } \mu > k.$$

Then the flock still has the following emergent behavior:

$$\max_{0 \le i,j \le k} |v_i - v_j|(t) = O\left((1+t)^{-(\mu-k)}\right).$$

We first make two comments regarding why one should expect to put some regularity conditions on the leader's behavior in order for a coherent pattern to emerge asymptotically.

(1) Intuitively, if the leader keeps changing its velocity substantially, it will be more difficult for the entire flock to follow and behave coherently. An extreme example is a flock with a *drunken* leader which flies in a Brownian random path. Then the entire flock cannot be expected to synchronize with the unpredictable motion of the leader instantaneously.

(2) In the theorem, the decaying constraint $\mu > k$ depends on the size k of the flock. Thus qualitatively speaking, it requires the leader to exert less free will when the flock is larger, in order to lead a coherent flock asymptotically. Consider the special hierarchical leadership under a *linear chain of command*:

$$k \to k-1 \to \cdots \to 1 \to 0.$$

The tail agent k has to go through all the k intermediate stages to sense any move that the leader is making. Thus intuitively, there will be a long time delay in between, and the leader has to be tempered enough to allow its distantly connected followers to respond coherently.

We now prepare a lemma that is similar to Lemma 4.3. Since the new nonautonomous system does not necessarily have the positivity property, we take a slightly different approach.

LEMMA 5.2. *Let $x, v, g \in \mathbb{R}^3$ satisfy*

$$\begin{cases} \dot{x} = v(t), \\ \dot{v} = -a(x,t)v(t) + g(t). \end{cases}$$

Suppose that

$$a(x,t) \ge \frac{H}{(1+|x|^2)^\beta} \qquad \textit{for some } \beta < 1/2, \textit{ and}$$
$$|g(t)| = O\left((1+t)^{-\eta}\right), \qquad \textit{with some constant } \eta > 1.$$

Then $|v(t)| = O((1+t)^{-(\eta-1)})$ *with the order constant depending only on the initial conditions* $x(t=0)$, $v(t=0)$, *and* H, β, *and* η.

Proof. From the second equation, one has

$$|v| \cdot |v|_t = \left(\frac{v^2}{2}\right)_t = \langle v, v_t\rangle = -a\langle v, v\rangle + \langle v, g\rangle \leq -a|v|^2 + |v| \cdot |g|.$$

Assume that v does not vanish identically on any nonempty open intervals for the same reason as in the proof of Lemma 4.3. Then one has

$$|v|_t \leq -a|v| + |g|, \qquad t > 0.$$

Fix any time $T > 0$, and define

$$|x|_* = \sup_{t \leq T} |x|(t) \quad \text{and} \quad a_* = \inf_{t \leq T} \frac{H}{(1+|x|^2)^\beta} = \frac{H}{(1+|x|_*^2)^\beta}. \tag{5.3}$$

Then one has

$$|v|_t \leq -a_*|v| + |g|, \qquad t \in [0, T]. \tag{5.4}$$

Since a_* is constant, integration yields

$$|v|(t) \leq |v|(0)e^{-a_* t} + \int_0^t |g|(\tau)e^{-a_*(t-\tau)}d\tau.$$

In particular, for any $t < T$,

$$|v|(t) \leq |v|(0) + \int_0^t |g|(\tau)d\tau \leq |v|(0) + \int_0^\infty |g(\tau)|d\tau := A_0.$$

(Since $\eta > 1$ by assumption, the integral of $|g|$ is finite.) Now that A_0 is independent of the time mark T, we conclude that the last upper bound must hold for *any* $t > 0$: $|v|(t) \leq A_0$, $t > 0$. Therefore, from the first equation $\dot{x} = v(t)$, one has

$$|x|(t) \leq |x|(0) + \int_0^t |v|(\tau)d\tau \leq B_0 + A_0 t, \quad t > 0,$$

where $B_0 = |x|(0)$. In particular, for any time mark $T > 0$, the quantities in (5.3) are subject to

$$|x|_* \leq B_0 + A_0 T \quad \text{and} \quad a_* \geq \frac{H}{[1+(B_0+A_0T)^2]^\beta}.$$

We then go back and integrate inequality (5.4) again, but from $T/2$ to T this time:

$$\begin{aligned}
|v|(T) &\leq |v|(T/2)e^{-\frac{a_*T}{2}} + \int_{T/2}^T |g|(\tau)e^{-a_*(T-t)}dt \\
&\leq A_0 e^{-\frac{HT/2}{[1+(B_0+A_0T)^2]^\beta}} + \int_{T/2}^\infty |g|(t)dt \\
&\leq A_0 e^{-\tilde{H}(A_0,B_0,\beta)(1+T)^{1-2\beta}} + \int_{T/2}^\infty O\left((1+t)^{-\mu}\right)dt \\
&= A_0 e^{-\tilde{H}(1+T)^{1-2\beta}} + O\left((1+T)^{-(\mu-1)}\right).
\end{aligned}$$

Since $\beta < 1/2$, we conclude that

$$|v|(T) = O\left((1+T)^{-(\mu-1)}\right),$$

where the constant in $O(\cdot)$ is independent of T. Since T is arbitrary, the lemma is established. □

We are now ready to prove Theorem 5.1. Details on some similar calculations will be directed to the proof of Theorem 4.4.

Proof. It suffices to prove the following more general result:

$$\max_{0\le i,j\le l} |v_i - v_j|(t) = O\left((1+t)^{-(\mu-l)}\right), \qquad t > 0, \tag{5.5}$$

for any subflock $[0, 1, \ldots, l]$ and $l \ge 1$.

When $l = 1$, define $x = x_1 - x_0$ and $v = v_1 - v_0$. Then $\dot{x} = v$, and

$$\dot{v} = \dot{v}_1 - \dot{v}_0 = a_{10}(v_0 - v_1) - f = -a_{10}v - f.$$

By the definition of an HL flock, $\mathcal{L}(1) \ne \varnothing$, and it has to be agent 0, implying that a_{10} is subject to the Cucker–Smale formula. Then by the preceding lemma (with $\eta = \mu$), one has

$$|v|(t) = O\left((1+t)^{-(\mu-1)}\right),$$

and (5.5) holds.

Suppose now that (5.5) is true for the subflock $[0, 1, \ldots, l-1]$ with $2 \le l \le k$, so that

$$\max_{0\le i,j\le l-1} |v_i - v_j|(t) = O\left((1+t)^{-(\mu-l+1)}\right). \tag{5.6}$$

As in the proof of Theorem 4.4, define the average features of the direct leaders of agent l by

$$\hat{x}_l = \frac{1}{d_l}\sum_{j\in\mathcal{L}(l)} x_j \quad \text{and} \quad \hat{v}_l = \frac{1}{d_l}\sum_{j\in\mathcal{L}(l)} v_j, \qquad d_l = \#\mathcal{L}(l),$$

and $x = x_l - \hat{x}_l$ and $v = v_l - \hat{v}_l$.

Then as in the proof of Theorem 4.4, one has $\dot{x} = v$ and

$$\begin{aligned} \dot{v} &= -a(x,t)\cdot v + g_l(t), \qquad \text{with} \\ g_l(t) &= \sum_{j\in\mathcal{L}(l)} a_{lj}\cdot(v_j - \hat{v}_l) - \frac{d\hat{v}_l}{dt}, \\ a(x,t) &= \sum_{j\in\mathcal{L}(l)} a_{lj}(x_l - x_j). \end{aligned}$$

We first estimate g_l. Since $|a_{lj}| \le H$ and $\mathcal{L}(l) \subseteq [0, 1, \ldots, l-1]$, by the induction assumption (5.6), the first term in g_l must be of the order $O((1+t)^{-\eta})$ with $\eta = \mu - l + 1$. For the remaining second term in g_l, let $1_{0\in\mathcal{L}(l)}$ denote the logical variable, which is 1 when agent 0 belongs to $\mathcal{L}(l)$, and 0 otherwise. Then

$$\frac{d\hat{v}_l}{dt} = \frac{1}{d_l}\sum_{j\in\mathcal{L}(l)} \dot{v}_j = 1_{0\in\mathcal{L}(l)}\cdot\frac{1}{d_l}\dot{v}_0 + \frac{1}{d_l}\sum_{j\in\mathcal{L}(l)\setminus\{0\}} \dot{v}_j.$$

Notice that $\dot{v}_0 = f(t) = O((1+t)^{-\mu})$, and each $\dot{v}_j$ with $j \in \mathcal{L}(l) \setminus \{0\}$ is some linear combination of $(v_s - v_j)$ with the s's in $\mathcal{L}(j) \subseteq [0, 1, \ldots, l-1]$. Thus by the induction assumption (5.6), one must have $\frac{d\hat{v}_l}{dt} = O((1+t)^{-\eta})$ with $\eta = \mu - l + 1$.

We now estimate $a(x,t)$. Since $\mu > k$ by the given condition, we have $\mu - l + 1 > k - l + 1 = 1$. As a result, by the induction assumption on the subflock $[0, 1, \ldots, l-1]$, for any $i, j \le l-1$,

$$\begin{aligned} |x_i - x_j|(t) &\le |x_i - x_j|(0) + \int_0^t |v_i - v_j|(\tau)d\tau \\ &\le |x_i - x_j|(0) + \int_0^\infty O\left((1+\tau)^{-(\mu-l+1)}\right)d\tau < \infty \qquad \forall\, t > 0. \end{aligned}$$

Therefore the boundedness property in (4.17) still holds, and the same calculation in the proof of Theorem 4.4 leads to

$$a(x,t) \ge \frac{\tilde{H}}{(1+|x|^2)^\beta}$$

for some constant $\tilde{H} = \tilde{H}(H, d_l, \beta, f, \text{initial conditions of } [0, \ldots, l-1])$.

Combining the estimations on g_l and a, one sees that $x(t)$ and $v(t)$ satisfy a perturbed system as in Lemma 5.2 with $\eta = \mu - l + 1$. Therefore, by Lemma 5.2,

$$|v_l - \hat{v}_l|(t) = |v|(t) = O\left((1+t)^{-(\eta-1)}\right) = O\left((1+t)^{-(\mu-l)}\right).$$

Now that by the induction assumption, for any $j \le l-1$, one must have

$$|v_j - \hat{v}_l|(t) = O\left((1+t)^{-(\mu-l+1)}\right), \quad \text{since } |v_j - v_i| = O\left((1+t)^{-(\mu-l+1)}\right) \quad \forall\, i \in \mathcal{L}(l).$$

Therefore, for any $j \le l-1$,

$$|v_l - v_j| \le |v_l - \hat{v}_l| + |v_j - \hat{v}_l| = O\left((1+t)^{-(\mu-l)}\right).$$

This completes the proof of (5.6), and thus the entire theorem. □

COROLLARY 5.3. *Under the same statements as in the preceding theorem, suppose $\mu > k+1$. Then there exists a constant configuration $(d_{ij})_{0 \le i,j \le k}$ with $d_{ij} \in \mathbb{R}^3$, such that*

$$\lim_{t \to \infty} (x_i(t) - x_j(t)) = d_{i,j}, \qquad 0 \le i, j \le k,$$

and the convergence rate is $O\left((1+t)^{-(\mu-k-1)}\right)$.

6. Conclusion. In this paper, we have investigated the emergent behavior of Cucker–Smale flocking under the structure of *hierarchical leadership* (HL). The convergence rates are established for both discrete-time and continuous-time HL flocks, as well as for HL flocks under an overall leader with free-will accelerations. In all these scenarios, the consistent convergence towards some asymptotically coherent patterns may reveal the advantages and necessities of having leaders and leadership in a complex (biological, technological, economic, or social) system with sufficient intelligence and memory.

As reviewed in the introduction section, there have been explosive multidisciplinary efforts in recent years in advancing this emerging area of multiagent complex systems with biological or intelligence signatures (as opposed to more traditional and *passive* particle systems in physics). The innate complexities (associated with multiagency, probing and sensing, intelligence, and dynamical reorganization and transience) have made rigorous mathematical analysis of such systems highly challenging. Therefore, to no one's surprise, the aim of most contemporary works has been to

(i) construct complex models that reproduce specific biological phenomena or design technology-driven applications, e.g., for ant armies, elephant herds, or robotic or sensor networks; and

(ii) simulate (via computing) these complex models, fine-tuning model parameters to obtain good approximations of the targeted biological or technological flocking phenomena.

As is well known in information theory [5] or the most recent learning theory (e.g., Vapnik [26] and Cucker and Smale [6]), a universal tradeoff always exists between the *complexity* of a model and its so-called *generalization* power, or the power of interpretation. Most practical flocking models in the several disciplines mentioned above have mainly attempted to reproduce faithful real-life flocking behavior, i.e., the *interpretation* power. Typically, these models involve many interacting terms, often nonlinear, nonconvex, or nonlocal. They are built upon good heuristic intuition but defy attempts at rigorous mathematical analysis.

On the other hand, recently the collective efforts of a group of mathematicians, including, for example, Cucker, Smale, Bertozzi, and their colleagues, have been focused on scaling down some model complexity at the necessary loss of certain interpretation power. The reduced models become more tractable mathematically, and also offer their developers deeper insights into more involved versions of the models. As in other major branches of complexity study, e.g., fluid dynamics or statistical mechanics [1], complexity reduction has often been made possible by

(a) incorporating suitable symmetries and invariants;

(b) reducing the number of significant signals or governing variables; and

(c) introducing topological, combinatorial, or dynamical structures,

as witnessed in the Cucker–Smale model [7, 8] and the current work.

It is within this general picture of multidisciplinary flocking modeling that the present work should be embedded. This author, and probably other interested researchers from the applied mathematics community as well, will continue to increase model complexities subject to mathematical rigor and analytic tractability.

Acknowledgments. The author thanks both Professors Steve Smale and Felipe Cucker for their generosity in sharing their ideas on their developing frameworks. This work would not have been possible without the patient daily guidance, encouragement, and numerous suggestions from Prof. Smale. The author is profoundly grateful for the tender care of Prof. Smale and Prof. David McAllester, as well as for the generous visiting support from the Toyota Technological Institute (TTI-C) on the campus of the University of Chicago during the fall semester of 2006. The author also thanks his former Ph.D. advisor, Prof. Gil Strang, for directing him to the works of Prof. Iain Couzin on effective leadership. Without the initial inspiration from Professor Andrea Bertozzi or her constant support and encouragement to the author, the present work also could not have come into existence. Finally, the author acknowledges delightfully the professional help from the referees and managing editor.

REFERENCES

[1] D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York, Oxford, 1987.

[2] Y.-L. Chuang, M. R. D'Orsogna, D. Marthaler, A. L. Bertozzi, and L. Chayes, *State transitions and the continuum limit for a 2D interacting, self-propelled particle system*, Phys. D, submitted.

[3] F. R. K. Chung, *Spectral Graph Theory*, AMS, Providence, RI, 1997.

[4] I. D. Couzin, J. Krause, N. R. Franks, and S. Levin, *Effective leadership and decision making in animal groups on the move*, Nature, 433 (2005), pp. 513–516.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[6] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2001), pp. 1–49.

[7] F. Cucker and S. Smale, *Emergent behavior in flocks*, IEEE Trans. Automat. Control, 52 (2007), pp. 852–862.

[8] F. Cucker and S. Smale, *Lectures on emergence*, Japan J. Math., 2 (2007), pp. 197–227.

[9] F. Cucker, S. Smale, and D. Zhou, *Modeling language evolution*, Found. Comput. Math., 4 (2004), pp. 315–343.

[10] I. Daubechies and J. C. Lagarias, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[11] M. R. D'Orsogna, Y.-L. Chuang, A. L. Bertozzi, and L. Chayes, *Self-propelled particles with soft-core interactions: Patterns, stability, and collapse*, Phys. Rev. Lett., 96 (2006), 104302.

[12] G. Flierl, D. Grünbaum, S. Levin, and D. Olson, *From individuals to aggregations: The interplay between behavior and physics*, J. Theoret. Biol., 196 (1999), pp. 397–454.

[13] M. W. Hirsch, S. Smale, and R. Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, 2nd ed., Academic Press, New York, 2003.

[14] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[15] H. Levine and W.-J. Rappel, *Self-organization in systems of self-propelled particles*, Phys. Rev. E, 63 (2000), article 017101.

[16] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., AMS, Providence, RI, 2001.

[17] Y. Liu and K. Passino, *Stable social foraging swarms in a noisy environment*, IEEE Trans. Automat. Control, 49 (2004), pp. 30–44.

[18] J. K. Parrish, S. V. Viscido, and D. Grünbaum, *Self-organized fish schools: An examination of emergent properties*, Biol. Bull., 202 (2002), pp. 296–305.

[19] G.-C. Rota and G. Strang, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.

[20] J. Shen, *Compactification of a set of matrices with convergent infinite products*, Linear Algebra Appl., 311 (2000), pp. 177–186.

[21] J. Shen, *A geometric approach to ergodic non-homogeneous Markov chains*, in Wavelet Analysis and Multiresolution Methods, Lecture Notes in Pure and Appl. Math. 212, Marcel Dekker, New York, 2000, pp. 341–366.

[22] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.

[23] C. M. Topaz and A. L. Bertozzi, *Swarming patterns in a two-dimensional kinematic model for biological groups*, SIAM J. Appl. Math., 65 (2004), pp. 152–174.

[24] C. M. Topaz, A. L. Bertozzi, and M. A. Lewis, *A nonlocal continuum model for biological aggregation*, Bull. Math. Bio., 68 (2006), pp. 1601–1623.

[25] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 803–812.

[26] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[27] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Schochet, *Novel type of phase transition transition in a system of self-driven particles*, Phys. Rev. Lett., 75 (1995), pp. 1226–1229.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 720–737

NEURAL TIMING IN HIGHLY CONVERGENT SYSTEMS*

COLLEEN MITCHELL† AND MICHAEL REED‡

Abstract. In order to study how the convergence of many variable-response neurons on a single target can sharpen timing information, we investigate the limit as the number of input neurons and the number of incoming spikes required to fire the target both get large with the ratio fixed. We prove that the standard deviation of the firing time of the target cell goes to zero in this limit, and we derive the asymptotic forms of the density and the standard deviation near the limit. We use the theorems to understand the behavior of octopus cells in the mammalian cochlear nucleus.

Key words. neural networks, precision, timing, convergence, octopus cells

AMS subject classifications. 92, 60

DOI. 10.1137/07068775X

1. Introduction.

A fundamental question in neurobiology is to understand how the central nervous system (CNS) can perform accurate and reliable calculations with neurons that are intrinsically variable and unreliable devices. Three more concrete versions of the question have received much attention: (1) How can network and/or cellular properties sharpen timing information or create accurate coincidence detectors? (2) How can synchronous activity in large groups of neurons be created and maintained? (3) Under what circumstances can intrinsic noise improve information processing capabilities?

The first question has long been studied in the auditory brainstem because it is experimentally accessible and because cellular, behavioral, and psychoacoustic experiments show that the auditory system can make extremely fine timing distinctions in the microsecond (or even nanosecond) range, even though individual neurons in the auditory nerve (AN) show latency standard deviations of approximately one millisecond in repeated trials with the same sound [25, 10, 45, 33, 50]. Lord Rayleigh [34] first proposed that the auditory system uses binaural timing distinctions to localize sound, and Jeffress [21] proposed the first neural mechanism based on delay lines and coincidence detection. Colburn [8] clearly formulated the question of how the auditory system can detect small time differences, given the noise in the AN, and went on to create some of the first mathematical models [9]. Important experimental studies include those of Rhode and Smith [37, 38] and Goldberg and Brown [14].

All fibers of the AN synapse on cells of the cochlear nucleus (CN). There are many different cell types in the CN that receive different numbers of AN synapses and have different response properties. Two experimental properties have received continuing attention from experimentalists and modelers. First, several CN cell types show "onset" responses; that is, they fire a single spike shortly after the initiation of the sound; the time lag is called the latency. The standard deviation of latency in AN fibers under repeated trials is of the order of 1 msec, but the standard deviation of latency in some

*Received by the editors May 22, 2006; accepted for publication (in revised form) April 11, 2007; published electronically December 20, 2007. The research in this paper and its preparation were supported by NSF grants DMS-0109872 and DMS-0616710.

http://www.siam.org/journals/siap/68-3/66050.html

†Department of Mathematics, University of Iowa, 14 MacLean Hall, Iowa City, IA 52242-1419 (colleen-mitchell@uiowa.edu).

‡Department of Mathematics, Duke University, Science Drive, Durham, NC 27708 (reed@duke.math.edu).

onset units of the CN is as much as an order of magnitude lower. Second, AN fibers phase lock to low frequency sounds, and this phase locking is even better in some CN units. Burkitt and Clark [4, 5] use numerical simulations of leaky integrate-and-fire models to study how convergence of inputs effects both the onset response and the increase in synchrony seen in target cells. Kalluri and Delgutte [23, 24] have created a computational model using leaky integrate-and-fire for the CN target cells and an adaptively filtered Poisson processes to model spike trains along each of the convergent AN fibers. They are interested in determining what properties of the target cell—the filtered Poisson process in AN fibers, the convergence from AN fibers to the target, and adaptation in the hair cells—cause the target neuron to have an onset response with low spontaneous rate. Young, Rothman, and coworkers [51, 39, 40, 41, 42, 43] have conducted experiments and used numerical simulations of biophysical models to investigate how the response properties of CN neurons depend on the details of their channel kinetics. Similarly, Cai, Walsh, and McGee [6, 7] used simulations of biophysical models to investigate the onset response of octopus cells in the CN using the physiological properties discovered by Oertel and coworkers [15, 13, 32].

The overall goal of this computational modeling was to investigate how convergence, detailed biophysics of CN neurons, and the known properties of auditory spike trains give rise to onset responses, higher synchrony, and the sharpening of timing in CN neurons. The numerical computations of the above investigators suggest strongly that there is a connection between the amount of convergence and the sharpening of timing information. However, their models are so elaborate and have so many parameters that it is difficult to make precise the mechanisms by which convergence sharpens timing. For this reason, we have been studying the much simpler model described below in which convergence and the sharpening of timing are isolated as the objects of study. The simplicity of the model allows us to use the tools of probability theory and mathematical statistics to prove theorems that make precise the relationship between convergence and the sharpening of timing. The previous numerical modeling and our theorems not only contribute to understanding the auditory brainstem but also address question (1) and by implication question (2) if several of these systems are connected in series. In other auditory brainstem work, Svirskis et al. [48] investigate through experiment and biophysical modeling the properties of medial superior olive neurons that make them excellent coincidence detectors and propose that coincidence detection is improved in some circumstances by a noisy background. Thus their work relates directly to the ideas of stochastic resonance in neural systems put forward by Greenwood et al. [17] and Stemmler [47] and so addresses question (3).

In our simple convergence model (see Figure 1), there are n identical input neurons, and all receive the same stimulus. Each fires a single action potential at a time selected independently from a probability density f with standard deviation 1 msec. The axons of the n neurons are of equal length and project to one target neuron that fires a single action potential the first time that it has received m inputs in the previous ε msec. Of course, the target neuron may not fire at all in response to a particular stimulus. We denote the conditional density of the time of firing of the target neuron, given that it fires, by $g_{n,m,\varepsilon,f}$ and its standard deviation by $\sigma_{n,m,\varepsilon,f}$, since both will depend on n, m, ε, and f. If $\sigma_{n,m,\varepsilon,f} < 1$ msec, then we say that timing has been sharpened. A change of variables shows that there is a scaling law $\sigma_{n,m,\varepsilon,f} = s\sigma_{n,m,\frac{\varepsilon}{s},f_s}$, where $f_s(t) \equiv sf(st)$, so there is no loss in generality in taking the standard deviation of the input density f to be 1 msec [35, 29].

Although the formulation of the problem is simple, it is difficult or impossible to derive closed form expressions for $g_{n,m,\varepsilon,f}$ and $\sigma_{n,m,\varepsilon,f}$ except in very special cases.

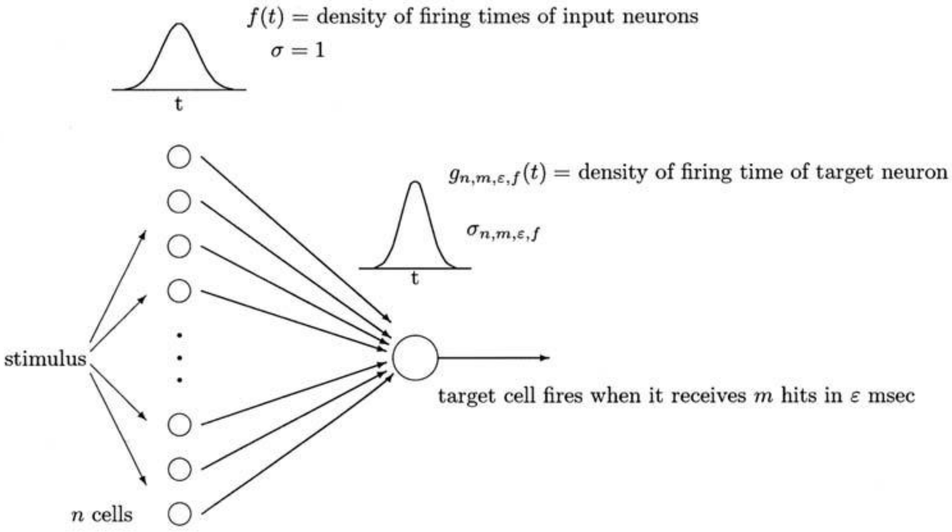

$f(t)$ = density of firing times of input neurons
$\sigma = 1$
t
$g_{n,m,\varepsilon,f}(t)$ = density of firing time of target neuron
$\sigma_{n,m,\varepsilon,f}$
t
stimulus
target cell fires when it receives m hits in ε msec
n cells

FIG. 1. *The basic model. The target cell receives n independently and identically distributed inputs and fires the first time it receives m within ε msec.*

Additionally, Monte Carlo simulations [35] show that, for n and f fixed, $\sigma_{n,m,\varepsilon,f}$ can have surprisingly complicated behavior as a function of m, the numbers of hits required, and ε, the size of the time window. For example, even for simple choices of f (uniform, exponential, normal) and $n = 10$, $\sigma_{n,m,\varepsilon,f}$ is sometimes monotone and sometimes nonmonotone as a function of either m or ε with the other held fixed. In these circumstances, it is natural to ask whether the behavior of $\sigma_{n,m,\varepsilon,f}$ is simpler in certain asymptotic limits. As $\varepsilon \to \infty$, the neuron will surely fire at the time of the mth hit, so $\sigma_{n,m,\infty,f}$ is given by order statistics, which is well understood. In the literature this model is called the (nonleaky) integrate-and-fire model. It is used by Marsalek, Koch, and Maunsell [27], it is the simplest model used by Burkitt and Clark [4], and it is the "analytic coincidence detector model" of Kalluri and Delgutte [23]. Thus those models are a special case of our model. Mitchell [29, 30] considered the singular asymptotic limit $\varepsilon \to 0$, proved that

$$g_{n,m,\varepsilon,f} \to \frac{f^m}{\int f^m} \tag{1}$$

independent of n, and derived an asymptotic form for $g_{n,m,\varepsilon,f}$ and $\sigma_{n,m,\varepsilon,f}$ near the limit.

In this paper, we study the limit as $n \to \infty$, $m \to \infty$ with the ratio $\frac{m}{n}$ held fixed. There are good theoretical and experimental reasons to think that this limit is important. First, it is relatively easy to see (Theorem 2.1 below), under reasonable hypotheses on f, that $\sigma_{n,m,\varepsilon,f} \to 0$ as $n \to \infty$ with m held fixed. That is, one can sharpen up timing as much as one wants by assuming a model with n large, for example by choosing $m = 1$, in which case the target always fires at first hit. Young, Robert, and Schofner [51] already pointed out that if f is exponential and $m = 1$, then $\sigma_{n,1,\varepsilon,f} = \frac{1}{n}$. The trouble is that many neurons (in particular those in the auditory nerve) have high spontaneous firing rates, and so with high n and low m the target cell will have a high spontaneous rate, which ruins its role as a neuron that measures the time since the stimulus. Thus it is natural to ask whether taking both n *and* m large can produce a system that sharpens timing dramatically but also has a very low spontaneous rate. The approximate calculations of Burkitt and Clark [4] and the numerical simulations of Kalluri and Delgutte [23] suggest strongly that there should be a clean asymptotic limit as $n \to \infty$ with $\frac{m}{n}$ held fixed. Furthermore, there is good reason to think that octopus cells in the cochlear nucleus operate near this asymptotic limit. Oertel [32] has found that the octopus cells receive up to 100 inputs from AN fibers and that between 20 and 50 hits within a small time window are required to make them fire.

In section 2, we answer the above question by showing that $\sigma_{n,m,\varepsilon,f} \to 0$ as $n \to \infty$, $m \to \infty$, with $\frac{m}{n}$ fixed. In section 3, we derive the asymptotic forms of $g_{n,m,\varepsilon,f}$ and $\sigma_{n,m,\varepsilon,f}$ near the limit. And in section 4, we apply the results to octopus cells.

2. Limit theorems as $n \to \infty$. Let $\{X_i\}_{i=1}^n$ denote the n independently and identically distributed random variables for the firing times of the inputs. Assume that the X_i's have density $f(x)$, continuous distribution $F(x)$, and finite mean and standard deviation. We will use T_n to denote the random variable for the firing time of the output (a formal definition is given in (6)). For some of the results below we also assume that there is an x_o so that

$$\int_{-\infty}^{x_o} f(x)\,dx = 0 \quad \text{and} \quad \int_{x_o}^{x+a} f(x)\,dx \ > 0 \quad \text{for all } \ a > 0. \tag{2}$$

This is reasonable biologically since the input neurons cannot respond before the stimulus (and perhaps not for some fixed delay afterwards).

We first consider the case where $n \to \infty$ while m and ε are fixed.

THEOREM 2.1. *Let f be a given probability density satisfying* (2), *and let $0 < \varepsilon \leq \infty$ and m be fixed. Then, as $n \to \infty$, the following hold*:

(i) *The probability that the target cell fires $\longrightarrow 1$.*

(ii) *$g_{n,m,\varepsilon,f} \longrightarrow \delta_{x_o}$; that is, $T_n \longrightarrow x_0$ in distribution. Further, T_n converges in probability to the point mass at x_o.*

(iii) *If f has compact support, then $\sigma_{n,m,\varepsilon,f} \longrightarrow 0$.*

Proof. Let $a > 0$ be given, and define $\gamma \equiv \int_{x_o}^{x_o+a} f(x)\,dx \ > 0$. Let Y_i be the random variable with value 1 if $x_o \leq X_i \leq x_o + a$ and zero otherwise. Then, for each k,

$$P\left\{\sum_{i=1}^n Y_i = k\right\} \ = \ \gamma^k(1-\gamma)^{n-k}\binom{n}{k},$$

so that

$$P\left\{\sum_{i=1}^n Y_i < m\right\} = \sum_{k=0}^{m-1} \gamma^k(1-\gamma)^{n-k}\binom{n}{k} \ \leq \ C\beta^n, \tag{3}$$

where β satisfies $1-\gamma < \beta < 1$. The constants β and C depend on m and γ but not on n, and so

$$P\left\{\sum_{i=1}^n Y_i \geq m\right\} \to 1 \quad \text{ as } \ n \to \infty. \tag{4}$$

Let S denote the event that the target cell fires and T_n be the time of firing. If we choose $a = \varepsilon$, then

$$P\{S\} \;\geq\; P\{S \cap \{x_o \leq T_n \leq x_o + \varepsilon\}\} \;=\; P\{\textstyle\sum_{i=1}^{n} Y_i \geq m\},$$

so (i) follows from (4). For all $0 < a \leq \varepsilon$,

$$P\{\{x_o \leq T_n \leq x_o + a\}|S\} = P\{S \cap \{x_o \leq T_n \leq x_o + a\}\}/P\{S\} \quad \to 1 \tag{5}$$

as $n \to \infty$, which proves (ii).

To prove (iii) without assuming compact support, we would need to prove a uniform integrability condition. We will prove such a condition in the case of the main theorem of this section (Theorem 2.3) but omit it here. If we assume that f does have compact support, which is a reasonable hypothesis from a biological perspective, then it is easy to check that $E(T_n|S) \to x_o$ and $E(T_n^2|S) \to x_o^2$ as $n \to \infty$, which gives (iii).

Theorem 2.1 shows, as expected, that if it takes a fixed number of hits in an ε time window to fire the target cell, then one can achieve any improvement in accuracy one wants by taking n large enough. This confirms the general belief in the literature [51, 4, 23] that greater convergence sharpens timing information. However, notice the important hypothesis that m is held fixed. Example 2.1 shows that if m is not fixed, then increasing convergence may make timing worse. Example 2.2 shows that, given a firing mechanism at the target cell, the answer to the question of whether timing gets better of worse depends on f. Thus one should be very careful about drawing general conclusions from simulations, since the results may depend on the form chosen for the noise.

Example 2.1. If $m = n$, $\varepsilon = \infty$, and f is exponential, the standard deviation of the firing times is monotone increasing. This is because the cell will fire when the last (nth) hit arrives. For large n this will be somewhere out in the tail of the distribution. See Table 1 for values. The final entry is computed using the asymptotic behavior of the nth order statistic [44]. Note that this case is not covered by Theorem 2.2 below because $\varepsilon = \infty$. It is also not covered by Theorem 2.3 below because the set $\{x \,|\, F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is empty.

Example 2.2. If $m = n$, $\varepsilon = \infty$, and f is uniform, the standard deviation of the firing times is monotone decreasing. This is because the nth hit out of n will be likely

TABLE 1
Values of σ for different n.

n	Exponential	Uniform
1	1.000 msec	1.000 msec
2	1.118 msec	0.816 msec
3	1.166 msec	0.671 msec
4	1.193 msec	0.566 msec
5	1.210 msec	0.488 msec
6	1.221 msec	0.429 msec
7	1.230 msec	0.382 msec
8	1.236 msec	0.344 msec
9	1.241 msec	0.313 msec
10	1.245 msec	0.287 msec
15	1.257 msec	0.203 msec
20	1.263 msec	0.157 msec
30	1.270 msec	0.108 msec
∞	1.283 msec	0.000 msec

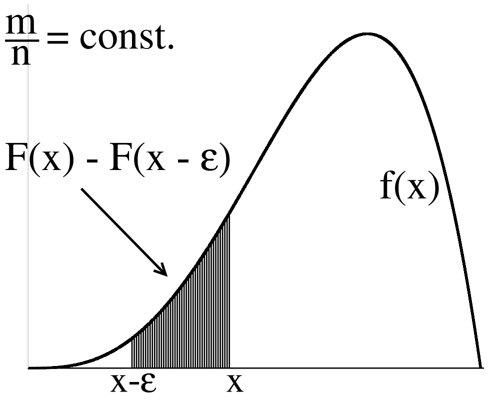


FIG. 2. *An example density to illustrate the derivation of T. T is the smallest value for x so that the area under the density curve between $x - \varepsilon$ and x is at least $\frac{m}{n}$.*

to be close to the right edge of the distribution. Note that this case is not covered by Theorem 2.2 below because $\{x \,|\, F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is not empty. It is also not covered by Theorem 2.3 below because $F(x) - F(x-\varepsilon)$ is not increasing at T.

We now consider the more interesting case where $n \to \infty$ with $\frac{m}{n}$ fixed. To see the intuition, consider the particular f depicted in Figure 2. For any x, we expect that approximately the fraction $F(x) - F(x-\varepsilon)$ of n selections from f should lie in the interval $[x-\varepsilon, x]$. Thus if $F(x) - F(x-\varepsilon) \geq \frac{m}{n}$, we expect that m or more selections will lie in $[x-\varepsilon, x]$, and thus x is certainly a candidate for the firing time of the target cell. Recall that the cell fires the *first* time that it gets m hits in an ε interval. Therefore, we define

$$T \;=\; \inf_x \left\{ x \,|\, F(x) - F(x-\varepsilon) \geq \frac{m}{n} \right\}$$

and expect that for large n the firing time should be close to T. Of course, depending on f, there may be no points in the set $\{x \,|\, F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$.

For the proofs below, we need to introduce some machinery. For each set of n independent selections, $\{X_i\}_{i=1}^n$, from f, we consider the *sample distribution function*

$$F_n(x) \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

where I is the indicator function taking value 1 if $X_i \leq x$ and 0 otherwise. We can now define the random variable for the output firing time T_n in terms of F_n. Choose any $M > T$, and define the random variable T_n to be

$$T_n \;=\; \inf_x \left\{ x \,|\, F_n(x) - F_n(x-\varepsilon) \geq \frac{m}{n} \right\} \tag{6}$$

if the set is nonempty and $T_n = M$ otherwise.

For n large it is known that $F_n(x)$ is a good approximation to $F(x)$. This is expressed via the *Kolmogorov–Smirnov distance*,

$$D_n \;\equiv\; \sup_{-\infty < x < \infty} |F_n(x) - F(x)|.$$

The classical Glivenko–Cantelli lemma (for a proof, see [2]) states that $D_n \to 0$ with probability one.

Theorem 2.2. *Suppose that $0 < \varepsilon < \infty$ and $0 < \frac{m}{n} \leq 1$ are fixed and that the set $\{x \mid F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is empty. Then the probability that the target cell fires converges to zero as $n \to \infty$.*

Proof. Since F is continuous and $\{x \mid F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is empty, there exists $\alpha > 0$ so that

$$F(x) - F(x-\varepsilon) \ \leq \ \frac{m}{n} - \alpha \quad \text{for all } x.$$

Therefore,

$$\begin{aligned} P\{\text{target cell fires}\} &= P\{\text{at least } m \text{ hits in } \ [x-\varepsilon, x] \text{ for some } x\} \\ &= P\{\text{at least } m \text{ hits in } \ (x-\varepsilon, x] \text{ for some } x\} \\ &= P\left\{F_n(x) - F_n(x-\varepsilon) \geq \frac{m}{n} \ \text{ for some } \ x\right\} \\ &\leq P\{F_n(x) - F(x) + F(x-\varepsilon) - F_n(x-\varepsilon) \geq \alpha \ \text{ for some } \ x\} \\ &\leq P\{2D_n \geq \alpha\}, \end{aligned}$$

which converges to zero by the Glivenko–Cantelli lemma. Thus, the probability that the target cell fires goes to zero as $n \to \infty$.

Theorem 2.3. *Suppose that the set $\{x \mid F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is nonempty, and define T as above. Suppose that $F(x) - F(x-\varepsilon)$ is strictly increasing at T and that $0 < \varepsilon \leq \infty$ and the ratio $0 < \frac{m}{n} \leq 1$ are fixed. Then, as $n \to \infty$, the following hold*:

(i) *The probability that the target cell fires* $\longrightarrow 1$.

(ii) $g_{n,m,\varepsilon,f} \longrightarrow \delta_T$; *that is*, $T_n \longrightarrow T$ *in distribution. Further*, T_n *converges to the point mass at T with probability one.*

(iii) $\sigma_{n,m,\varepsilon,f} \longrightarrow 0$.

Proof. To prove (i), note that, by the strict monotonicity at T, the set $\{x \mid F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is nonempty. Note that here we have used only that the set is not empty, but we use the stronger monotonicity hypothesis for the proof of (ii). Thus, there is an $\bar{x}$ so that $F(\bar{x}) - F(\bar{x}-\varepsilon) \equiv \gamma > \frac{m}{n}$. Let Y_i be the random variable that has value 1 if X_i is in $[\bar{x}-\varepsilon, \bar{x}]$ and 0 otherwise. The Y_i are independent Bernoulli random variables with mean γ. Thus,

$$\begin{aligned} P\{Y_1 + \cdots + Y_n < m\} \ &= \ P\left\{\frac{Y_1 + \cdots + Y_n}{n} - \gamma < \frac{m}{n} - \gamma\right\} \\ &\longrightarrow 0 \end{aligned}$$

as $n \to \infty$ by the weak law of large numbers. If there are m or more of the X_i in the particular interval $[\bar{x}-\varepsilon, \bar{x}]$, the target cell fires, and so (i) is proved.

To prove (ii), let $0 < \mu < M - T$ be given, where M is the constant from the definition of T_n (6). We note that the continuity of F and the fact that $F(x) \to 0$ as $x \to -\infty$ imply that there is an $\alpha > 0$ so that

$$\sup_{x < T-\mu} (F(x) - F(x-\varepsilon)) < \frac{m}{n} \ - \ \alpha. \tag{7}$$

Further, by the continuity of F and monotonicity at T, there is a $\beta > 0$ so that

$$\sup_{x < T+\mu} (F(x) - F(x-\varepsilon)) > \frac{m}{n} \ + \ \beta. \tag{8}$$

We will show that $P\{|T_n - T| > \mu\} \to 0$ as $n \to \infty$. First,

$$\begin{aligned} P\{T_n < T - \mu\} &= P\left\{\exists x < T - \mu \,|\, F_n(x) - F_n(x-\varepsilon) \geq \frac{m}{n}\right\} \\ &\leq P\{\exists x < T - \mu \,|\, F_n(x) - F_n(x-\varepsilon) - (F(x) - F(x-\varepsilon)) \geq \alpha\} \\ &\leq P\{2D_n > \alpha\}. \end{aligned}$$

Similarly,

$$\begin{aligned} P\{T_n > T + \mu\} &= P\left\{F_n(x) - F_n(x-\varepsilon) < \frac{m}{n} \ \ \forall x < T + \mu\right\} \\ &\leq P\{F(x) - F(x-\varepsilon) - F_n(x) + F_n(x-\varepsilon) > \beta \ \ \forall x < T + \mu\} \\ &\leq P\{2D_n > \beta\}. \end{aligned}$$

In each case the probability converges to zero by the Glivenko–Cantelli lemma, which proves (ii). Note that since D_n converges with probability one, so does T_n. This in turn implies convergence in distribution.

To prove (iii) we need to show that the sequences $\{T_n\}$ and $\{T_n^2\}$ are uniformly integrable (for a proof, see [44]), i.e., that $\lim_{c\to\infty} \sup_n \int_c^\infty x g_{n,m,\varepsilon,f} dx = 0$ and that $\lim_{c\to\infty} \sup_n \int_c^\infty x^2 g_{n,m,\varepsilon,f} dx = 0$. We begin by bounding the density of T_n, $g_{n,m,\varepsilon,f}$, which we will abbreviate $g_n(x)$. In [30] we derived an explicit integral formula for $g_n(x)$. It is more convenient to write this expression in terms of the ordered inputs known as the order statistics. Let Y_i be the ith order statistic, i.e., the random variable which is the ith smallest of the X_i's. Let $f_{\{Y_i|T_n=Y_i\}}(x)$ denote the conditional density of Y_i given that $T_n = Y_i$, and let P_i be the probability that $T_n = Y_i$. The density $g_n(x)$ of T_n is the normalized sum from m to n of these conditional densities:

$$g_n(x) \;=\; \frac{\sum_{i=m}^n f_{\{Y_i|T_n=Y_i\}}(x) P_i}{P(\text{success})}. \tag{9}$$

Using the joint density of the Y_i's, we can compute $f_{\{Y_i|T_n=Y_i\}}(x)$ by integrating over the appropriate event:

$$\begin{aligned} &f_{\{Y_i|T_n=Y_i\}}(x) \\ &= \frac{1}{P_i} n! f(x) \int_{\Omega_3} \prod_{j=i+1}^{n} f(y_j) \int_{\Omega_2} \prod_{j=i-m+1}^{i-1} f(y_j) \int_{\Omega_1} \prod_{j=1}^{j=i-m} f(y_j) \prod_{j=1}^{i-1} dy_j \prod_{j=i+1}^{n} dy_j, \end{aligned} \tag{10}$$

where Ω_1, Ω_2, and Ω_3 are the sets

$$\begin{aligned} \Omega_1 &= \{y_1 < \cdots < y_{i-m} \ \text{ and } y_k < y_{k+m-1} - \varepsilon \text{ for } k = 1, \ldots, i-m\}, \\ \Omega_2 &= \{x - \varepsilon < y_{i-m+1} < \cdots < y_{i-1} < x\}, \\ \Omega_3 &= \{x < y_{i+1} < \cdots < y_n\}. \end{aligned}$$

The sets Ω_1, Ω_2, and Ω_3 correspond to the statements that the first $i-1$ arriving hits do not fire the cell, that the ith does, and that the remaining times can be anything.

We can attain a bound on the integral by replacing Ω_1 with the larger set

$$\Omega_1 = \{y_1 < \cdots < y_{i-m} \ \text{ and } \ y_{i-m} < x - \varepsilon\}.$$

Next we can integrate explicitly to obtain the bound on the conditional density

$$f_{\{Y_i|T_n=Y_i\}}(x) \leq \frac{1}{P_i} n! f(x) \frac{(1-F(x))^{n-i}}{(n-i)!} \frac{(F(x) - F(x-\varepsilon))^{m-1}}{(m-1)!} \frac{F(x-\varepsilon)^{i-m}}{(i-m)!}.$$

Summing over i in (9) gives the bound

$$g_n(x) \leq \frac{1}{P(\text{success})} n! \frac{(1-(F(x)-F(x-\varepsilon)))^{n-m}}{(n-m)!} \frac{(F(x)-F(x-\varepsilon))^{m-1}}{(m-1)!} f(x).$$

For large c, $F(x)-F(x-\varepsilon)$ goes to zero. We can make a straightforward calculation using Stirling's formula to show that $n!\frac{(1-(F(x)-F(x-\varepsilon)))^{n-m}}{(n-m)!}\frac{(F(x)-F(x-\varepsilon))^{m-1}}{(m-1)!}$ is uniformly bounded independent of n for $F(x)-F(x-\varepsilon)$ sufficiently small. Specifically if c is large enough so that $F(x)-F(x-\varepsilon) < \frac{a^a}{(a-1)^{a-1}}$ for all $x > c$, where a is the ratio n/m, then there is a constant A so that $g_n(x) \leq \frac{A}{P(\text{success})} f(x)$ for all $x > c$. Since f has a finite mean and standard deviation, this bound implies that $\{T_n\}$ and $\{T_n^2\}$ are both uniformly integrable, and so $E(T_n) \to T$ and $E(T_n^2) \to T^2$, which proves (iii).

Theorem 2.3 shows that if $F(x) - F(x-\varepsilon)$ crosses $\frac{m}{n}$ the first time it reaches $\frac{m}{n}$, then the firing time will converge to the point mass at this time as $n \to \infty$. The following two examples show that while most sets of parameters will be covered in the above theorems, we have not addressed what will happen if $F(x) - F(x-\varepsilon)$ is not increasing at T but rather reaches $\frac{m}{n}$ and is constant for some time or reaches $\frac{m}{n}$ and immediately drops back down. This situation is unlikely in the biological context but is of mathematical interest. Example 2.3 shows that if f is exponential, then either Theorem 2.2 or 2.3 will apply unless $F(x_o+\varepsilon)$ is exactly $\frac{m}{n}$. In Example 2.4 we discuss various cases in which the hypotheses of Theorem 2.3 are, or are not, satisfied.

Example 2.3. If f is exponential, then $F(x)-F(x-\varepsilon)$ will be monotone increasing from x_o to $x_o+\varepsilon$ and monotone decreasing for $x > x_o+\varepsilon$. If the value of $F(x)-F(x-\varepsilon)$ at its peak, namely $F(x_o+\varepsilon)$, is greater than $\frac{m}{n}$, then T will be less than ε and the hypotheses of Theorem 2.3 will be satisfied, so the standard deviation will go to zero (see Figure 3(A)). If, on the other hand, the value at the peak is less than $\frac{m}{n}$, then the set $\{x \,|\, F(x) - F(x-\varepsilon) \geq \frac{m}{n}\}$ is empty and the hypotheses of Theorem 2.2 will be satisfied, so the probability of firing will go to zero. It is of mathematical interest to study what will happen in the borderline case where $F(x_o+\varepsilon)$ is exactly $\frac{m}{n}$.

Example 2.4. Let $x_o < x_1 < y_o < y_1$ and suppose that the support of the density f consists of the two intervals $[x_o, x_1]$ and $[y_o, y_1]$. In addition, suppose that $x_1 - x_o < \varepsilon$, $y_1 - y_o < \varepsilon$, and $y_o - x_1 > \varepsilon$, so the intervals are small and well separated compared to ε. Let $p = \int_{x_o}^{x_1} f(x)\,dx$ and $q = \int_{y_o}^{y_1} f(x)\,dx$. There are several cases to consider. If $p > \frac{m}{n}$, then $x_o < T < x_1$ and the "strictly increasing" hypothesis holds, so the conclusions of Theorem 2.3 hold. If $p < \frac{m}{n}$ and $q < \frac{m}{n}$, then the probability of firing goes to zero as $n \to \infty$ with $\frac{m}{n}$ fixed by Theorem 2.2. If $p < \frac{m}{n}$ and $q > \frac{m}{n}$, then $y_o < T < y_1$ and again the "strictly increasing" hypothesis holds, so the conclusions of Theorem 2.3 hold. Finally, suppose that $\frac{m}{n} = \frac{1}{2}$ and that $p = q = \frac{1}{2}$. Then $T = x_1$ but the "strictly increasing" hypothesis does not hold. The number of hits in the first region is given by the binomial $B(n,p)$. Thus the probability of firing in this first interval, $P(B(n,p) \geq m)$, converges to $\frac{1}{2}$ as $n \to \infty$ with $\frac{m}{n}$ fixed. A straightforward argument shows that the conditional density (conditioned on firing in the first interval) converges to δ_{x_1}. The same arguments show that if the neuron does not fire in the first interval, then it has probability $\frac{1}{2}$ of firing in the second interval, and the conditional density (conditioned on firing in the second interval) converges to δ_{y_1}. Therefore, as $n \to \infty$ with $\frac{m}{n}$ fixed, the density (conditioned on firing) converges to $\frac{2}{3}\delta_{x_1} + \frac{1}{3}\delta_{y_1}$. Thus, if the "strictly increasing at T" hypothesis does not hold, the conclusions of Theorem 2.3 may not hold.

3. The asymptotic form. We will now prove the asymptotic normality of T_n. We will consider the case where, in addition to the left edge hypothesis (2) we assume that $T < x_0 + \varepsilon$. This means that there is more than m/n probability in the interval $(x_0, x_0 + \varepsilon)$, i.e., $F(x_o + \varepsilon) - F(x_o) > m/n$, and therefore that $F(x - \varepsilon) = 0$ for all $x < x_o + \varepsilon$.

THEOREM 3.1. *Suppose that $T < x_0 + \varepsilon$; then T_n is asymptotically normal with mean T and standard deviation*

$$\sigma_c = \frac{\left(\frac{m}{n}\left(1 - \frac{m}{n}\right)\right)^{1/2}}{f(T)n^{1/2}}. \tag{11}$$

We call the standard deviation σ_c because it is a result of the convergence studied in sections 1 and 2.

Proof. Fix t and let

$$G_n(t) = P\left(\frac{T_n - T}{\sigma_n} \le t\right).$$

We wish to show that $G_n(t) \to \Phi(t)$, where Φ is the cumulative distribution for the standard normal. We begin by rewriting $G_n(t)$ using the definition of T_n given in (6):

$$\begin{aligned} G_n(t) &= P(T_n \le T + t\sigma_n) \\ &= P\left(\exists x \le T + t\sigma_n | F_n(x) - F_n(x - \varepsilon) \ge \frac{m}{n}\right). \end{aligned}$$

Since $\sigma_n \to 0$ as $n \to \infty$ and $T < x_0 + \varepsilon$, there is an $\bar{n}$ such that for all $n \ge \bar{n}$, $T + t\sigma_n < x_0 + \varepsilon$. Therefore if $n \ge \bar{n}$ and $x \le T + t\sigma_n$, $F(x - \varepsilon) = 0$. Note that, by the definition of F_n, if $F(x - \varepsilon)$, then the probability of a hit before $x - \varepsilon$ is zero and $F_n(x - \varepsilon)$ is also zero for all n. So, for $n \ge \bar{n}$,

$$\begin{aligned} G_n(t) &= P\left(\exists x \le T + t\sigma_n | F_n(x) \ge \frac{m}{n}\right) \\ &= P\left(F_n(T + t\sigma_n) \ge \frac{m}{n}\right), \end{aligned}$$

where we have used the monotonicity of F_n. Notice that since $F(T - \varepsilon) = 0$, the value T is just the value of the $\frac{m}{n}$th quantile, which we will denote ξ. The $\frac{m}{n}$th quantile is defined by $\xi = \inf\{x : F(x) \ge \frac{m}{n}\}$. The proof rests upon the asymptotic normality of the sample $\frac{m}{n}$th quantile, ξ_n, defined by $\xi_n = \inf\{x : F_n(x) \ge \frac{m}{n}\}$. ξ_n is asymptotically normal with mean ξ and standard deviation σ_n [44]. Now we can write G_n in terms of the quantiles:

$$G_n(t) = P(\xi_n \le \xi + t\sigma_n).$$

Therefore the asymptotic normality of the sample quantile implies the asymptotic normality of T_n.

It is the hypothesis that $T < x_0 + \varepsilon$ that makes the proof of Theorem 3.1 easy by reducing the question to the asymptotic behavior of quantiles. Intuitively, the hypothesis means that there is a lot of probability close to the initial point x_0. Example 3.1 shows that this hypothesis is biologically reasonable. Example 3.2 shows what can happen if this hypothesis is violated and gives a conjecture for the general case.

Example 3.1. For AN neurons, the density f looks like a (translated, smoothed) exponential [25, 51, 49] with standard deviation approximately 1 msec. In Figure 3(A)

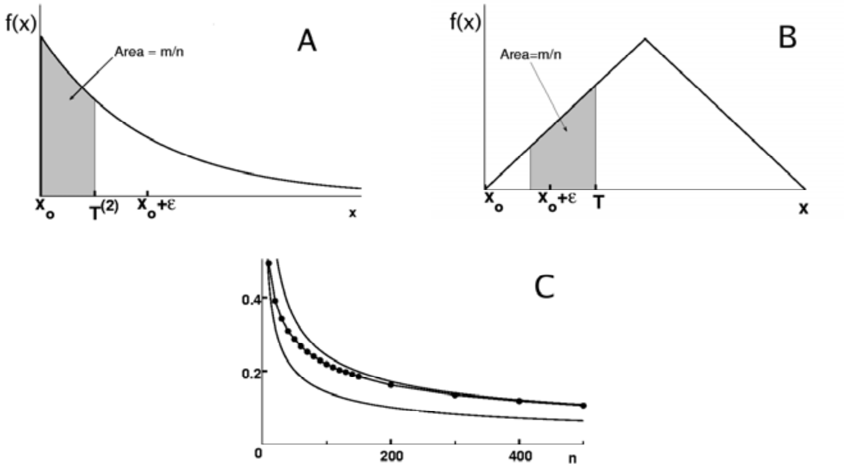


Fig. 3. *Examples illustrating the importance of the hypothesis* $T < x_o + \varepsilon$ *in Theorem* 3.1. *Panel* (A) *shows the density for the exponential as in Example* 3.1. *The shaded portion has area* $\frac{m}{n}$, *and one can see that* $T^{(2)} = x_o + \ln\frac{5}{3} < x_o + \varepsilon$ *so that Example* 3.1 *satisfies the hypotheses of Theorem* 3.1. *Similarly, the shaded portion in panel* (B) *has area* $\frac{m}{n}$, *but in this case* $T = 1.7 - \sqrt{6} > x_o + \varepsilon$ *so that Theorem* 3.1 *does not apply. Panel* (C) *compares results of Monte Carlo simulations (data are the dots on the middle curve) to the predictions of Theorem* 3.1 *(bottom curve,* σ_c*). We see that the theorem does not apply, but that our conjecture (top curve,* $\bar{\sigma}_c$*) for this more general case is supported.*

we show the exponential distribution (starting at x_0) with standard deviation 1 msec. It is easy to check that $T = x_0 + \ln\frac{n}{n-m}$. Consider three cases $\frac{m_1}{n_1} = .2$, $\frac{m_2}{n_2} = .4$, and $\frac{m_3}{n_3} = .5$, which will have corresponding $T^{(1)} = x_0 + \ln\frac{5}{4}$, $T^{(2)} = x_0 + \ln\frac{5}{3}$, and $T^{(3)} = x_0 + \ln 2$. Thus, if $\varepsilon = 1$ msec, which is reasonable for octopus cells [15, 32] and many other neurons, we will have $T^{(i)} < x_0 + \varepsilon$ in all three cases, so Theorem 3.1 applies.

Example 3.2. On the other hand, suppose that f is the piecewise linear "hat" distribution (see Figure 3(B)). In order to have standard deviation 1 msec, the density is supported on the interval $[-\sqrt{6}\sqrt{6}]$ (so $x_o = \sqrt{6}$). In all of the cases above $T^{(i)} > x_o + \varepsilon$, and so Theorem 3.1 does not apply. For example if $\frac{m}{n} = .2$, then $T = -\sqrt{6} + 1.7 \approx -0.75$. In this case the standard deviation does not converge to the value σ_c given in Theorem 3.1 but instead to another higher value. We conjecture that in this case T_n will be asymptotically normal with mean T and standard deviation

$$\bar{\sigma}_c = \frac{\left(\frac{m}{n}\left(1 - \frac{m}{n}\right)\right)^{1/2}}{(f(T) - f(T - \varepsilon))n^{1/2}}.$$

Figure 3(C) shows the values for σ_c (bottom curve), $\bar{\sigma}_c$ (top curve), and the standard deviation computed using Monte Carlo simulations as in [35] (middle curve; each dot was computed using 100,000 trials). We can see that for large n the values do not approach σ_c but rather $\bar{\sigma}_c$, supporting our conjecture.

4. Applications to octopus cells. The latency of a neuron in the auditory system is the length of time between the start of a sound and the time of the first action potential produced by the neuron. In mammals, AN neurons, which provide the input to the auditory brainstem, have latencies in the range 2 to 8 msec, with standard deviations of approximately 1 msec under repeated trials [25]. The auditory system must use group properties of these highly variable inputs to extract sharp timing information so that the animal can make time distinctions in the low microsecond range. According to Oertel et al. [32], much of this processing is done by octopus cells in the cochlear nucleus that "detect coincident firing within populations of auditory nerve fibers and convey acoustic information in precisely timed action potentials." It is estimated that octopus cells receive synapses from roughly 60 to 100 AN neurons (i.e., $60 \leq n \leq 100$) and require that 20% to 50% of these synapses be activated by incoming action potentials within 1 msec in order fire an action potential (i.e., $\varepsilon =$ 1 msec and $0.2 \leq \frac{m}{n} \leq 0.5$) [32, 15]. It is therefore of interest to test whether the predictions of the theorems in this paper are consistent with the observed in vivo and in vitro behavior of octopus cells. For some parameter choices we can also make more specific predictions for optimal values of m and n.

The histograms of latencies in AN neurons are quite variable but look roughly like smoothed exponential distributions. Thus, we shall assume that f is exponential with standard deviation 1 msec, and it follows that $T = x_o + \log \frac{n}{n-m}$. If $\varepsilon \geq T$, which holds for all cases considered below (see Example 3.1 above), then $F(T-\varepsilon) = 0$, and the asymptotic formula (11) has the simple form

$$\sigma_c^2 = \frac{\frac{m}{n}}{(1-\frac{m}{n})n}. \tag{12}$$

If we evaluate σ_c for n and m in the physiological ranges given above, we obtain values in the range 0.05 msec to 0.13 msec (Table 2). Oertel et al. report that the standard deviations of latencies of octopus cells in response to sounds are approximately 0.1 msec [32], so the formula (11) certainly predicts the order of magnitude improvement of timing seen in vivo in octopus cells.

TABLE 2
Values of σ_c.

n	m	σ_c (msec)
100	20	0.05
100	33	0.07
100	50	0.10
60	30	0.13

We can use (12) to explore a variety of questions about the physiology of octopus cells. First we ask why n isn't larger than the range 60–100. σ_c is the standard deviation in the latency of the octopus cell due to the variation in firing times of the inputs. However, there are other sources of variation. The AN neurons that synapse on the octopus cell may have somewhat different axonal lengths and diameters, both of which will affect arrival times. Second, due to such factors as the diffusion of neurotransmitter across the synaptic cleft, the finite number of postsynaptic receptors, and the variation in the local membrane chemistry, the integration of synaptic inputs by the octopus cell will have variation under repeated trials even if the timing of the inputs is the same. Assuming that these other factors are independent of the noise in the firing times of the inputs, and denoting the corresponding standard deviation by σ_{other}, we have $\sigma_o^2 = \sigma_c^2 + \sigma_{other}^2$. Fortunately, the experiments in [15], where shocks are applied to the nerve root, give a good estimate, $\sigma_{other} = 0.05$ msec. Figure 4 shows the behavior of σ_o as a function of n in two cases, $\frac{m}{n} = \frac{1}{2}$ and $\frac{m}{n} = \frac{1}{5}$, that are

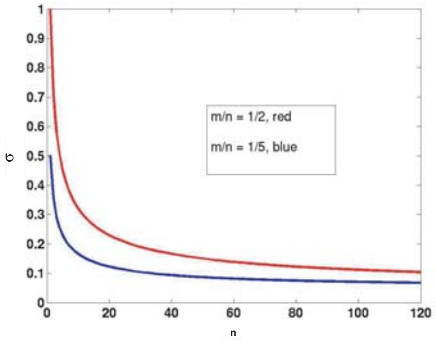


FIG. 4. *Predicted values of the standard deviation of the latency of octopus cells, σ_o, for different values of n.*

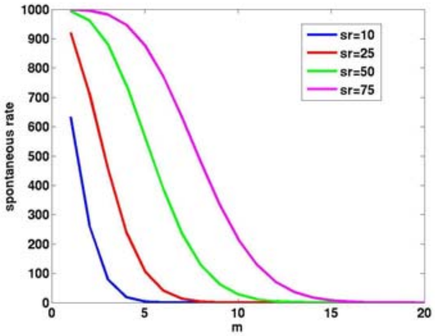


FIG. 5. *Predicted values of the spontaneous rate of the octopus cell as a function of m with $n = 100$ and four different assumptions about the spontaneous rate, r, of AN neurons.*

the expected extremes for the ratio $\frac{m}{n}$. In both cases there is not much extra decrease in σ_o after $n = 60$ and very little after $n = 100$.

Many AN neurons have high or very high spontaneous rates, even ranging as high as 100 spikes/sec [25, 19]. If n is high and m is low, then many of the successful firings of the octopus cell will be spontaneous, i.e., unrelated to input. However, it is known that octopus cells have essentially no spontaneous rate [37, 38, 46]. Thus, it is a natural question to ask how large m must be so that the spontaneous rate of the octopus cell in our model is 1 spike/sec or less. Assume $\varepsilon = 1$ msec, and suppose that each incoming AN neuron has a spontaneous rate of r spikes/msec. Then, the probability that any particular AN neuron delivers a spike within a 1 msec interval is approximately r. Assuming that the AN neurons are independent, the probability of m or more incoming spikes within the 1 msec interval is approximately

$$Prob\{\#\text{incoming hits} \geq m\} = \sum_{k=m}^{n} \binom{n}{k} r^k (1-r)^{n-k}, \tag{13}$$

and thus the spontaneous firing rate of the octopus cell in spikes/sec, denoted by $SR(r, n, m)$, will be approximately 1000 times the probability in (13). Figure 5 shows the graphs of $SR(r, n, m)$ as functions of m for $n = 100$ and for four different choices of r. If the incoming AN fibers have spontaneous rates of $r = .075$ spikes/msec

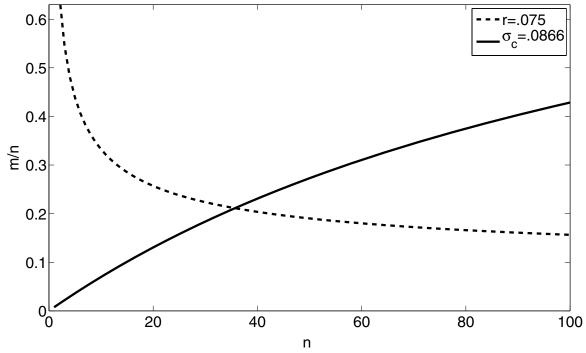


FIG. 6. *The region of the* $(\frac{m}{n}, n)$ *plane for which* $\sigma_o \leq 0.1$ *and the octopus cell has a spontaneous firing rate* ≤ 1 *spike/sec is the region below the solid curve and above the dashed curve.*

(75 spikes/sec), $SR(.075, 100, m)$ does not go below 1 spike/sec until $m = 18$. Thus the model, with $n = 100$, predicts that m is 18 or higher, which corresponds well with the estimates of experimentalists [15, 32, 13].

We can also allow both $\frac{m}{n}$ and n to be free and ask what is the region in the $(n, \frac{m}{n})$ plane that gives the observed physiological behavior. First, we require that the standard deviation of the latency of the octopus cell satisfy $\sigma_o \leq 0.1$ msec. Since $\sigma_o^2 = \sigma_c^2 + \sigma_{other}^2$, this is equivalent to the requirement that $\sigma_c^2 \leq .1^2 - .05^2$. Equation (12) can be rearranged to give a bound on $\frac{m}{n}$ in terms of this maximum allowable standard deviation:

$$\frac{m}{n} \leq \frac{(\sigma_c^2)n}{1 + (\sigma_c^2)n}. \tag{14}$$

Second, we require that the spontaneous firing rate of the octopus cell be less than the maximum value $r_{max} = .001$ spike/msec or 1 spike/sec. Using (13) and the normal approximation to the binomial gives

$$\frac{m}{n} \geq \Phi^{-1}(1 - r_{max})\sqrt{\frac{r(1-r)}{n}} + r, \tag{15}$$

where Φ is the cumulative distribution function of the standard normal and r is the spontaneous rate of the incoming neurons. The points below the solid curve in Figure 6 satisfy (14), and the points above the dashed curve satisfy (15) in the case $r = .075$ spikes/msec.

In Figure 6 we assumed that the standard deviation of f is 1 msec and that the spontaneous rate of the AN neurons $r = 75$spikes/sec. For other assumptions the curves are somewhat different. For example, instead of choosing $\sigma_c = .0866$, we could recognize that it is not particularly beneficial to require that σ_c be smaller than σ_{other}. In this case we require that σ_c be roughly the same as σ_{other} so that both are equal to .05 msec and repeat the above calculations (shown in Figure 7), predicting a much smaller region of possible values for m and n. In this case we require $n \geq 80$ and that $\frac{m}{n}$ be between 18% and 20%. This is within the experimental estimates but suggests a much smaller range of optimal values. From this calculation we can predict that ideally n should be near the high end of its range, close to 100, and that m should be near the low end of its range, close to 20.

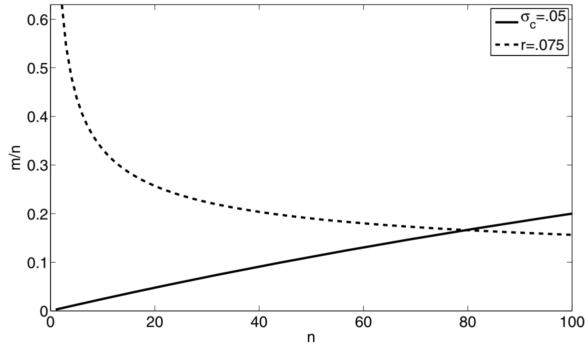


FIG. 7. *The region of the* $(\frac{m}{n}, n)$ *plane for which* $\sigma_c \leq 0.05$ *and the octopus cell has a spontaneous firing rate* ≤ 1 *spike/sec is the region below the solid curve and above the dashed curve. In both Figures* 6 *and* 7*, we assume that the standard deviation of* f *is* 1 *msec and that the spontaneous rate of the AN neurons* $r = 75$ *spikes/sec.*

5. Discussion. The model given in Figure 1 was formulated to allow a mathematical investigation of how convergence (the number of incoming neurons, n) and the number of hits required to make the target cell fire, m, affect the sharpening of timing information when the firing times of the incoming neurons are noisy. The main theorem (Theorem 2.3) shows that if $n \to \infty$ and $m \to \infty$ with $\frac{m}{n}$ fixed, then the standard deviation of the time of firing of the target cell goes to zero. The physiological significance of the result is that timing can be sharpened by taking both n large and m large (to avoid spontaneous firing of the target). That there should be a theorem like this was suggested by the approximate calculations of Burkitt and Clark [4] and the numerical simulations of Kalluri and Delgutte [23]. In section 3 we derived approximate formulas for $g_{n,m,\varepsilon,f}$ and $\sigma_{n,m,\varepsilon,f}$ near the limit. In section 4, we used the asymptotic formula for $\sigma_{n,m,\varepsilon,f}$ to study octopus cells of the mammalian cochlear nucleus and saw that the predictions of the mathematical model correspond quite well to experimental observations.

We hope that the theorems proved in this paper can be a first step in proving theorems about more complicated and difficult neurophysiological questions. One such question is the improved phase locking of CN neurons compared to the phase locking in the AN [19, 22, 43, 4, 23], which is universally believed to occur because of convergence of many AN fibers on CN target cells. The mathematical situation here is more complicated, since typically one models the firing pattern in individual AN fibers by a Poisson process whose parameter $\lambda(t)$ depends on the sound; for example, $\lambda(t)$ would be periodic for a pure tone. The quantity of interest is the distribution of spike times of the target cell modulo the nearest multiple of the period. For some cells the target cell may fire at first hit, while for other cells many subthreshold hits in a small time window may be necessary for firing. Because of the background noise caused by the high spontaneous rates of many AN fibers, this may be an excellent use of the theory of stochastic resonance [17, 47]. To prove such theorems one would need to represent the noise as stochastic processes rather than making the simple approximations that we have used in section 4. Another such question is how synchronous firing of large groups of neurons in the CNS is created and maintained. Such firing has been proposed as central to "binding" mechanisms in the visual system [12, 26, 28], the improvement of coordination of motor systems in the cerebellum [20], and the creation of the γ

rhythm [3]. The approximation theorem in section 3 is a natural starting point for studying synfire chains [1, 11, 36, 18] that have both noise and high convergence from level to level. Many of the models for these systems involve inhibition, so an important step would be the extension of the results in [35, 30] and this paper to include inhibitory neurons.

In applying our model to octopus cells we have simplified the biological situation in several ways. First, AN neurons synapse serially on the large dendrites of octopus cells, not directly on the cell body as in our model. Oertel has shown [32] that the AN neurons that carry higher frequency sounds (they fire on average earlier) synapse further out on the dendrite, and the AN neurons that carry lower frequency sounds (they fire on average later) synapse closer to the cell body. Golding, Ferragamo, and Oertel [16] conclude that this arrangement on the dendrite, as well as the thickness of the dendrite and special channel properties, insure that the influence of each AN neuron arrives (on average) at the cell body at the same time, which justifies our assumption that all the AN neurons synapse directly on the cell body.

A much more serious simplification is that we have ignored the detailed biophysics of synapses and the postsynaptic membrane. All the biophysics is contained in the two parameters, ε, the time window, and m, the number of hits required in that time window to fire the target cell. We believe that the results in section 4 show conclusively that our model, with these two simple parameters, explains why octopus cells improve the standard deviation of timing by one order of magnitude and why octopus cells have no spontaneous rates. Of course, the *values* of these two parameters arise from the detailed biophysics of the synapses and postsynaptic membrane.

Finally, it is reasonable to ask whether octopus cells, or indeed any neurons, have sharp time windows as we assume in our model. Ferragamo and Oertel [13] have conducted a detailed study of the potential of the postsynaptic membrane of octopus cells. They showed that it is the rate of rise of the potential (dependent on the rate of arrival of incoming spikes) that determines whether the octopus cell fires. This is exactly what one would expect if the octopus cell had a sharp time window. If the rate of rise is high enough, then there will be enough incoming spikes in the time window, and if the rate of rise is too slow, then there will not be enough incoming spikes in the time window. More generally, we have studied a number of frequently used nonlinear models for the biophysics of postsynaptic membranes and have shown that, in reasonable parameter ranges, they have quite sharp time windows. These results will appear in a subsequent publication [31].

Acknowledgments. The authors are very grateful to Professor Donata Oertel, who raised many of the questions discussed in section 4. We would also like to thank Matthew Moehlmann, who wrote the code for the Monte Carlo simulation in Example 3.2.

REFERENCES

[1] M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex*, Cambridge University Press, Cambridge, UK, 1991.

[2] P. Billingsley, *Probability and Measure*, Wiley, New York, 1975.

[3] C. E. Borgers and N. Kopell, *Synchronization in networks of excitatory and inhibitory neurons with sparse, random connectivity*, Neural Comput., 15 (2003), pp. 509–538.

[4] A. Burkitt and G. Clark, *Analysis of integrate-and-fire neurons: Synchronization of synaptic input and spike output*, Neural Comput., 11 (1999), pp. 871–901.

[5] A. Burkitt and G. Clark, *Synchronization of the neural response to noisy periodic synaptic input*, Neural Comput., 13 (2001), pp. 2639–2672.

[6] Y. Cai, E. Walsh, and J. McGee, *Mechanisms of onset responses in octopus cells of the cochlear nucleus: Implications of a model*, J. Neurophysiol., 78 (1997), pp. 872–883.

[7] Y. Cai, J. McGee, and E. Walsh, *Contributions of ion conductances to the onset responses of octopus cells in the ventral cochlear nucleus: Simulation results*, J. Neurophysiol., 83 (2000), pp. 301–314.

[8] H. Colburn, *Theory of binaural interaction based on auditory-nerve data.* I. *General strategy and preliminary results on interaural discrimination*, J. Acoust. Soc. Amer., 54 (1973), pp. 1458–1470.

[9] H. Colburn, *Theory of binaural interaction based on auditory-nerve data.* II. *Detection of tones in noise*, J. Acoust. Soc. Amer., 61 (1977), pp. 525–533.

[10] E. Covey and J. H. Casseday, *The monaural nuclei of the lateral lemniscus of the echolocating bat: Parallel pathways for analyzing the temporal features of sounds*, J. Neurosci., 11 (1991), pp. 3456–3470.

[11] M. Diesmann, M.-O. Gewaltig, S. Rotter, and A. Aersted, *Stable propagation of synchronous spiking in cortical neural networks*, Nature, 402 (1999), pp. 529–533.

[12] A. K. Engel, P. Konig, and W. Singer, *Direct physiological evidence for scene segmentation by temporal coding*, Proc. Natl. Acad. Sci. USA, 88 (1991), pp. 9136–9140.

[13] M. Ferragamo and D. Oertel, *Octopus cells of the mammalian cochlear nucleus sense the dynamic properties of synaptic excitation*, J. Neurophysiol., 87 (2003), pp. 2262–2270.

[14] J. M. Goldberg and P. B. Brown, *Response of binaural neurons of dog superior olivary complex to dichotic tone stimuli: Some physiological mechanisms of sound localization*, J. Neurophysiol., 32 (1969), pp. 613–636.

[15] N. Golding, D. Robertson, and D. Oertel, *Recordings from slices indicate that octopus cells of the cochlear nucleus detect coincident firing of auditory nerve fibers with temporal precision*, J. Neurosci., 15 (1995), pp. 3138–3153.

[16] N. Golding, M. Ferragamo, and D. Oertel, *Role of intrinsic conductances underlying responses to transients in octopus cells of the cochlear nucleus*, J. Neurosci., 19 (1999), pp. 2897–2905.

[17] P. E. Greenwood, U. U. Muller, L. M. Ward, and W. Wefelmeyer, *Statistical analysis of stochastic resonance in a thresholded detector*, Aust. J. Stat., 32 (2003), pp. 49–70.

[18] G. Hayon, M. Abeles, and D. Lehmann, *A model for representing the dynamics of a system of synfire chains*, J. Comput. Neurosci., 18 (2005), pp. 41–53.

[19] D. R. F. Irvine, *The Auditory Brainstem*, Springer-Verlag, New York, 1986.

[20] R. Ivry, *Cerebellar timing systems*, Internat. Rev. Neurobiol., 41 (1997), pp. 555–573.

[21] L. A. Jeffress, *A place theory of sound localization*, J. Comput. Psychol., 41 (1947), pp. 35–39.

[22] P. X. Joris, L. H. Carney, P. H. Smith, and T. Yin, *Enhancement of neural synchronization in the anteroventral cochlear nucleus.* I. *Responses to tones at the characteristic frequency*, J. Acoust. Soc. Amer., 71 (1994), pp. 1022–1036.

[23] S. Kalluri and B. Delgutte, *Mathematical models of cochlear onset neurons:* I. *Point neuron with many synaptic inputs*, J. Comput. Neurosci., 14 (2003), pp. 71–90.

[24] S. Kalluri and B. Delgutte, *Mathematical models of cochlear onset neurons:* II. *Model with dynamic spike-blocking state*, J. Comput. Neurosci., 14 (2003), pp. 91–110.

[25] N. Kiang, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, MIT Press, Cambridge, MA, 1996.

[26] P. Konig, A. K. Engel, and W. Singer, *Integrator or coincidence detector? The role of cortical neurons revisited*, Trends. Neurosci., 19 (1996), pp. 130–137.

[27] P. Marsalek, C. Koch, and J. Maunsell, *On the relationship between synaptic input and spike output jitter in individual neurons*, Proc. Natl. Acad. Sci. USA, (1997), pp. 735–740.

[28] M. S. Matell and W. H. Meck, *Neurophysiological mechanisms of interval timing behavior*, BioEssays, 22 (2000), pp. 94–103.

[29] C. Mitchell, *Mathematical Properties of Time-Windowing in Neural Systems*, Ph.D. thesis, Department of Mathematics, Duke University, Durham, NC, 2003.

[30] C. Mitchell, *Precision of neural timing: The small ε limit*, J. Math. Anal. Appl., 309 (2005), pp. 567–582.

[31] C. Mitchell and M. Reed, *Mathematical Properties of Time-Windowing in Neural Systems, Emergent Time Windows in Nonlinear Neural Models*, in preparation, 2007.

[32] D. Oertel, R. Bal, S. Gardner, P. Smith, and P. Joris, *Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 11773–11779.

[33] G. D. Pollak, *Some comments on the perception of phase and nanosecond time disparities by echolocating bats*, J. Comput. Physiol. A, 172 (1993), pp. 523–531.

[34] L. Rayleigh, *On our perception of sound direction*, Philos. Mag., 13 (1907), pp. 214–232.

[35] M. Reed, J. Blum, and C. Mitchell, *Precision of neural timing: Effects of convergence and time-windowing*, J. Comput. Neurosci., 13 (2002), pp. 35–47.

[36] A. Reyes, *Synchrony-dependent propagation of firing rate in iteratively constructed networks* in vitro, Nature Neurosci., 6 (2003), pp. 593–599.

[37] W. S. Rhode, D. Oertel, and P. H. Smith, *Physiological response properties of cells labelled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus*, J. Comput. Neurol., 213 (1983), pp. 448–463.

[38] W. S. Rhode and P. H. Smith, *Encoding timing and intensity in the ventral cochlear nucleus of the cat*, J. Neurophysiol., 56 (1986), pp. 261–286.

[39] J. S. Rothman and P. B. Manis, *Differential expression of three distinct potassium currents in the ventral cochlear nucleus*, J. Neurophysiol., 89 (2003), pp. 3070–3082.

[40] J. S. Rothman and P. B. Manis, *Kinetic analyses of three distinct potassium conductances in ventral cochlear nucleus neurons*, J. Neurophysiol., 89 (2003), pp. 3083–3096.

[41] J. S. Rothman and P. B. Manis, *The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons*, J. Neurophysiol., 89 (2003), pp. 3097–3113.

[42] J. Rothman, E. Young, and P. Manis, *Convergence of auditory nerve fibers onto bushy cells in the ventral cochlear nucleus: Implications of a computational model*, J. Neurophysiol., 70 (1993), pp. 2562–2582.

[43] J. Rothman and E. Young, *Enhancement of neural synchronization in computational models of ventral cochlear nucleus bushy cells*, J. Auditory Neurosci., 2 (1996), pp. 47–62.

[44] R. Serfling, *Approximation Theorems of Mathematics Statistics*, Wiley, New York, 1976.

[45] J. A. Simmons, M. Ferragamo, C. F. Moss, S. B. Stevenson, and R. A. Altes, *Discrimination of jittered sonar echoes by the echolocating bat,* Eptesicus fuscus*: The shape of target images in echolocation*, J. Comput. Physiol. A, 167 (1990), pp. 589–616.

[46] P. H. Smith, P. X. Joris, M. I. Banks, and T. Yin, *Responses of cochlear nucleus cells and projections of their axons*, in The Mammalian Cochlear Nuclei: Organization and Function, M. A. Merchan, J. M. Juiz, D. A. Godfrey, and E. Mugnaini, eds., Plenum, New York, 1993, pp. 349–360.

[47] M. Stemmler, *A single spike suffices: The simplest form of stochastic resonance in model neurons*, Computation in Neural Systems, 7 (1996), pp. 687–716.

[48] G. Svirskis, V. Kotak, D. Sanes, and J. Rinzel, *Enhancement of signal-to-noise ratio and phase locking for small inputs by a low threshold outward current in auditory neurons*, J. Neurosci., 22 (2002), pp. 11019–11025.

[49] T. Yin, *personal communication*.

[50] W. Yost and G. Gourevitch, eds., *Directional Hearing*, Springer-Verlag, New York, 1987.

[51] E. Young, J.-M. Robert, and W. Schofner, *Regularity and latency of units in ventral cochlear nucleus: Implications for unit classification and generation of response properties*, J. Neurophysiol., 60 (1988), pp. 1–29.

SIAM J. APPL. MATH.
Vol. 68, No. 3, pp. 738–759

STABILITY OF TRAFFIC FLOW BEHAVIOR WITH DISTRIBUTED DELAYS MODELING THE MEMORY EFFECTS OF THE DRIVERS[*]

RIFAT SIPAHI[†], FATIHCAN M. ATAY[‡], AND SILVIU-IULIAN NICULESCU[§]

Abstract. Stability analysis of a single-lane microscopic car-following model is studied analytically from the perspective of delayed reactions of human drivers. In the literature, the delayed reactions of the drivers are modeled with discrete delays, which assume that drivers make their control decisions based on the stimuli they receive from a point of time in the history. We improve this model by introducing a distribution of delays, which assumes that the control actions are based on information distributed over an interval of time in history. Such an assumption is more realistic, as it takes into consideration the memory capabilities of the drivers and the inevitable heterogeneity of their delay times. We calculate exact stability regions in the parameter space of some realistic delay distributions. Case studies are provided demonstrating the application of the results.

Key words. memory, distributed delay, traffic dynamics, traffic flow stability

AMS subject classifications. 34K20, 34K60, 93D20

DOI. 10.1137/060673813

1. Introduction.

Traffic behavior has been an important research topic since the 1930s, with the of aim of reducing the undesirable social (e.g., vehicle accidents) and economic (for example, congestion and increasing pollution) effects of increasingly complex traffic loads. For this purpose, one needs a good understanding of traffic dynamics, in which many parameters or constraints play an important role, such as the physical conditions of highways, mechanical properties of vehicles, psychological states of the drivers, traffic laws, on- and off-ramps, multiple lanes, traffic density, etc. The literature contains various models addressing different phenomena; see, e.g., the survey [12] and the references therein.

Among the parameters that play a major role in traffic behavior, there exists a critical one which has been recognized as early as the 1950s [4], namely the *time delay.* It mainly originates due to the time needed by human drivers for sensing, being conscious, and performing control actions [7]. Consequently, traffic dynamics, and ultimately its mathematical models, *inherently* carry *time delays.* See, e.g., [2, 23, 19, 18, 26, 27] for some delay models and related discussions.

Stability characterizations of traffic models may be quite different when time delays are taken into account; for instance, a stable delay-free dynamics may become unstable when delays are considered. Therefore, a thorough stability analysis of the dynamics in the time delay domain is necessary. Without entering into details, we shall consider a continuous-time microscopic car-following model proposed in [4, 21] to describe traffic behavior. What distinguishes this study is the idea of incorporating *distributed delays* in order to represent the memory of the drivers.

[*] Received by the editors October 31, 2006; accepted for publication (in revised form) September 19, 2007; published electronically December 20, 2007.
http://www.siam.org/journals/siap/68-3/67381.html

[†] Corresponding author. Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115 (rifat@coe.neu.edu).

[‡] Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, Leipzig 04103, Germany (atay@member.ams.org).

[§] Laboratoire de Signaux et Systèmes (L2S), Supélec, 3, rue Joliot Curie, 91190, Gif-sur-Yvette, France (Silviu.Niculescu@lss.supelec.fr).

Human-in-the-loop modeling. From the perspective of control theory, traffic flow can be seen as human-in-the-loop dynamics [16], since human drivers play the major role in the evolution of the traffic flow. Such a dynamical structure exhibits richer and more complicated features not only due to the challenges of modeling human beings, but also due to the delayed control actions of the drivers, which distinguishes them from fast response characteristics of sensors and controllers. On the other hand, *the modeling of time-delays* representing the behavior of the drivers is a challenge. Models in the existing literature make use of discrete delays [2, 27, 23], which describe an action of the driver at time t that is based on what is experienced at a point of time $t - \tau$, $\tau \geq 0$, in the past. The stability of the arising dynamics has been studied by several authors from various perspectives. See [2, 6, 18, 19] for the utilization of nonlinear optimal velocity functions relating headway versus desired velocity; [19, 18, 20] for one- and two dimensional bifurcation analysis as well as phase diagrams characterizing oscillations, collision, and the stopping motion nature of the traffic flow; [27] for incorporating human driver modeling into traffic flow where human adaptation and anticipation are considered along with multiple vehicle following strategies and phase diagrams revealing collisions, oscillations, and accident-free traffic flow; and [23] for the stability analysis of traffic flow in which multiple discrete time-delays corresponding to different time scales of reaction of human drivers against velocity and headway changes are taken into account.

We note, however, that discrete-delay models can have their shortcomings. For instance, regarding τ as a fixed unchanging quantity with known value would ignore the possibility that the dynamics may possess inherent "memory" effects which use the past history of the received information. Especially for the traffic flow, the presence of human drivers suggests that memory effects should be taken into consideration. Moreover, the behavior of individual drivers and their reactions are not identical, and in reality exhibit a distribution of inhomogeneous behavior. Therefore, the use of a model taking into account the distributed nature of the delays will yield a better representation of reality. Such an argument has also some potential in modeling the dynamics of human-in-the-loop systems [16], which will ultimately open new research directions for designing driver assistance or semiactive controllers that can guide the human beings for a safer drive in uncertain environments.

In the present work, motivated by the above reasoning, we model the delayed action/decision of human drivers using distributed delays. The physical basis of the model is the fact that the drivers perform their decisions based on what they continuously observe (during a memory window) from the evolving traffic flow, during which some information is retained and used in the decision-making process. These decisions are clearly limited by the capacity of the memory, i.e., the size of the memory window, which will be an important parameter in the analysis.

Objective and approach. The main objectives of the paper are to (a) study the stability margin of the traffic flow dynamics over a microscopic car-following model in the parameter space defining the memory effects, and (b) discuss the analytical and physical interpretations of the results for several practical traffic scenarios.

For the stability analysis, we use a frequency-domain approach combined with some simple geometric ideas, and give necessary and sufficient conditions for the stability of the dynamics. In this context, we also explore whether the stability region consists of a single connected set or of several "islands" (also called *pockets*) of stability. We note that in the present text the stability is robust against small delay variations (section 2); thus small delay perturbations do not induce instability; see some discussions in [24, 10, 17, 15]. The focus of the work here is an analytical study of the effects of distributed delays on the stability of traffic flow.

Moreover, it is shown in [1] that distributed memory enhances the stability margin in a network stabilization scheme. It is interesting to see whether a similar stability-enhancing feature also exists in traffic flow dynamics when the drivers make use of the past history of the traffic evolution in their decision-making, which would clearly have significance in realistic traffic modeling.

To model memory effects one can use common delay distributions, namely γ-distribution with and without a gap, uniform distribution. As discussed below, the first two cases are easier to handle, whereas the uniform distribution needs a more careful treatment. This mainly originates from a characteristic equation which does not exactly fit into standard classes treated in the literature, in that (i) the delay terms appear not only in exponential terms but also in the coefficients, (ii) the characteristic equation includes complex coefficients, which does not carry the features of those with real coefficients as treated in the literature [24, 17, 10, 15], (iii) there exist two independent delay parameters; stability investigations under the presence of more than one delay is quite complicated [9, 17, 22, 25, 24, 15].

The key ideas of our analysis are based on decoupling the dynamics into lower dimensions and introducing an interconnection scheme interpretation from control feedback systems theory. Such an analysis is the key to the complete analytical treatment of the arising characteristic equation. Connections with existing approaches and methodologies will be made in the following sections. In section 2, we present the microscopic car-following model and the spatial configuration of vehicles in traffic. Section 3 introduces some preliminaries, and section 4 is devoted to the main results on the stability analysis, where analytical and geometrical arguments for deriving the stability results are developed. Section 5 illustrates the results with numerical examples for several traffic scenarios, and a brief summary in section 6 concludes the presentation.

Notation. We use $\mathbb{R}$ for real numbers, where an additional $+$ or $-$ sign as a subscript indicates the positive and negative real numbers, respectively. Similarly, $\mathbb{C}_+$ and $\mathbb{C}_-$ denote complex numbers with positive and negative real parts, respectively. The imaginary axis of the complex plane is denoted by $j\mathbb{R}$, where $j = \sqrt{-1}$. We use $s \in \mathbb{C}$ for the Laplace variable, whose values on the imaginary axis are denoted by $s = j\omega$ for $\omega \in \mathbb{R}$. The eigenvalues of the matrix A are represented by $\lambda_i(A)$. $\angle z$ denotes the argument of the complex number z.

2. Microscopic car-following model with distributed delays. Although undesirable, the presence of time-delays in the process of decision making and performing a control action by human drivers is neither avoidable nor negligible (e.g., [27, 26, 2, 23]). In this section, we develop memory effects on a conceptual microscopic car-following model.

Microscopic car-following model. In order to understand the behavior of traffic flow and to propose ways to avoid its undesirable effects (congestion, accidents, economic losses, time losses, degradation of the quality of the environment), various mathematical models have been proposed in the literature [8, 12, 6, 18, 27, 26]. Despite the available results, the topic still offers open problems today.

A main direction in traffic research is focused on highway traffic models [27, 19, 12, 6], since travel times on highways are longer and travel speed is relatively high, and health and economic issues are largely at risk [12]. Furthermore, in most cases a single-lane problem is considered. Such an assumption is quite realistic and also allows one to obtain further insight on the problem due to its simpler mathematical formulation. This type of model also forms the main focus in our study. A single-lane traffic flow, in which a chain of vehicles travels at a constant velocity (so-called quasi steady-state) without changing lanes, is considered. We use a *microscopic model* in which the dynamics of individual vehicles and drivers is taken into account. Despite the simplicity of the model, an *analytical* stability investigation becomes nontrivial due to the presence of time-delays, as we present below. Furthermore, there exist various complications in realistically modeling human beings and their delayed reactions.

The primary interest in this work is to shed light, from the perspective of memory effects of human drivers, on the stability of traffic behavior. As a starting point in this new direction, the linear stability analysis of the perturbations around the constant-velocity solutions will be of particular interest in order to reveal the stability features with respect to the parametric domain defining the memory effects. In order to achieve this, a linear mathematical model inspired by earlier work is introduced in what follows. Due to the linear stability analysis, however, the mathematical model considered here is not dependent on some additional parameters defining the traffic, such as dependence on headway, drivers' sensitivity as a function of headway, acceleration and deceleration characteristics, human anticipation, etc. Readers are directed to the work in [6, 12, 20, 27] for more elaborate models.

Discrete-delay models. Many studies in the literature model the time-delayed actions of the drivers by a discrete delay τ. One such microscopic car-following model is given as [4, 12, 21, 23]

$$\dot{v}_i(t) = \kappa_i(v_{i-1}(t-\tau) - v_i(t-\tau)). \tag{2.1}$$

This model represents the dynamics of the velocity perturbations v_i around constant-velocity solutions (the equilibrium at which *all* vehicles travel at a constant velocity V) of the traffic flow. The parameter $\kappa_i > 0$ denotes the sensitivity of a driver to the velocity difference between his vehicle and the one in front, and gives a measure of the driver's reactivity (aggressiveness) based on his experience and knowledge of the environment. For further information on discrete-delay models, refer to [2, 6, 18, 20, 27]; on the links between the stability features of discrete-delay models and Hopf bifurcations, some interesting arguments can be found in [18, 20, 19]; and on the phase diagrams of traffic flow characterizing collisions, oscillations, and accident-free flow, see [20, 27]. The cited references are also valuable sources for various mathematical models over which additional parameters not considered here, such as headway, sensitivity, acceleration, and deceleration dependence can be studied.

Distributed delay model. By incorporating a general memory effect, $f(\tau)$, into the system (2.1), we arrive at the following generalized model:

$$\dot{v}_i(t) = \kappa_i \int_0^\infty f(\tau)(v_{i-1}(t-\tau) - v_i(t-\tau))\, d\tau, \tag{2.2}$$

where we assume that the delay kernel f is a measurable function of exponential order. When f is a Dirac delta function, (2.2) reduces to the discrete delay model (2.1). See above for discussions of earlier work on discrete delay models.

Delay distributions. We will consider the following common distribution functions f in the model (2.2):

(1) *Uniform distribution.* For $\delta > 0$ and $h \geq 0$, the distribution

$$f(\tau) = \begin{cases} 1/\delta & \text{if } h < \tau < h + \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

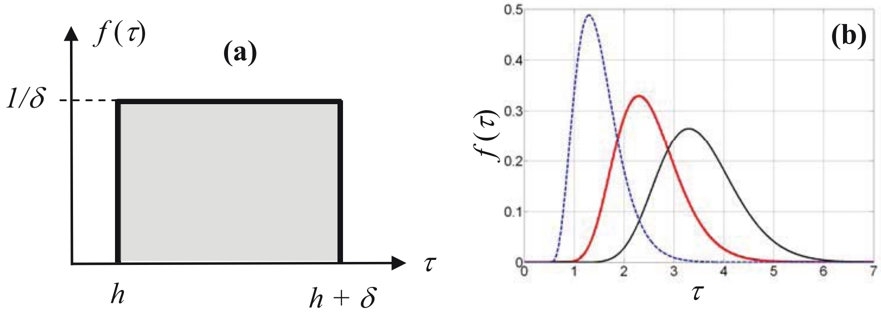


FIG. 2.1. (a) *Uniform distribution with h being dead-time and δ being size of the memory window.* (b) *γ-distribution with gap, $h = 0.5$, for $p = 5$ (dashed curve), $p = 10$ (thick curve), $p = 20$ (thin curve), where $pq = 1$.*

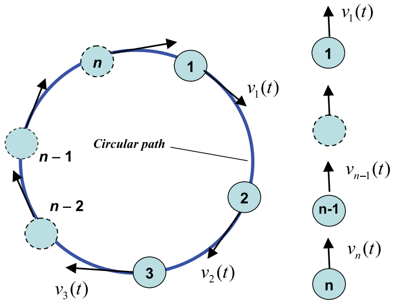


FIG. 2.2. *Ring and linear configurations of the traffic model with n vehicles.*

represents an average of the information available in the short-term memory, and can be considered as a first order approximation to a more complicated distribution; see Figure 2.1(a). It will be of interest to see how h (corresponding to some dead-time before the perception of the sensory input by the driver) and the window size δ affect the stability.

(2) *γ-distribution with and without gap.* We have

$$f(\tau) = \begin{cases} \dfrac{(\tau-h)^{p-1}e^{-(\tau-h)/q}}{q^p\Gamma(p)} & \text{if } \tau \geq h, \\ 0 & \text{if } \tau < h, \end{cases} \tag{2.4}$$

where p, q are positive parameters of the distribution, Γ denotes the gamma function, and the gap $h \geq 0$ represents the dead-time. The mean value of f is $h + pq$, and the variance is pq^2 (which exist for $p > 2$), which is a measure of the length of the memory window, q being a scaling factor ($q = 1$, mean value $= h + p$, variance $= p$). It will be of interest to see how the quantities h, p, and q affect stability. See in Figure 2.1(b) the γ-distribution with gap for various p values keeping pq fixed.

Spatial configuration of the model. We consider two widely utilized configurations (Figure 2.2) with n number of vehicles: (a) vehicles traveling around a ring, and (b) vehicles arranged along a linear path. In the linear configuration, the vehicle in front (for which we use the convention of labeling with index $i = 1$) travels at a constant velocity, i.e., $\dot{v}_1(t) = 0$. Hence, the linear configuration can be derived from the circular one by setting $\kappa_1 = 0$.

3. Problem statement and preliminaries. The system (2.2) can be expressed in vector form as

$$\dot{\mathbf{v}}(t) = \int_0^\infty J\mathbf{v}(t-\tau)f(\tau)\,d\tau, \tag{3.1}$$

where $\mathbf{v} = (v_1, \dots, v_n)$ and $J \in \mathbf{R}^{n\times n}$ is a configuration matrix weighted by the driver sensitivities κ_i. For the *circular* form of the configuration in Figure 2.2, it is obvious that $v_0 = v_n$; thus the appropriate index selection in (2.2) becomes $i = 1, \dots, n$. Consequently, the matrix J takes the form

$$J = \begin{pmatrix} -\kappa_1 & 0 & \cdots & 0 & \kappa_1 \\ \kappa_2 & -\kappa_2 & 0 & \cdots & 0 \\ 0 & \kappa_3 & -\kappa_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \kappa_n & -\kappa_n \end{pmatrix}. \tag{3.2}$$

The configuration matrix for the linear arrangement of vehicles is obtained by setting $\kappa_1 = 0$, and it is denoted by J'.

The characteristic equation for (3.1) is given by

$$\chi(s) = \det[sI - JF(s)] = 0, \tag{3.3}$$

where $F(s)$ is the Laplace transform of f. We assume the generic case that J is diagonalizable; that is, its eigenvectors form a basis for $\mathbb{R}^n$. Then (3.3) can be expressed as $\chi(s) = \prod_{i=1}^n (s - \lambda_i(J)F(s)) = 0$, where $\lambda_i(J)$ is the ith eigenvalue of J. In the remainder of the text, λ_i will denote $\lambda_i(J)$ unless otherwise stated. The roots corresponding to the ith factor of χ, i.e., the solutions s of the equation

$$\chi_i(s) \triangleq s - \lambda_i F(s) = 0, \tag{3.4}$$

determine the fate of perturbations along the ith eigenvector of J. The perturbations die out if and only if Re(s) < 0 for all solutions s of (3.4). As the stability depends on the spectrum of J, the following properties will be needed in the analysis.

LEMMA 3.1. *The configuration matrix J has a simple zero eigenvalue, and all its remaining eigenvalues have negative real parts. Furthermore, $|\lambda_i| \le 2\kappa_{\max}$, $i = 1, \dots, n$, where $\kappa_{\max} = \max\{\kappa_i\}$. If $\kappa_i = \kappa$ for all i, then the eigenvalues of J are given by*

$$\lambda_i = \kappa(e^{j2\pi(i-1)/n} - 1), \qquad i = 1, \dots, n. \tag{3.5}$$

If J' denotes the matrix obtained from J by setting $\kappa_1 = 0$, then the eigenvalues of J' are real and given by $0, -\kappa_2, -\kappa_3, \dots, -\kappa_n$.

Proof. The rows of J sum to zero, implying that zero is an eigenvalue corresponding to the eigenvector $(1, 1, \dots, 1)^\top$. The circular configuration corresponds to a (weighted) directed graph which is strongly connected; that is, there is a directed path from every vertex to any other vertex. It follows that zero is a simple eigenvalue of the graph Laplacian matrix, which is equal to J in the present case. Furthermore, by Gershgorin's theorem [13], the eigenvalues of J are located in the union of discs

$$\lambda_i \in \bigcup_{\ell=1}^n \mathcal{B}(-\kappa_\ell) = \mathcal{B}(-\kappa_{\max}), \tag{3.6}$$

where $\mathcal{B}(r) = \{s \in \mathbb{C} : |s - r| \leq |r|\}$ is the disc centered at $r \in \mathbb{R}$ with radius $|r|$. Hence, all nonzero eigenvalues of J have negative real parts since $\kappa_i > 0$ for all i. Moreover, $|\lambda_i| \leq 2\kappa_{\max}$ by (3.6). One can also directly calculate the characteristic polynomial of J by an easy expansion by cofactors to obtain

$$\prod_{i=1}^{n}(\lambda + \kappa_i) - \prod_{i=1}^{n}\kappa_i = 0. \tag{3.7}$$

So, if $\kappa_i = \kappa$ for all i, then $(\lambda/\kappa + 1)^n = 1$; i.e., $\lambda/\kappa + 1$ are the nth roots of unity, which yields (3.5). The spectrum of J' follows from (3.7) by setting $\kappa_1 = 0$. □

We make the convention of labeling the zero eigenvalue of J as λ_1, and denote the corresponding eigenvector as $v_1 = (1, 1, \ldots, 1)^\top$. Hence, $\mathrm{Re}(\lambda_i) < 0$ for $i \geq 2$. Note that $s = 0$ is a solution of (3.4) when $i = 1$. This is the indicator of the rigid body rotation of the whole configuration of vehicles, and consequently the stability of the configuration is determined by the roots of the modified characteristic equation

$$\hat{\chi}(s) = \prod_{i=2}^{n}(s - \lambda_i F(s)) = 0. \tag{3.8}$$

Hence, car-following dynamics given by (2.2) is stable if all solutions s of (3.8) have negative real parts.

Remark 3.2. *Notice that all roots of the delay-free form of* (3.8) *are stable, as per Lemma* 3.1. *This forms only the starting point of the stability analysis. The main results presented in the following section reveal the necessary and sufficient conditions for the complete stability analysis of* (2.2) *with respect to the parametric domain of interest: delay distribution $f(\tau)$ and the spectrum of J. Readers are directed to some other models in the earlier cited references for the incorporation of additional parameters that may play a role in the stability of traffic dynamics.*

4. Main results.

We first introduce an interconnection scheme idea for the characteristic equation (3.8), which will allow a geometric approach to the stability analysis by an appropriate separation of parameters.

4.1. Interconnection scheme interpretation.

Consider first the uniform distribution (2.3), whose Laplace transform is

$$F(s) = \frac{e^{-sh}(1 - e^{-s\delta})}{s\delta}. \tag{4.1}$$

Note that $F(s) \to e^{-sh}$ as $\delta \to 0$, corresponding to the fact that f approaches a Dirac delta at h, where one recovers the discrete-delay model (2.1) with $\tau = h$. Using (4.1) in (3.4) gives

$$\chi_i(s) = s - \lambda_i e^{-sh}\frac{1 - e^{-s\delta}}{s\delta} = 0. \tag{4.2}$$

The singularity of χ_i at zero is removable since $\lim_{s\to 0}\chi_i(s) = -\lambda_i$. Hence by defining $\chi_i(0) = -\lambda_i$, we can treat χ_i as an analytical function and use arguments based on the continuous dependence of its roots on the parameters. In particular, since $\lambda_i \neq 0$ for $i \geq 2$, by Lemma 3.1, the following result is immediate.

Lemma 4.1. *$s = 0$ is not a solution to* (4.2) *for $i = 2, \ldots, n$.*

We rearrange (4.2) as

$$H_i(s)\cdot\Delta(s) = 1, \quad i = 2,\dots,n, \tag{4.3}$$

$$\text{where}\quad H_i(s) = \frac{\lambda_i e^{-sh}}{s}, \qquad \Delta(s) = \frac{1-e^{-s\delta}}{s\delta}. \tag{4.4}$$

The advantage of the form (4.3) lies mainly in the separation of the parameter δ and the eigenvalues of J, which will simplify the analysis. Similarly, for the γ-distribution we have $F(s) = e^{-hs}(qs+1)^{-p}$, and the corresponding characteristic equation can be expressed as an interconnection scheme,

$$H_i(s)\cdot\Delta_\gamma(s) = \frac{\lambda_i e^{-sh}}{s}\cdot\frac{1}{(qs+1)^p} = 1. \tag{4.5}$$

Note that $|\Delta_\gamma(j\omega)| \leq 1\ \forall\omega\in\mathbf{R}$, for any q.

4.2. Stability analysis. Given λ_i, the roots of (3.8) depend continuously on the parameters of the distribution $f(\tau)$ (see [5]). The method for stability analysis can then be summarized as follows. (a) Check the stability of the delay-free dynamics (3.8). (b) Calculate the characteristic roots on the imaginary axis, $s = j\omega$, of the interconnection scheme $H_i(s)\Delta(s) - 1 = 0$. (c) Check in which direction $s = j\omega$ crosses the imaginary axis with respect to the parametric domain of interest. Notice that the characteristic equation in (4.2) does not necessarily exhibit complex conjugate s solutions if $\lambda_i \in \mathbb{C}$.

We present below the stability analysis of (4.3) for the case of uniform distribution. The technique is easily extendable to γ-distribution with and without a gap with the construct of the interconnection scheme. The challenges in assessing the stability are (i) analysis should be performed in the multiple parameter space (h,δ,λ_i); (ii) the interconnection scheme carries complex coefficients; (iii) one of the parameters from (i), δ, appears both in an exponent and in the denominator of the interconnection scheme. The complications (i)–(iii) prevent our benefiting from many stability analysis techniques [3, 24, 17, 22, 15]. Before we introduce how to tackle these difficulties, we start with some conditions for the roots of (4.3) to exhibit imaginary axis crossings.

The following algorithm enables the calculation of the curves in the (h,δ) domain on which (4.3) has a solution of the form $s = j\omega$.

Computation algorithm for $s = j\omega$ and (h,δ) pairs. Given λ_i, the characteristic roots $s = j\omega$ crossing the imaginary axis and the corresponding parameter pairs (h,δ) can be exhaustively computed as follows. First, we write the magnitude condition in (4.3), at $s = j\omega$, which allows us to compute δ independently from h. (This is a direct consequence of the separation feature introduced by the interconnection scheme interpretation.) Second, on the same equation (4.2), we write the argument condition, which yields h.

Step 1. Define the continuous function $f_\Delta : \mathbb{R}\to\mathbb{R}_+$ by

$$f_\Delta(u) = \begin{cases} \sin^2 u/u^2, & u\neq 0,\\ 1, & u = 0.\end{cases} \tag{4.6}$$

Let

$$u = \frac{\delta\omega}{2}\in\mathbb{R}. \tag{4.7}$$

Step 2. By substituting (4.7) into the magnitude of (4.3),

$$|H_i(j\omega)|^2|\Delta(j\omega)|^2 = 1, \tag{4.8}$$

one obtains

$$\frac{|\lambda_i|^2}{(\omega)^2} f_\Delta(u) = 1, \tag{4.9}$$

which can be alternatively written as

$$f_\Delta(u)\frac{|\lambda_i|^2\delta^2}{4u^2} = 1 \Rightarrow f_\Delta(u) = \frac{4u^2}{|\lambda_i|^2\delta^2} = ku^2, \tag{4.10}$$

where $k = 4/(|\lambda_i|^2\delta^2)$.

The following proposition is immediate using Steps 1–2 above.

Proposition 4.2 (frequency-sweeping characterization [17]). *$s = j\omega$ is a root of (4.3) if and only if $\omega \leq |\lambda_i|$.*

Proof. It is clear from (4.6) and (4.7) that $|H_i(j\omega)| = |\lambda_i|/|\omega| > 1$ should be satisfied such that (4.8) holds. □

Step 3. One simply *sweeps* u and obtains ω from (4.9).

Step 4. Using u and ω from Step 3, solve for δ from (4.7).

Step 5. Define next

$$\lambda_i = |\lambda_i|e^{j\phi_i} \quad \text{with } \phi_i = \angle\lambda_i, \tag{4.11}$$

where $\phi \in (\pi/2, 3\pi/2)$ as per Lemma 3.1, and use the argument condition on (4.3) to compute h.

Step 6. Rearrange (4.3) as

$$e^{-j\omega h} = \frac{-\delta(\omega)^2}{\lambda_i(1 - e^{-j\delta\omega})},$$

from which one obtains the following by equating the arguments of both sides:

$$h = \frac{1}{\omega}[-\pi + \phi_i + \angle(1 - \cos(\delta\omega) + j\sin(\delta\omega)) + 2\pi\ell], \quad \ell \in \mathbb{Z}. \tag{4.12}$$

Step 7. Using (4.7), it is clear that $\angle(1 - \cos(\delta\omega) + j\sin(\delta\omega)) = \tan^{-1}\left(\frac{\cos u}{\sin u}\right) = \frac{\pi}{2} - u$, simplifying (4.12) to

$$h = \frac{1}{\omega}\left(-\frac{\pi}{2} + \phi_i - u + 2\pi\ell\right). \tag{4.13}$$

Proposition 4.3. *There exists a connected stability region of the traffic flow dynamics (2.2) in the parameter space (h, δ) that includes the origin $(h, \delta) = (0, 0)$. The bounds of this region on the δ and h axes are respectively given by*

$$\bar{h} = \min_{2\leq i\leq n}\left(\frac{2\phi_i - \pi}{2|\lambda_i|}\right) \quad \textit{and} \quad \bar{\delta} = \min_{2\leq i\leq n}\left(-\frac{(2\phi_i - \pi)^2}{2|\lambda_i|\cos(\phi_i)}\right). \tag{4.14}$$

Remark 4.4. *We shall show later that the stability region mentioned in the above proposition is actually the unique stability region in the parameter space.*

Proof. When $h = \delta = 0$, the distribution function in (2.3) is a Dirac delta whose Laplace transform equals 1. From (4.2), the characteristic roots are $s = \lambda_i$, $i = 2, \ldots, n$, where $\mathrm{Re}(\lambda_i) < 0$ by Lemma 3.1. Thus, (2.2) is stable for $(h, \delta) = (0, 0)$. Consequently, the stability region in the (h, δ) parameter domain contains an open neighborhood of the origin $(h, \delta) = (0, 0)$ [5]. We next calculate the stability margins along the δ- and h-axes.

Stability when $\delta = 0$. For $\delta \to 0$ the characteristic equation (4.2) becomes

$$s = \lambda_i e^{-sh}. \tag{4.15}$$

Equation (4.15) has a stable root for $h = 0$ (see Proposition 4.3), and stability is lost if a characteristic root crosses the imaginary axis for some $h \neq 0$. Thus, the stability is preserved between $h = 0$ and the minimum positive h for which (4.15) has a solution $s = j\omega$. This minimum h is computed as follows. By Lemma 4.1, the magnitude condition on (4.15) yields $\omega = \mp|\lambda_i|$, and the argument condition requires

$$h = \frac{1}{|\lambda_i|}\left(-\frac{\pi}{2} \mp \phi_i + 2\pi\ell\right), \quad \ell \in \mathbb{Z}, \tag{4.16}$$

with $+\phi_i$ when $\omega > 0$. Using the above equation, one obtains the smallest positive h, $\bar{h}$ as given in (4.14). Hence, when $\delta = 0$, (4.2) is stable for $h \in [0, \bar{h})$.

Stability when $h = 0$. The smallest positive value of δ for which stability is lost is denoted by $\bar{\delta}$; that is, the stability interval along δ-axis is $\delta \in [0, \bar{\delta})$. The characteristic equation (4.2) when $h = 0$ is

$$\frac{\delta s^2}{\lambda_i} + e^{-s\delta} - 1 = 0. \tag{4.17}$$

Letting $s = j\omega$, solving for the exponential term, and equating the magnitude conditions of both sides, one easily obtains δ:

$$\delta = -\frac{2|\lambda_i|\cos(\phi_i)}{\omega^2}. \tag{4.18}$$

Moreover, solving the exponential term in (4.17), equating the arguments of both sides, substituting δ from (4.18), and noting that $\omega \neq 0$ by Lemma 4.1, we obtain

$$\omega = \frac{2|\lambda_i|\cos(\phi_i)}{\angle(-\cos(2\phi_i) + j\sin(2\phi_i)) \pm 2\pi\ell}, \quad \ell = 0, 1, 2, \ldots,$$

which simplifies and gives rise to δ

$$\omega = \frac{2|\lambda_i|\cos(\phi_i)}{\pi - 2\phi_i \mp 2\pi\ell}, \quad \delta = -\frac{(\pi - 2\phi_i \mp 2\pi\ell)^2}{2|\lambda_i|\cos(\phi_i)}. \tag{4.19}$$

When $\delta = 0$ the dynamics is stable. Following the root continuity arguments [5], the stability is lost for the smallest positive δ, which is attained when $\ell = 0$ and equal to $\bar{\delta}$ as given in (4.14). □

Characterization of the geometry of the stability boundaries. The stability boundaries are among those curves in the (h, δ) parametric domain which give rise to an $s = j\omega$ solution in the interconnection scheme. In order to characterize the geometry of the stability boundaries, one has to establish the link from u domain

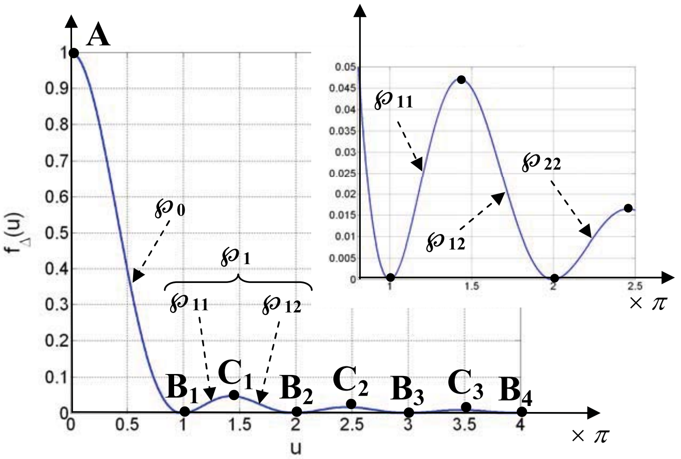


FIG. 4.1. *$f_\Delta(u)$ versus u. B_ℓ and C_ℓ are end points and the local maxima of the segments $\wp_\ell$, $\ell > 0$, respectively.*

to (h, δ). To achieve this, we give some definitions first. Define the extrema points of f_Δ function with $A(f_\Delta(0))$, $B_\ell(f_\Delta(\ell\pi))$, and $C_\ell(f_\Delta((2\ell+1)\pi/2))$, $\ell = 1, 2, 3, \ldots$; see Figure 4.1. Partition $f_\Delta(u)$ into segments and label each one of them as follows: $\wp_0 = f_\Delta(u)$ with $u \in [0, \pi]$; $\wp_{\ell,\ell} = f_\Delta(u)$ with $u \in [\ell\pi, (2\ell+1)\pi/2]$; $\wp_{\ell,\ell+1} = f_\Delta(u)$ with $u \in [(2\ell+1)\pi/2, 2\ell\pi]$, $\ell = 1, 2, 3, \ldots$. Obviously,

$$f_\Delta = \bigcup_{\ell \geq 0} \wp_\ell, \quad \text{where } \wp_\ell = \wp_{\ell,\ell} \bigcup \wp_{\ell,\ell+1} \quad \forall \ell > 0. \tag{4.20}$$

Notice that each of the "segments" $\wp_0$, $\wp_{\ell,\ell}$, and $\wp_{\ell,\ell+1}$ as a function is monotonic with respect to the variable u, as seen in Figure 4.1. The points B_ℓ and C_ℓ in this figure correspond to the end points and the local maxima of the segments $\wp_\ell$, $\ell > 0$, respectively. Due to the symmetry of the function f_Δ, $f(u) = f(-u)$, and by (4.7), we will restrict our subsequent analysis to positive ω, $\omega > 0$.

Monotonicity properties. We will now utilize the monotonicity properties of the "segments" of $f_\Delta(u)$ in order to explain how their mapping in (h, δ) forms. We first comment on the extremities in u and (h, δ) domains. Given λ_i, at point A we have $f_\Delta(u = 0) = 1$; hence from (4.8), $\omega = |\lambda_i|$ and from (4.7), $\delta = 0$. At the end points of $\wp_\ell$, that is B_ℓ and $B_{\ell+1}$, we have $\lim_{u \to \ell\pi} f_\Delta(u) \to 0^+$; thus for a solution of the interconnection scheme in the form of $s = j\omega$ to exist, it is clear from (4.9) that $\omega \to 0^+$. Therefore, the image of B_ℓ and $B_{\ell+1}$ on (h, δ) space corresponds to infinity, since these curves are open-ended curves.

Recall that if $s = j\omega$ solution of the interconnection scheme exists, then (4.10) holds. This indicates that u solutions of (4.8) lie at the intersection points of the curve $f_\Delta(u)$ and the parabola u^2 parameterized by k. This relationship is generically depicted in Figure 4.2(a) for various k values and for positive $u > 0$. Notice from (4.10) that δ is inversely proportional to k.

Let us now elaborate on Figure 4.2(a), since it depicts the geometry of the solution points in u (such as points C, D, E, and F), which we will use for characterizing the corresponding geometry in the (h, δ) domain. Denote this correspondence from u domain to (h, δ) domain by $f_\Delta(u) \mapsto \zeta(h, \delta) : \mathbb{R}_+ \mapsto \mathbb{R} \times \mathbb{R}_+$. Next, define ζ_0, $\zeta_{\ell,\ell}$,

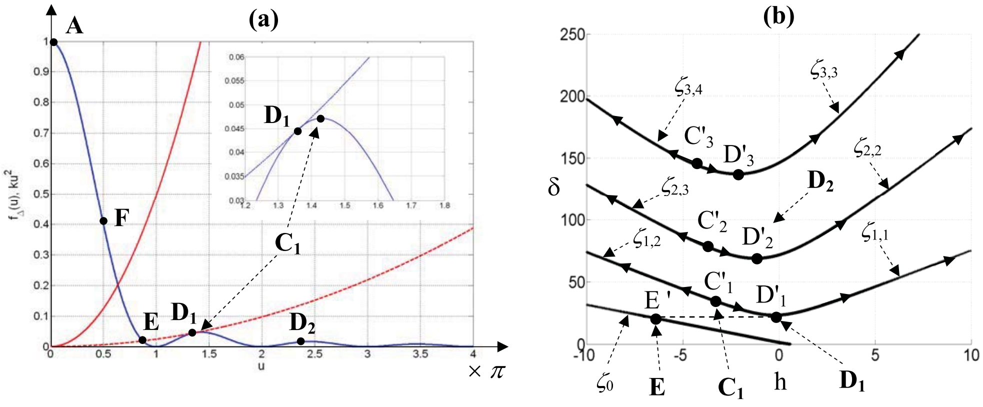


FIG. 4.2. (a) *Generic depiction of f_Δ and ku^2 versus u, where the parabola sweeps the first quadrant counterclockwise with increasing k. Inset: Distinction between points D_1 and C_1.* (b) *The mapping of the intersection points between f_Δ and ku^2 to the (h, δ) domain ($C_i \mapsto C'_i$, $D_i \mapsto D'_i$, $E \mapsto E'$, etc.). Arrows indicate decreasing direction of ω.*

and $\zeta_{\ell,\ell+1}$ segments in the (h, δ) domain that correspond to the segments $\wp_0$, $\wp_{\ell,\ell}$, and $\wp_{\ell,\ell+1}$, respectively, in u domain, and similarly, let $\zeta_\ell = \zeta_{\ell,\ell} \bigcup \zeta_{\ell,\ell+1}$ $\forall \ell > 0$. Figure 4.2(b) presents generically some of these curves in the (h, δ) domain, where E' corresponds to E of Figure 4.2(a); D'_1, D'_2, D'_3, etc., are the mappings of the points at which the parabola in Figure 4.2(a) is tangent to $\wp_{11}$, $\wp_{22}$, $\wp_{33}$, etc., segments (shown in Figure 4.1), respectively; C'_1, C'_2, C'_3, etc., are the mappings of the local maxima points C'_1, C'_2, C'_3, etc., at which the maximum omega on a respective ζ_ℓ, $\ell > 0$, curve. The points C_ℓ can also be seen as the end points of $\zeta_{\ell,\ell}$ and $\zeta_{\ell,\ell+1}$ curves.

Notice that, for a given λ_i, intersection points in Figure 4.2(a) for fixed k give rise to fixed δ, as per (4.10). For instance, for a particular selection of k, the parabola may intersect $\wp_{1,1}$ and $\wp_{1,2}$, giving rise to two points, one on $\zeta_{1,1}$ and the other on the $\zeta_{1,2}$ curve. Furthermore, all such points are earmarked by ω. See the two points in Figure 4.3(c)–(d) labeled with ω_1 and ω_2 as an example.

REMARK 4.5. *The ζ_0 curve in Figure* 4.2(b) *is due to mapping of the $\wp_0$ segment; see also Figure* 4.3(a)–(b)*. For the remaining segments, we state the following. For a given k, any two points at the intersection of the parabola and the segment $\wp_\ell$, $\ell > 0$, give rise to two points on the ζ_ℓ curve; see Figure* 4.3(c)–(d)*. These two points are generated by two distinct u values, the larger of which corresponds to the larger ω as per* (4.9)*. It is clear from* (4.13) *that larger u corresponds to smaller h. Consequently, the point marked by ω_1 in Figure* 4.3(c)–(d) *arises from the intersection between ku^2 and $\wp_{1,2}$; thus it lies on the $\zeta_{1,2}$ curve, and $\omega_1 > \omega_2$ holds. The arrows on ζ curves in Figure* 4.2(b) *and Figure* 4.3(c)–(d) *indicate the direction of decreasing ω on the respective curves. Finally, the point marked with C_ℓ indicates the location of maximum ω attained on the curve ζ_ℓ, $\ell > 0$, which is clearly due to the local maximum of the f_Δ curve in the interval $u \in (\pi, 2\pi)$ (maximum u thus maximum ω as per* (4.13)*). Figure* 4.3 *presents separately the way in which the first two curves ($\wp_0 \mapsto \zeta_0$ and $\wp_1 \mapsto \zeta_1$) are generated.*

PROPOSITION 4.6 (crossing curve characterization). *The boundary of the stability region defined in Proposition* 4.3 *and depicted in Figure* 4.3 *arises from the correspondence $\wp_0(u) \mapsto \zeta_0(h, \delta)$. The necessary and sufficient condition forming this boundary in the (h, δ) domain is obtained only from the interval $u \in [0, \pi/2)$.*

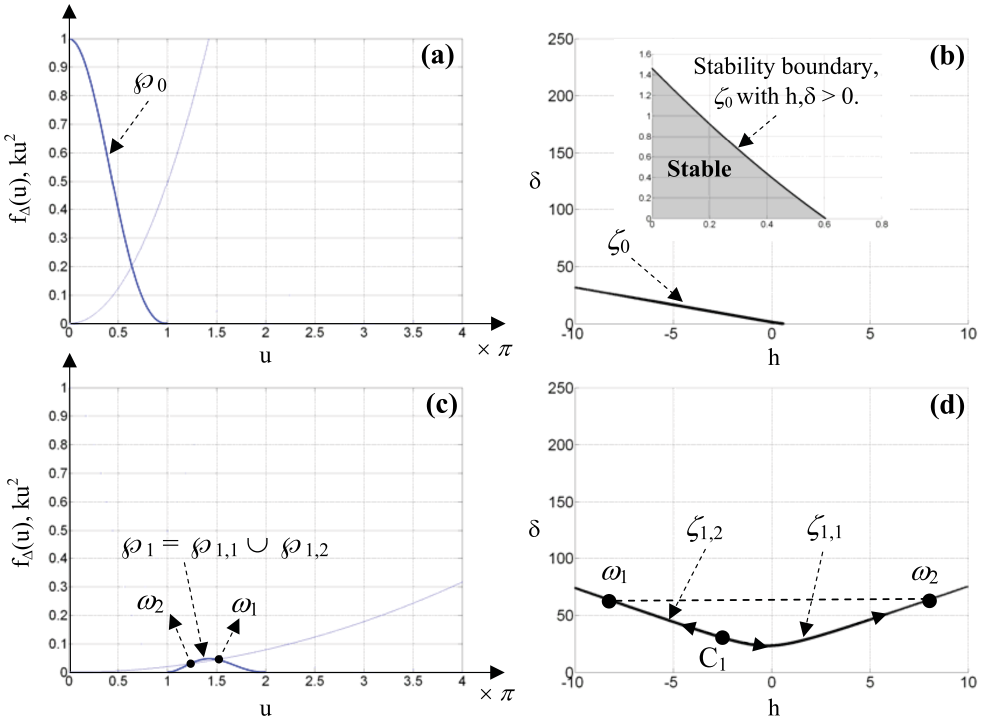


FIG. 4.3. (a)–(b) *Mapping of* $\wp_0$ *to* ζ_0. (c)–(d) *Mapping of* $\wp_1 = \wp_{1,1} \cup \wp_{1,2}$ *to* $\zeta_1 = \zeta_{1,1} \cup \zeta_{1,2}$ *(*$\wp_{1,i} \mapsto \zeta_{1,i}$*, for* $i = 1, 2$*). Arrows indicate the decreasing direction of* ω*, which attains its maximum on a respective* ζ_ℓ *curve at point* C_ℓ*,* $\ell > 0$.

Proof. (i) *Condition* $u \in [0, \pi/2)$. By Proposition 4.3, the stability boundary intersecting the h axis, $\bar{h}$, is created when $\delta = 0$. Since $\delta = 0$ corresponds to $u = 0$, starting from $u = 0$, tracing the intersection points of the parabola ku^2 and the curve $\wp_0$, one extracts the boundary of the single connected stability region in (h, δ). On this boundary, δ is monotonically increasing considering (4.7) and Proposition 4.8. To complete the proof, one should take into account the constraint of $h > 0$. Hence, δ should increase until h becomes zero. This occurs when the boundary of the stability region intersects the δ-axis at $\bar{\delta}$; see Proposition 4.3. The u value corresponding to $\bar{\delta}$ is $u = \omega\delta/2 < \pi/2$ as per (4.19).

(ii) *Necessary and sufficient condition.* There exists no u interval other than $u \in [0, \pi/2)$ that gives rise to a (h, δ) point on the stability boundary, since $f_\Delta(u)$ values attained in the interval $u \in [0, \pi/2)$ cannot be attained by any $u > \pi/2$. Hence $u \in [0, \pi/2)$ is necessary and sufficient for the computation of this boundary. □

REMARK 4.7. *Geometrically, the interval* $u \in [0, \pi/2)$ *corresponds to the segment* AF *in Figure* 4.2(a).

PROPOSITION 4.8 (local monotonicity property). *On the stability boundary, increasing* k *monotonically decreases* u *and* δ *solutions arising from the intersection points between* $f_\Delta(u)$ *and the parabola* ku^2.

Proof. From (4.10), $k = f_\Delta(u)/u^2$. The variation of k with respect to u is $\frac{dk}{du} = f'_\Delta(u)\frac{1}{u^2} - \frac{2}{u^3} f_\Delta(u)$. Since $f_\Delta(u) > 0$ and $f'_\Delta(u) < 0$ in the interval $u \in [0, \pi/2)$ defining the stability boundary, dk/du is negative. Recall that δ and k are inversely proportional; thus δ on the stability boundary is decreasing for increasing k. □

Readers are directed to the work in [11] as an example of the deployment of monotonicity ideas which gives rise to the characterization of the geometry of stability regions of a class of delay differential equations.

LEMMA 4.9. *dh/du on the stability boundary is nonpositive for all $u \in [0, \frac{\pi}{2})$.*

Proof. From (4.9), $\omega^2 = f_\Delta |\lambda_i|^2$. For $u \in [0, \pi/2)$, one can take $\omega = \sqrt{f_\Delta}|\lambda_i| = \sin(u)|\lambda_i|/u$ for studying the variations of h in (4.13). Note that the h variation on the stability boundary is obtained for $\ell = 0$; see Proposition 4.3. Hence, the variation of h along the stability boundary is found as

$$\frac{d}{du}\frac{(\phi_i - \pi/2 - u)u}{|\lambda_i|\sin(u)} = -\frac{A(u)\phi_i + B(u)}{2|\lambda_i|\sin^2(u)}, \tag{4.21}$$

where $A(u) = 2u\cos u - 2\sin u$ and $B(u) = 4u\sin u - u\pi\cos u - 2u^2\cos u + \pi\sin u$. Notice that $A(u) < 0$ for all $u \in (0, \pi/2)$ since $u < \tan(u)$ in this interval; thus the inequality in (4.21) becomes $A(u)\phi_i + B(u) > 0$, which can be rewritten as

$$-\frac{B(u)}{A(u)} > \phi_i \quad \forall u \in \left(0, \frac{\pi}{2}\right),$$

where in our case $\phi_i \in (\pi/2, 3\pi/2)$. Consequently, it is sufficient to prove that $-B(u)/A(u) > 3\pi/2$, that is,

$$\frac{2u - \pi}{u - \pi} - \frac{u}{\tan(u)} < 0 \quad \forall u \in \left(0, \frac{\pi}{2}\right).$$

The fact that the above inequality holds can be seen from Figure 4.4. Although an analytical proof can be given, the algebraic manipulations are rather involved and thus omitted. □

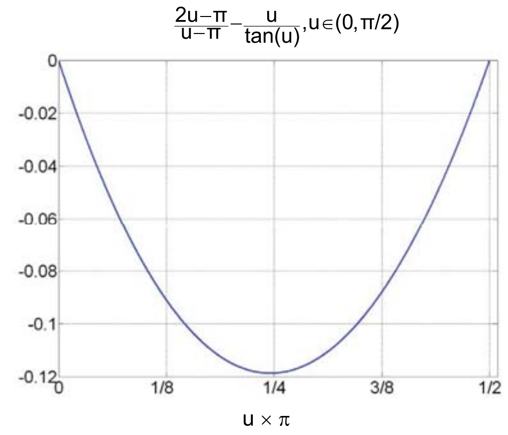


FIG. 4.4. *Plot of* $\frac{2u-\pi}{u-\pi} - \frac{u}{\tan(u)}$ *versus* u*,* $u \in (0, \pi/2)$*.*

REMARK 4.10. *By Proposition* 4.8 *and Lemma* 4.9*, increasing u corresponds to increasing δ and decreasing h. Hence, on the stability boundary, we have the property* $\frac{\partial\delta}{\partial u}\frac{\partial u}{\partial h} = \frac{\partial\delta}{\partial h} < 0$.

REMARK 4.11. *Since parameter u depicting the stability boundary is bounded as per Proposition* 4.6*, all ω satisfying the interconnection scheme belong to the set* $\Gamma_\omega = \{\omega \mid H_i(j\omega)\Delta(j\omega) - 1 = 0,\ u = \delta\omega/2,\ u \in [0, \pi/2)\}$. *As such, the imaginary roots on the stability boundary are given by the set $j\Gamma_\omega$. From* (4.9)*, one can obtain*

Γ_ω as $\Gamma_\omega = (\sqrt{2}|\lambda_i|/\pi, |\lambda_i|]$. This interval of ω is obviously a subset of the interval claimed in Proposition 4.2, since the interval of u creating the stability boundary is now restricted as $u \in [0, \pi/2)$.

Let us comment on the remaining part of the $\wp_0$ curve, $u \in (\pi/2, \pi)$. Following part (i) of the above proof, we state that corresponding h values obtained by u, $u \in (\pi/2, \pi)$, are all negative and thus ignored. Values of h on the curves obtained by the shifting $2\pi\ell/\omega$, $\ell = 1, 2, \ldots$, as per (4.13), however, may become positive, and thus should be carefully treated.

PROPOSITION 4.12 (crossing curves property). *The curves $\zeta_{\ell,\ell}$ and $\zeta_{\ell,\ell+1}$, $\ell \geq 1$, do not intersect the stability boundary formed by ζ_0.*

Proof. It is sufficient to prove $\inf_{\delta\in(\zeta_{\ell,\ell}\cup\zeta_{\ell,\ell+1})} \delta > \sup_{\delta\in\zeta_0} \delta$. From Proposition 4.8, it is clear that $\inf_{\delta\in(\zeta_{\ell,\ell}\cup\zeta_{\ell,\ell+1})} \delta$ arises from the $u = u_1$ value at point D_1 in Figure 4.2. As per Propositions 4.6 and 4.8, $\sup_{\delta\in\zeta_0} \delta$ is found at $u = \pi/2$ (point F in Figure 4.2). Since $u_1 > \pi/2$, using Proposition 4.8, one can see that the inequality $\inf_{\delta\in(\zeta_{\ell,\ell}\cup\zeta_{\ell,\ell+1})} \delta > \sup_{\delta\in\zeta_0} \delta$ holds. □

It is now proven that ζ_0 does not intersect with the remaining $\zeta_{\ell,i}$ ($i = \ell, \ell+1$, $\ell \geq 1$) segments; however, one also needs to show *stability properties* around the regions bordered by all the ζ curves in order to reveal the stability regions in the parametric domain of (h, δ). The following proposition helps achieve this.

PROPOSITION 4.13 (crossing direction along h, for δ fixed). *Given λ_i and δ, the crossing direction of the imaginary root(s) $s = j\omega$ along the h-axis is independent of the delays $h + \frac{2\pi}{\omega}\ell$, $\ell = 0, 1, \ldots$, creating these imaginary roots.*

Proof. The characteristic equation χ_i in (4.2) can be expressed as $\chi_i = s - \lambda_i e^{-sh}\Delta(s) = 0$ using (4.4). Then, the sensitivity expression at $s = j\omega$, after suppressing the arguments to conserve space, becomes

(4.22)

$$S(j\omega) = \left.\frac{ds}{dh}\right|_{s=j\omega} = \left(-\frac{\partial\chi_i}{\partial h}\left(\frac{\partial\chi_i}{\partial s}\right)^{-1}\right)_{s=j\omega} = -\left.\frac{s\lambda_i e^{-sh}\Delta}{1+\lambda_i e^{-sh}(h\Delta - \Delta')}\right|_{s=j\omega},$$

where $\Delta' = \frac{\partial\Delta}{\partial s}$. The above equation simplifies to

$$S(j\omega) = -\left.\frac{s\Delta}{\lambda_i^{-1}e^{sh} + h\Delta - \Delta'}\right|_{s=j\omega}. \tag{4.23}$$

Let $S(j\omega) = \frac{S_{NR}(\omega)+jS_{NI}(\omega)}{S_{DR}(\omega)+jS_{DI}(\omega)}$ with $S_{NR}, S_{NI}, S_{DR}, S_{DI} \in \mathbb{R}$. Then the real part of (4.23), which indicates the crossing direction of the $s = j\omega$ root across the imaginary axis, becomes

$$\Re\left(S(j\omega)\right) = \frac{S_{NR}S_{DR} + S_{NI}S_{DI}}{S_{DR}^2 + S_{DI}^2}. \tag{4.24}$$

Notice that if $\operatorname{sgn}(\Re\left(S(j\omega)\right)) = +1$ (or -1), this will indicate an imaginary axis crossing from left to right (or from right to left) half of the complex plane. Since the denominator of (4.24) is positive, it is dropped for studying the sign. Consequently, the numerator of (4.24) will determine the crossing direction, which becomes

$$\omega\left(\Im(\Delta)\left(h\Re(\Delta) - \Re(\Delta') + \Re\left(\frac{e^{jh\omega}}{\lambda_i}\right)\right) - \Re(\Delta)\left(h\Im(\Delta) - \Im(\Delta') + \Im\left(\frac{e^{jh\omega}}{\lambda_i}\right)\right)\right)$$

$$= \omega\left(\Im(\Delta)\left(-\Re(\Delta') + \Re\left(\frac{e^{jh\omega}}{\lambda_i}\right)\right) - \Re(\Delta)\left(-\Im(\Delta') + \Im\left(\frac{e^{jh\omega}}{\lambda_i}\right)\right)\right). \tag{4.25}$$

After some manipulations and dropping $\omega > 0$, this yields

$$\operatorname{sgn}\left(\Re\left(S(j\omega)\right)\right) = \operatorname{sgn}\left(\Im\left(\bar{\Delta}(j\omega)(\Delta'(j\omega) - \lambda_i^{-1}e^{jh\omega})\right)\right), \tag{4.26}$$

where $\bar{\Delta}(j\omega)$ is the complex conjugate of $\Delta(j\omega)$. Notice that the above equation is independent of h due to the exponential terms $e^{jh\omega}$ remaining unchanged for the selection of $h + \frac{2\pi}{\omega}\ell$, $\ell = 0, 1, \ldots$. □

REMARK 4.14. *The invariance feature of the root sensitivity expression in* (4.26) *enables an effective tool which helps reveal the stability/instability regions of the traffic dynamics in the entire* $(h, \delta) \in \mathbb{R}_+^2$ *parametric domain.*

We finally develop some ideas in the following borrowing from the implicit function theorem in order to study the smoothness and stability transition behavior around the ζ_0 curve in (h, δ) domain.

4.3. Local vs. global characterization. We elaborate on the geometry of ζ_0 from the implicit function theorem [14], which states that, on the imaginary axis, the characteristic equation $a(s, h, \delta)|_{s=j\omega} = H_i(j\omega)\Delta(j\omega) - 1 = 0$ may be used to locally express h and δ in terms of ω as an implicit function, in the form of $(h, \delta) = \varphi(\omega)$, if $da/d\omega \neq 0$ holds. This condition is nothing but the definition of the existence of *regular points* that allows the implementation of the implicit function theorem locally.

By the help of the interconnection scheme interpretation and the monotonicity ideas, the local representation can be extended. Following the theorem, one first classifies the points on ζ_0 in the (h, δ) domain. Such a classification identifies whether the implicit function theorem is applicable or not on the function $a(s, h, \delta) = 0$. Parameter δ is continuous on ζ_0, and its variation with respect to k is nonzero. Similar arguments also hold for h; see Proposition 4.6). The smoothness and nonzero derivative of the variables h on ζ_0 indicate that ζ_0 consists of only "regular points." This suffices to show that the theorem is applicable around *any* local point of ζ_0. By introducing the variable u along with the interconnection scheme, which only scales ω, we manage to separately express $h = \varphi_1(u)$ and $\delta = \varphi_2(u)$ *globally* on the stability boundary ζ_0.

So far we have shown the existence and the geometry of the stability region connected to the origin of the parametric domain (h, δ), intersecting the h and δ axis. In the following, the characterization of the stability/instability transition of the dynamics in (h, δ) using (4.2) is presented. For this objective, we use the idea based on the implicit function theorem [9], along with the separation of variables h and δ via the interconnection scheme that we have considered.

Take the characteristic equation $a(s, h, \delta) = H_i\Delta - 1 = 0$. When s moves along the imaginary axis, an (h, δ) pair moves along $\zeta_0(h, \delta)$. Let us first define the following for a given $\omega \in \Gamma_\omega$ on the $\zeta_0(h, \delta)$ curve:

$$R_0 = \Re\left(\frac{j}{s}\frac{\partial a(s, h, \delta)}{\partial s}\right)_{s=j\omega}, \quad R_1 = -\Re\left(\frac{1}{s}\frac{\partial a(s, h, \delta)}{\partial h}\right)_{s=j\omega},$$

$$I_0 = \Im\left(\frac{j}{s}\frac{\partial a(s, h, \delta)}{\partial s}\right)_{s=j\omega}, \quad I_1 = -\Im\left(\frac{1}{s}\frac{\partial a(s, h, \delta)}{\partial \tau_\nu}\right)_{s=j\omega},$$

$$R_2 = -\Re\left(\frac{1}{s}\frac{\partial a(s, h, \delta)}{\partial \delta}\right)_{s=j\omega}, \quad I_2 = -\Im\left(\frac{1}{s}\frac{\partial a(s, h, \delta)}{\partial \delta}\right)_{s=j\omega}.$$

By Lemma 4.1, $a(s,h,\delta)=0$ is an analytical function, and using the implicit function theorem, the tangent of $\zeta_0(h,\delta)$ is expressed as

$$\text{(4.27)} \qquad \begin{pmatrix} dh/d\omega \\ d\delta/d\omega \end{pmatrix} = \begin{pmatrix} R_1 & R_2 \\ I_1 & I_2 \end{pmatrix}^{-1} \begin{pmatrix} R_0 \\ I_0 \end{pmatrix} = \frac{1}{R_1I_2-R_2I_1} \begin{pmatrix} R_0I_2-I_0R_2 \\ I_0R_1-R_0I_1 \end{pmatrix},$$

provided that $R_1I_2-R_2I_1\neq 0$. In order to characterize the stability transition, one needs to consider h and δ as functions of $s=\mu+j\omega$. Since the tangent of $\zeta_0(h,\delta)$ along the positive direction (i.e., increasing ω direction) is $(dh/d\omega, d\delta/d\omega)$, the normal to the $\zeta_0(h,\delta)$ curve pointing to the left-hand side of the positive direction is $(-d\delta/d\omega, dh/d\omega)$. Also, as a pair of complex conjugate roots of $a(s,h,\delta)=0$ crosses the imaginary axis at $s=j\omega$ to $\mathbb{C}_+$, (h,δ) moves along the direction $(\partial h/\partial\mu, \partial\delta/\partial\mu)$. So, if the inner product of this vector with the normal vector is positive, i.e.,

$$\text{(4.28)} \qquad \left(\frac{\partial h}{\partial\omega}\frac{\partial\delta}{\partial\mu}-\frac{\partial\delta}{\partial\omega}\frac{\delta h}{\partial\mu}\right)_{s=j\omega} > 0,$$

then the region on the left of the stability curve $\zeta_0(h,\delta)$ at $s=j\omega$ has two more unstable roots than the right of $\zeta_0(h,\delta)$ curve. On the other hand, if the inner product is negative, then the region on the left of the stability curve $\zeta_0(h,\delta)$ has two fewer unstable roots than the region on its right. Similar to the tangency condition defined in (4.27), we can express $(\partial h/\partial\mu, \partial\delta/\partial\mu)$ as in the following:

(4.29)

$$\begin{pmatrix} \partial h/\partial\mu \\ \partial\delta/\partial\mu \end{pmatrix} = \begin{pmatrix} R_1 & R_2 \\ I_1 & I_2 \end{pmatrix}^{-1} \begin{pmatrix} I_0 \\ -R_0 \end{pmatrix} = \frac{1}{R_1I_2-R_2I_1} \begin{pmatrix} R_0R_2+I_0I_2 \\ -R_0R_1-I_0I_1 \end{pmatrix}.$$

Proposition 4.15. *Given any (h,δ) pair on the stability curve $\zeta_0(h,\delta)$, the inequality* (4.28) *is always satisfied.*

Proof. Simple manipulations show that $\left(\frac{\partial h}{\partial\omega}\frac{\partial\delta}{\partial\mu}-\frac{\partial\delta}{\partial\omega}\frac{\delta h}{\partial\mu}\right)_{s=j\omega}>0$ if

$$\text{(4.30)} \qquad R_2I_1-R_1I_2=\left(\frac{|\lambda_i|}{\omega}\right)^2\frac{4\sin^2(\delta\omega/2)-\delta\omega\sin(\delta\omega)}{(\delta\omega)^3}>0.$$

As per (4.7) and from Proposition 4.6, for $u\in(0,\pi/2)$, the inequality in (4.30) can be alternatively studied over

$$\text{(4.31)} \qquad 2\sin^2(u)-u\sin(2u)>0, \quad u\neq 0.$$

This inequality always holds since it can be rewritten in the form of a well-known trigonometric property, $\sin(u)/u>\cos(u)$, in the interval $u\in(0,\pi/2)$. For the case when $u=0$, we state that f_Δ is continuous and its derivative exists at $u=0$; therefore similar arguments on the smoothness of ζ_0 hold for $u=0$ as well. □

Remark 4.16. *Since the region below ζ_0 is known to be stable, Proposition* 4.15 *indicates that the region on the other side of ζ_0 has two unstable roots. Also, with the above proof, the smoothness of the stability curve ζ_0 is guaranteed.*

γ-distribution with and without a gap. We finally comment on the γ-distribution with and without gap. By (4.5), the characteristic equation is

$$\text{(4.32)} \qquad s(qs+1)^p-e^{-hs}\lambda_i=0.$$

When $h=0$, the above equation becomes a polynomial in s whose stability is easy to determine. At the origin of the parameter domain $(p,q)=(0,0)$, the characteristic equation becomes $s - \lambda_i = 0$, which is stable by Lemma 4.1. For $p \neq 0$ or $q \neq 0$, the stability can be studied by numerical computation of the roots or by using the jury test. In the example case studies, we will study the stability of (4.32) via numerical computations.

When the γ-distribution has a positive gap h, the analysis is more complicated. In the parametric domain of (p, q), the problem reduces to assuring the stability of the dynamics (2.2) with respect to λ_i and gap h. We study the stability in the (h, q) parametric domain by taking fixed p values. The procedure is as follows. Using the fact that eigenvalues of J are in complex conjugate pairs, one can reform the characteristic equation directly from (3.8) with real coefficients and perform the stability analysis for various selections of fixed p in the domain of (h, q). We mention that there exist various techniques in the literature to handle this analysis [17, 10, 22, 24].

5. Illustrative examples. In the following, some example cases are presented to demonstrate the memory effects on the stability of traffic flow dynamics and their physical interpretations. The developed theory equally allows one to study various scenarios such as the influence of the number of vehicles, presence of nonidentical drivers, aggressiveness of drivers, etc. In order to preserve coherence among the example cases, we will present the case when the drivers are identical (thus $\kappa_i = \kappa$) and compare the arising stability features of traffic dynamics with respect to aggressiveness of the drivers κ and the two spatial configurations given in Figure 2.2.

5.1. Uniform distribution. We take $\kappa_i = \kappa = 1.5$ and $\kappa_i = \kappa = 2$, respectively, which are in the same order of magnitude with those given in [2]. Figure 5.1(a) depicts the stability region for the ring configuration. The region shaded by light gray, Φ_1, represents the stability domain when $\kappa = 2$. When $\kappa = 1.5$, the stability region is enlarged by an additional region labeled as Φ_2 (dark grey); hence $\Phi_1 \cup \Phi_2$ becomes the stability region.

From Figure 5.1(a) we conclude that for both κ values the size of the memory window that is "tolerable" (to maintain stability) is the widest when $h = 0$. With a nonzero dead-time ($h \neq 0$) the allowable window size becomes narrower and eventually disappears.

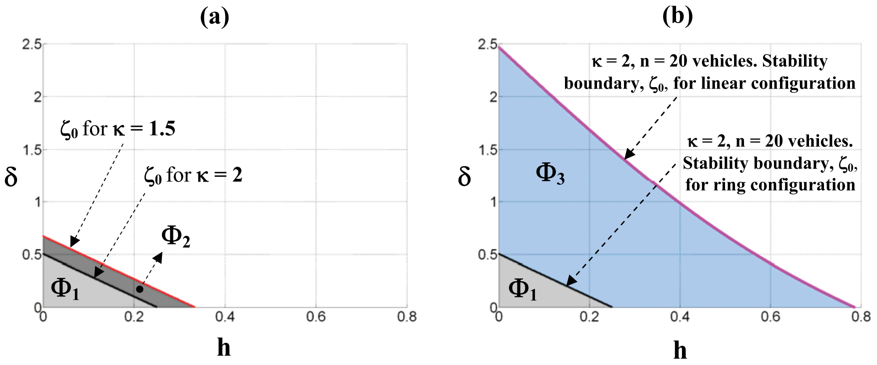


FIG. 5.1. (a) *The size of the memory window* δ *(sec) versus* h *(sec) for the system* (2.2) *with uniformly distributed delays* (2.3) *and* $n = 20$ *vehicles. The shaded area is the stability region.* (b) *Comparison of stability regions for ring and linear spatial configuration of vehicles, with* $\kappa = 2$ *and* $n = 20$ *vehicles.*

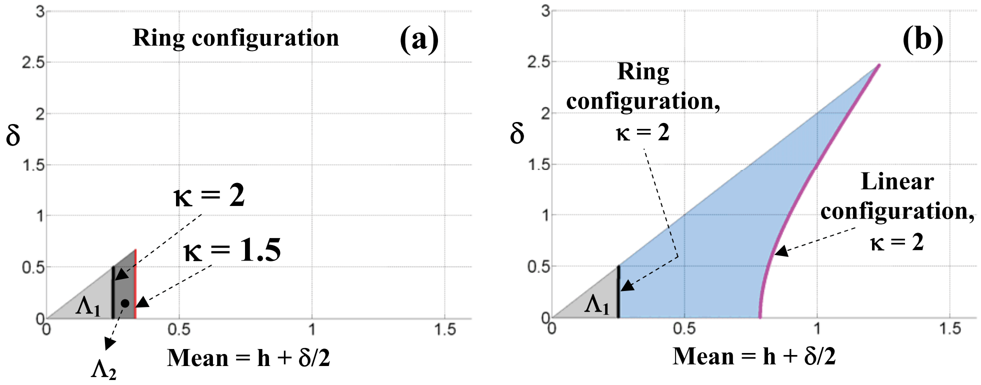


Fig. 5.2. *Stability regions in the parameters mean delay and memory size, for* $n = 20$ *vehicles.* (a) *Ring configuration for* $\kappa = 1.5$ *and* $\kappa = 2$, (b) *comparison between linear and ring configurations with* $\kappa = 2.0$.

5.1.1. Stability regions for ring and linear configurations. We now compare the stability regions for the different spatial configurations of the vehicles depicted in Figure 2.2. Figure 5.1(b) compares the stability regions for $n = 20$ vehicles. The regions labeled by Φ_1 in subfigures (a) and (b) are identical. Thus, the linear configuration offers an enhancement in the stability by the additional Φ_3 region.

The stability enhancement in the linear configuration can be explained mathematically. In a global sense, one can state a measure of the stability enhancement by how large $\bar{h}$ and $\bar{\delta}$ are. This can be easily checked by the results in Proposition 4.3. By Lemma 3.1, the eigenvalues of the configuration matrix J' are real in the linear configuration, implying $\phi = \pi$. For this setting of ϕ, it can be verified from (4.14) that $\bar{h}$ and $\bar{\delta}$ attain their maximum values: $\bar{h} = \frac{\pi}{2\kappa_{\max}}$ and $\bar{\delta} = \frac{\pi^2}{2\kappa_{\max}}$. This can be proven by assuming identical drivers with special form of λ_i as per (3.5), which can be used in (4.14) to show that as $\phi \to \pi$, the stability margins increase on both h and δ axes.

Physically, the linear configuration represents more degrees of freedom in the coupled dynamics, since the leading car is not restricted by the traffic, whereas in the ring configuration the motion of each vehicle plays a role in determining stability, thus limiting larger stability regions.

5.1.2. Stability with respect to mean delay versus memory size. We now present stability regions in the parameter domain *mean delay* $(h+\delta/2)$ versus *memory size* (δ). The mean of the uniform distribution represents the averaged effects of the memory, and it converges to the discrete delay case as the memory size approaches zero. Hence, the mean can also be seen as a link from discrete to distributed delays.

The traffic scenario is the same as in section 5.1. We first take a ring configuration of $n = 20$ vehicles with identical drivers and depict the stability regions in the new parametric domain. Thus, the stability pictures in Figure 5.1(a) correspond to those in Figure 5.2(a). The region Φ_1 is distorted in the new domain, $h \times \delta \to (h + \delta/2) \times \delta : \Phi_1 \mapsto \Lambda_1$, and similarly $(\Phi_1 \cup \Phi_2) \mapsto (\Lambda_1 \cup \Lambda_2)$. Figure 5.2(b) compares the stability regions for the ring and linear configurations, for $\kappa = 2.0$. The labeled regions correspond to those in Figure 5.1(b) as $\Phi_1 \mapsto \Omega_1$, and $(\Phi_1 \cup \Phi_2) \mapsto (\Omega_1 \cup \Omega_2)$.

Figure 5.2(a) shows how overaggressive drivers (large κ) cause instability, unless

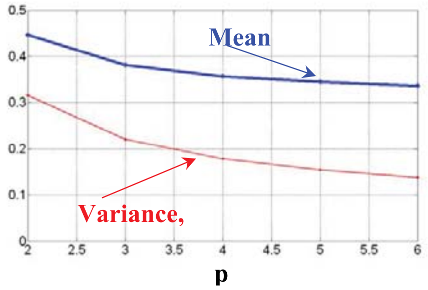


FIG. 5.3. (a) *Maximum allowable mean delay and the corresponding variance of the γ-distribution* ($h = 0$) *for varying p values.* (b) *Crossing boundary in mean vs. variance of the γ-distribution* ($h = 0$).

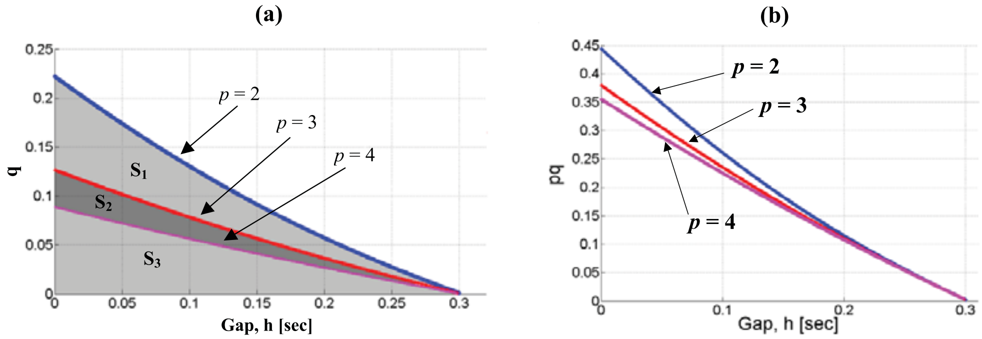


FIG. 5.4. (a) *Stability regions in* (h, q) *domain for various values of the parameter p of the γ-distribution with gap.* (b) *Corresponding stability boundaries in* (h, pq) *domain.*

the mean delay and the memory window are reduced, i.e., they can react almost instantaneously. This is analogous to feedback control systems where high gains may cause instability. Figure 5.2(b) shows that the linear configuration of the vehicles allows larger memory windows that can be utilized by the drivers without inducing instability. For this particular example, the allowable memory size (that preserves stability) in linear configuration of the vehicles is five times more than that of the ring configuration. An interesting observation from Figure 5.2(b) is that if the mean delay is relatively large (e.g., $h + \delta/2 = 1$ in linear configuration), stability is still possible, but too small or too large window sizes yield instability. In contrast, for smaller values of mean delay, the window size can even become zero, resuming the discrete delay case. This clearly shows the nontrivial qualitative effects of distributed delays.

5.2. γ-distribution with and without gap. We consider $n = 3$ vehicles and identical drivers with $\kappa = 2$. For $h = 0$ and $p \in [2, 6]$, we identify the maximum allowable q for which the characteristic equation (4.32) remains stable. Using these q values, the mean pq and the variance pq^2 of the distribution are plotted with respect to the parameter p in Figure 5.3. Notice that for smaller p the variance becomes larger, which corresponds to larger mean delay. This is also an indication of larger stability regions. In other words, increasing the variance of the delay distribution for a fixed mean delay enlarges the stability region.

In the case of a nonzero gap h, i.e., in the presence of a dead-time, the stability regions are shown in Figure 5.4(a) as p values. The shaded zones S_3, $S_2 \cup S_3$, and

$S_1 \cup S_2 \cup S_3$ correspond to stability regions for $p = 4$, $p = 3$, and $p = 2$, respectively. The subplots in Figure 5.4 have the same color coding. Note that the influence changing p on the stability region is more pronounced for small values of h than for larger values. We emphasize that, when modeling memory effects, the presence of a gap may not be negligible since the dynamics may become sensitive even to a small gap.

6. Conclusions. We have studied the stability of a single-lane microscopic car-following model in the parametric domain describing the delayed reactions of human drivers. In contrast to the literature, we have modeled such delayed reactions based on the "memory capabilities" of human drivers, which assumes that control actions are based on a "memory window" distributed over the the time history of the traffic flow dynamics. The resulting system with distributed delays offers a more realistic model, although the corresponding stability analysis becomes more difficult. In the analytical development of the paper, we have derived necessary and sufficient conditions for the stability of the traffic flow with distributed delays. Numerical examples have been given for two common delay distributions, two spatial configurations, and a realistic set of parameter values. The results show some nontrivial effects of distributed delays and reiterate that the modeling and analysis of traffic holds many mathematical challenges.

Acknowledgments. Part of this work was conducted at HeuDiaSyC (UMR CNRS 6599), Compiègne, France, in the academic year 2005–2006, during which Rifat Sipahi was supported by the Chateaubriand Scholarship of the French Government. R.S. and S.-I.N. are grateful to Martin Treiber, T.U. Dresden, Germany, for fruitful discussions on the dynamics of traffic flow, and acknowledge the financial support of Centre National de Recherche Scientifique (CNRS) for R.S.'s visit to F. M. Atay at Max Planck Institute on distributed delays and memory effects.

REFERENCES

[1] F. M. Atay, *Distributed delays facilitate amplitude death of coupled oscillators*, Phys. Rev. Lett., 91 (2003), paper 094101.

[2] M. Bando, K. Hasebe, K. Nakanishi, and A. Nakayama, *Analysis of optimal velocity model with explicit delay*, Phys. Rev. E, 58 (1998), pp. 5429–5435.

[3] E. Beretta and Y. Kuang, *Geometric stability switch criteria in delay differential systems with delay dependent parameters*, SIAM J. Math. Anal., 33 (2002), pp. 1144–1165.

[4] R. E. Chandler, R. Herman, and E. W. Montroll, *Traffic dynamics: Analysis of stability in car following*, Oper. Res., 7 (1958), pp. 165–184.

[5] R. Datko, *A procedure for determination of the exponential stability of certain differential-difference equations*, Quart. Appl. Math., 36 (1978), pp. 279–292.

[6] L. C. Davis, *Modifications of the optimal velocity traffic model to include delay due to driver reaction time*, Phys. A, 319 (2003), pp. 557–567.

[7] M. Green, *"How long does it take to stop?" Methodological analysis of driver perception-brake times*, Transportation Human Factors, 2 (2000), pp. 195–216.

[8] H. Greenberg, *An analysis of traffic flow*, Oper. Res., 7 (1959), pp. 78–85.

[9] K. Gu, S.-I. Niculescu, and J. Chen, *On stability crossing curves for general systems with two delays*, J. Math. Anal. Appl., 311 (2005), pp. 231–253.

[10] K. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Birkhäuser Boston, Cambridge, MA, 2003.

[11] J. K. Hale and W. Huang, *Global geometry of the stable regions for two delay differential equations*, J. Math. Anal. Appl., 178 (1993), pp. 344–362.

[12] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Modern Phys., 73 (2001), pp. 1067–1141.

[13] A. R. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[14] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, Undergrad. Texts in Math., Springer-Verlag, New York, 1980.

[15] W. Michiels and S.-I. Niculescu, *Stability and Stabilization of Time-Delay Systems: An Eigenvalue-Based Approach*, Adv. Des. Control 12, SIAM, Philadelphia, 2007.

[16] R. M. Murray, ed., *Control in an Information Rich World: Report of the Panel on Future Directions in Control, Dynamics, and Systems*, SIAM, Philadelphia, 2003.

[17] S.-I. Niculescu, *Delay Effects on Stability: A Robust Control Approach*, Lecture Notes in Control and Inform. Sci. 269, Springer-Verlag, Heidelberg, 2001.

[18] G. Orosz and G. Stepan, *Hopf bifurcation calculations in delayed systems with translational symmetry*, J. Nonlinear Sci., 14 (2004), pp. 505–528.

[19] G. Orosz, R. E. Wilson, and B. Krauskopf, *Global bifurcation investigation of an optimal velocity traffic model with driver reaction time*, Phys. Rev. E, 70 (2004), paper 026207.

[20] G. Orosz, B. Krauskopf, and R. E. Wilson, *Bifurcations and multiple traffic jams in a car-following model with reaction time delay*, Phys. D, 211 (2005), pp. 277–293.

[21] R. W. Rothery, in Traffic Flow Theory, 2nd ed., TRB Special Report Volume 165, N. H. Gartner, C. J. Messner, and A. J. Rathi, eds., Transportation Research Board (TRB), Washington, DC, 1992, Chapter 4.

[22] R. Sipahi and N. Olgac, *Complete stability robustness of third-order LTI multiple time-delay systems*, Automatica, 41 (2005), pp. 1413–1422.

[23] R. Sipahi and S.-I. Niculescu, *Analytical stability study of a deterministic car following model under multiple delay interactions*, in Proceedings of the 6th IFAC Time Delay Systems Workshop, L'Aquila, Italy, 2006, International Federation of Automatic Control/Elsevier, Oxford, UK, 2006.

[24] G. Stepan, *Retarded Dynamical Systems: Stability and Characteristic Function*, Longman Scientific & Technical/John Wiley & Sons, New York, 1989.

[25] O. Toker and H. Ozbay, *Complexity issues in robust stability of linear delay-differential systems*, Math. Control Signals Systems, 9 (1996), pp. 386–400.

[26] M. Treiber and D. Helbing, *Memory effects in microscopic traffic models and wide scattering in flow-density data*, Phys. Rev. E, 68 (2003), paper 046119.

[27] M. Treiber, A. Kesting, and D. Helbing, *Delays, inaccuracies and anticipation in microscopic traffic models*, Phys. A, 360 (2006), pp. 71–88.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 760–783

SHOCK SOLUTIONS FOR PARTICLE-LADEN THIN FILMS*

BENJAMIN P. COOK†, ANDREA L. BERTOZZI†, AND A. E. HOSOI‡

Abstract. We derive a lubrication model describing gravity-driven thin film flow of a suspension of heavy particles in viscous fluid. The main features of this continuum model are an effective mixture viscosity and a particle settling velocity, both depending on particle concentration. The resulting equations form a 2×2 system of conservation laws in the film thickness $h(x, t)$ and in ϕh, where $\phi(x, t)$ is the particle volume fraction. We study flows in one dimension under the constant flux boundary condition, which corresponds to the classical Riemann problem, and we find the system can have either double-shock or singular shock solutions. We present the details of both solutions and examine the effects of the particle settling model and of the microscopic length scale b at the contact line.

Key words. thin liquid films, contact lines, Riemann problem, systems of conservation laws, suspension flow, sedimentation, singular shocks

AMS subject classifications. 76D08, 76D07, 35L65, 76T20

DOI. 10.1137/060677811

1. Introduction.

The flow of thin viscous films [44] is of great importance to many problems in science and engineering. Flows in thin films can result from slow processes such as spreading [17] and evaporation [6] or stronger driving forces such as capillarity [38] or gravity [21]. As shown by Huppert in [21], gravity-driven films on an incline can be described roughly by the conservation law

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x} h^3 = 0 \tag{1.1}$$

for the film thickness h. However, in many cases, there exist large gradients in h and dry regions where $h = 0$, requiring more complex models that incorporate capillary forces and the thermodynamic wetting process.

Wetting occurs as a fluid domain evolves, moving in particular the contact line, where the solid, liquid, and vapor phases meet. Despite the fundamental importance of contact lines to fluid dynamic boundary conditions, many of their basic properties are not fully understood [2, 13]. The standard no-slip boundary condition is inadequate near a moving contact line [14, 20], and two common contact line models either allow a small slip velocity [20] or assume a thin "precursor" film rather than a dry substrate [55]. These models have contributed to understanding the capillary ridge that often develops near the contact line [3, 16, 19, 23, 28, 55], the rupture of thin films [39, 57], the contact angle that the free surface makes with the substrate [16, 23, 25, 50], and the relevance of the material composition of the fluid and substrate [12, 24, 50]. A "fingering" instability observed in [21] which deforms the contact line in

*Received by the editors December 15, 2006; accepted for publication (in revised form) September 19, 2007; published electronically January 4, 2008.
http://www.siam.org/journals/siap/68-3/67781.html

†Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095 (bencook@math.ucla.edu, bertozzi@math.ucla.edu). The first and second authors were supported by NSF VIGRE grant DMS-0502315, NSF grants ACI-0321917 and DMS-0244498, and ONR grant N000140410078.

‡Hatsopoulos Microfluids Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (peko@mit.edu). This author was supported by NSF grant ACI-0323672.

some driven films has also motivated analysis [3, 55], simulation [28], and experiments [12, 24, 50] on thin film flow.

Film flows of more complex materials are much less understood. For dry granular flows, air can frequently be neglected, and the central modeling challenge is to determine an appropriate constitutive relation [26]. Replacing the fluid with a suspension, however, introduces the possibility of phase segregation, allowing new behaviors not seen in pure fluids and only recently observed in film flows [54, 59]. Segregation of viscous suspensions can be driven by gravity [59], but has also been observed with neutrally buoyant particles in thin films [54] and in the related Hele–Shaw flow [53]. Direct numerical simulation of suspension flows considering individual particles is computationally demanding, and existing methods do not account for the complexities of a contact line [49, 58]; consequently, continuum models play an essential role in understanding these flows.

Continuum descriptions of viscous suspensions involve three main effects: an "effective viscosity" greater than that of the suspending fluid [30, 52], the settling of heavy particles due to gravity [11], and particle fluxes thought to result from particle interactions in the presence of shear [34]. Various models have incorporated some or all of these effects [40, 42, 45, 47]; however, only a limited number of flow geometries have been studied, most commonly the one-dimensional Couette and Poiseuille flows (for exceptions see [15]). In particular Schaflinger, Acrivos, and Zhang [47] model a gravity-driven thin film, though they do not consider variation in the flow direction caused by gravitational segregation that we model below.

Our work is motivated by the experiment and model described by Zhou et al. in [59]. The experiment consists of a gravity-driven film of a dense ($\geq 17\%$ by volume) suspension of glass beads in oil which flows down an inclined plane under constant flux upstream conditions. They observed three different particle behaviors in this experiment, depending on the inclination angle and particle concentration of the initially well-mixed suspension, which they summarized in a phase diagram. At low inclination angles and concentrations, the particles settle out of the flow, leaving a film of clear fluid, while at intermediate angles and concentrations the suspension appeared well mixed for the duration of the experiment. At high angles and concentrations the particles accumulate near the moving contact line in a "particle-rich ridge." They also observed that while the well-known fingering instability [21] occurs in the first two regimes, it is largely suppressed when the ridge appears. Their new model describes this third regime, characterized by spatially varying rheology, which appears to have no analogue in pure fluid motion.

Zhou et al. derived their model by treating the mixture locally as a Newtonian fluid, which allows the use of standard lubrication techniques. The two-phase flow is described by an overall velocity determined from the local value of a concentration-dependent effective viscosity, and a settling velocity of the heavy particles relative to the fluid. They derived a system of two coupled fourth-order evolution equations for the film thickness and particle concentration, and argued that the essential dynamics are determined by a system of conservation laws obtained by retaining only the first-order terms. They also presented double-shock solutions for this system depending on the parameter b appearing in their contact line model, which represents the thickness of a precursor film appearing ahead of the bulk flow. They compared these solutions to the experimentally observed ridge, and noted that the calculated speeds of the two shocks become nearly equal at small values of b. Their calculations, however, were not sufficient to determine whether the shock speeds actually coincide at some finite $b^* > 0$, an important issue since this would call into question the existence of solutions for $b < b^*$. Furthermore, they described the physical derivation and the shock solutions only briefly.

The purpose of the present work is to give a more complete derivation of this model describing the ridge regime, and to more thoroughly characterize its shock solutions, including their dependence on b. Aspects of these solutions motivate a revision of the particle settling model, which appears later in the manuscript. In section 2 we present a full derivation following the assumptions of Zhou et al., which was not included in their work. While the equations we derive are slightly different, they appear to have the same qualitative behavior. As in [59] we present two forms of the model equations. The "full system" including fourth-order terms due to surface tension is beyond the scope of our subsequent analysis, but nonetheless important for a faithful description and for modeling phenomena such as the capillary ridge and the fingering instability. The first-order "reduced system," which we study in section 4, is expected though not guaranteed to approximate the full system away from the contact line. In section 3 we recall the classical theory for hyperbolic systems of conservation laws, and in section 4 we apply these methods to the reduced system. For double-shock solutions, we find the two shock speeds do become equal at a certain precursor thickness b, below which the equations have no classical solution. In section 5 we compare this case to the mathematical theory of singular shocks, in which a delta mass is concentrated at the discontinuity. We study one approximate singular shock solution and find the particle concentration exceeds the limit of close packing, suggesting this form of the model is inaccurate at high concentrations. We propose a modified form for the settling velocity in section 6 which causes both the particle and fluid velocities to vanish at close packing, and we find the resulting equations appear to be well-posed for all precursor thicknesses. We summarize our results in section 7, concluding that the modified equations appear more realistic, though a comparison with the fourth-order system and/or with experiments is still needed to establish their ultimate validity.

2. Derivation. Two common methods for describing binary mixtures in a continuum framework are the "two-fluid" and the "mixture" or "one-fluid" models [56]. The two-fluid model balances forces on the two components separately, with the forces of interaction appearing explicitly as a function of the two velocities. It therefore requires a separate viscosity for each phase. The one-fluid model balances forces on the mixture as a whole, using an effective viscosity, and postulates a form for the relative velocity between the two components. Since empirical formulae are readily available for the effective mixture viscosity and settling velocity, we follow Zhou et al. in using the one-fluid model. We also note that the fluid and particle velocities are nearly equal, so the one-fluid equations describing average and relative velocities can be expected to be less strongly coupled than their two-fluid counterparts.

Deriving a one-fluid model involves balancing forces first for the mixture as a whole, without regard to interactions between the two components. In the present case inertia is negligible, and these forces are just gravity and viscous stress. We use an empirical expression for the latter in which the mixture is considered a Newtonian fluid, with an effective viscosity depending on the particle volume fraction ϕ. For a fluid of kinematic viscosity μ_f one form for this relation is [30, 52]

$$\mu(\phi) = \mu_f(1 - \phi/\phi_m)^{-2}, \tag{2.1}$$

where $\phi_m \approx 0.67$ is the random packing fraction of spheres. This viscosity leads to a stress tensor of the form

$$\mathbf{\Pi} = p\mathbf{I} - \frac{1}{2}\mu(\phi)\left[\nabla\mathbf{v} + (\nabla\mathbf{v})^T\right], \tag{2.2}$$

where p is the fluid pressure and $\mathbf{v}$ is a velocity characterizing the motion of the mixture. Since the two mixture components in general have different velocities, say v_f and v_p for the fluid and particulate phases, respectively, $\mathbf{v}$ must be some average of the two. Much of the experimental literature deals with neutrally buoyant mixtures, in which the two velocities are the same and the distinction is unnecessary, but in this case the question is relevant. We argue that since the constitutive model involves neither inertia nor gravity, it should be independent of the masses of the two phases. Therefore we select the volume-averaged velocity: defining

$$\mathbf{v} = (1-\phi)\mathbf{v_f} + \phi\mathbf{v_p}, \qquad \mathbf{v_{rel}} = \mathbf{v_p} - \mathbf{v_f} \tag{2.3}$$

thus comprises the one-fluid model for the mixture, and the individual phase velocities can be recovered by

$$\mathbf{v_p} = \mathbf{v} + (1-\phi)\mathbf{v_{rel}}, \qquad \mathbf{v_f} = \mathbf{v} - \phi\mathbf{v_{rel}}. \tag{2.4}$$

The average velocity satisfies the Stokes equations:

$$\nabla\cdot\mathbf{\Pi} = \rho(\phi)\mathbf{g}, \qquad \nabla\cdot\mathbf{v} = 0, \tag{2.5}$$

where $\rho(\phi)$ is the mixture density and $\mathbf{g}$ is the gravitational field. The density is given by $\rho(\phi) = \rho_f(1+\Delta\phi)$, where $\Delta = (\rho_p - \rho_f)/\rho_f$ and ρ_f and ρ_p are the densities of the fluid and particulate phases.

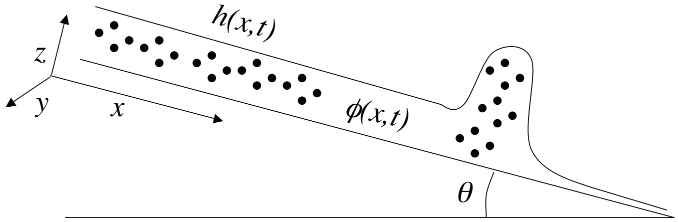


FIG. 2.1. *Geometry of the film problem. While our derivation will allow* y *dependence, we study the* y*-independent case. Our model assumes* ϕ *is independent of* z*.*

We now define the problem geometry as in Figure 2.1, considering an advancing film that coats a plane inclined at the angle θ. In deriving the equation for $\mathbf{v}$, we follow the standard methods used for pure fluid films [17, 44]. The lubrication approximation, valid at small Reynolds numbers and geometric aspect ratios, assumes $\mathbf{v}$ lies in the $\mathbf{x}$-$\mathbf{y}$ plane and $\left|\frac{\partial\mathbf{v}}{\partial z}\right| \gg \max\left(\left|\frac{\partial\mathbf{v}}{\partial x}\right|, \left|\frac{\partial\mathbf{v}}{\partial y}\right|\right)$. Correspondingly, we now consider all velocities to be two-dimensional vectors, as well as the gradient $\nabla = \mathbf{x}\frac{\partial}{\partial x} + \mathbf{y}\frac{\partial}{\partial y}$, and define $g_\perp = \mathbf{g}\cdot\mathbf{z} = |\mathbf{g}|\cos\theta$ and $\mathbf{g}_\parallel = \mathbf{g} - g_\perp\mathbf{z} = (|\mathbf{g}|\sin\theta)\mathbf{x}$. In this notation, the Stokes equations now read

$$\frac{\partial p}{\partial z} = -\rho(\phi)g_\perp, \tag{2.6a}$$

$$\nabla p = \mu(\phi)\frac{\partial^2\mathbf{v}}{\partial z^2} + \rho(\phi)\mathbf{g}_\parallel. \tag{2.6b}$$

The Laplace–Young boundary condition states that the pressure at the free surface, $z = h(x,y)$, is given by

$$p\left(x,y,h(x,y)\right) = -\gamma\nabla^2 h(x,y), \tag{2.7}$$

where γ is the coefficient of surface tension. The pressure is then determined by

$$p(x,y,z) = -\gamma\nabla^2 h(x,y) + \int_z^{h(x,y)} \rho(\phi(x,y,z'))g_\perp dz' \tag{2.8}$$

from the depth and particle concentration of the film. Here it is convenient to assume the particle concentration is independent of the z coordinate, so that the integral in (2.8) is merely $\rho(\phi)g_\perp(h-z)$. We will discuss this assumption further in our treatment below of particle motion.

Combining (2.6b) and (2.8) and defining $P(x,y) = -\gamma\nabla^2 h + \rho(\phi)g_\perp h$, we have

$$\nabla P - zg_\perp\rho'(\phi)\nabla\phi = \mu(\phi)\frac{\partial^2\mathbf{v}}{\partial z^2} + \rho(\phi)\mathbf{g}_\parallel. \tag{2.9}$$

The boundary conditions of interest are no stress ($\partial\mathbf{v}/\partial z = 0$) at the free interface and no slip ($\mathbf{v} = 0$) at the solid interface. Equation (2.9) can now be integrated twice in z with the constants of integration determined by these conditions, to arrive at the equation

$$\mu(\phi)\mathbf{v} = \left(hz - \frac{z^2}{2}\right)(\rho(\phi)\mathbf{g}_\parallel - \nabla P) + \frac{1}{2}(h^2 z - z^3/3)g_\perp\nabla\rho(\phi) \tag{2.10}$$

for the volume-averaged velocity. Integrating once more gives the depth-averaged velocity

$$\mathbf{v_{av}} = \frac{h^2}{3\mu(\phi)}\left[\gamma\nabla\nabla^2 h - g_\perp\left(\nabla(\rho(\phi)h) - \frac{5}{8}h\nabla\rho(\phi)\right) + \rho(\phi)\mathbf{g}_\parallel\right]. \tag{2.11}$$

Modeling the relative velocity due to particle settling turns out to be more difficult. Recall that in the above lubrication analysis, we have assumed the particles are evenly distributed across the film depth. This may seem unrealistic because the normal component of gravity is pulling the particles toward the solid substrate, but this model is concerned with the particle-rich ridge regime occurring at high angles and concentrations, in which Zhou et al. found that particles do not settle out of the flow. A similar effect was also observed in a thin film experiment [54] performed by Timberlake and Morris with neutrally buoyant particles: they found higher concentrations near the free surface, and attributed this to a shear-induced particle flux such as Leighton and Acrivos describe in [34]. This flux consists of a nonlinear diffusion in the presence of shear, and in inhomogeneous flows an additional migration of particles away from regions of high shear. Schaflinger, Acrivos, and Zhang, in their model for film flow [47], balance gravity-driven settling with only the diffusive flux, and find steady state solutions in which the concentration increases with depth. While the corresponding problem including both diffusion and migration remains unsolved, Carpen and Brady found nonmonotone concentration profiles in a model for the related inclined Poiseuille flow [9] and also showed that these profiles are unstable due to heavy material suspended above lighter material. Thus it is unclear whether the actual concentration profile for film flow increases or decreases with depth, so we find it reasonable to consider the simplest case, a uniform depth profile.

We begin our model of the relative motion with the settling velocity

$$\mathbf{v_s} = \frac{2a^2\Delta\rho_f\mathbf{g}_\parallel}{9\mu_f} \tag{2.12}$$

of a single sphere of radius a in $\mathbb{R}^3$, while noting that this expression neglects the effects of the solid boundary, the free surface, and other particles. Equation (2.12) uses $\mathbf{g}_\|$ because particles were not observed to settle vertically in the ridge regime. The problem of determining settling rates of concentrated mixtures is complex, even in the idealized case of monodisperse spherical particles in a large domain [11]. Some of the challenges are summing the interactions between spheres, which decay only as $1/r$ in Stokes flows, and interpreting theoretical results that imply divergent fluctuations about the mean particle velocity [5, 8]. Since there is no general agreement of theoretical and numerical results with experiments, sedimentation is commonly modeled by an empirical hindered settling function,

$$\mathbf{v_{rel}} = f(\phi)\mathbf{v_s}, \tag{2.13}$$

such as the Richardson–Zaki function (see [46])

$$f_{RZ}(\phi) = (1-\phi)^n, \qquad n \approx 5. \tag{2.14}$$

We also seek a correction to represent the impeding effect of the solid substrate on particle motion. A similar problem involving a sphere falling next to a vertical wall has been solved approximately by the method of images [18], leading to the series solution

$$\mathbf{v_{rel}} = \left(1 - \frac{259}{256}\left(\frac{a}{z}\right) + \frac{9}{16}\left(\frac{a}{z}\right)\log\left(\frac{a}{z}\right) - \frac{1}{16}\left(\frac{a}{z}\right)^3 + \frac{15}{256}\left(\frac{a}{z}\right)^4 + \cdots\right)\mathbf{v_s} \tag{2.15}$$

for the velocity, where $z > a$ is the distance from the center of the particle to the wall. The important quantity in a lubrication model is the depth-averaged velocity, which in the case of particle settling can be interpreted as $(1/h)\int_a^h \mathbf{v}(z)dz$. Figure 2.2 shows this average for a range of the nondimensional parameter h/a, along with the simpler function that we will use to approximate wall effects:

$$w(h) = \frac{A(h/a)^2}{\sqrt{1 + \left[A(h/a)^2\right]^2}} \tag{2.16}$$

with $A = 1/18$. This function has the desired properties $w \approx 0$ for $h < a$, $w \approx 1$ for $h \gg a$, and unlike (2.15) is differentiable and positive on $(0, \infty)$. We have chosen the parameter A so that this function resembles (2.15), but since the latter neglects the net flow and the effects of other particles it should mainly be viewed as a correction to ensure $\mathbf{v_{rel}} \to 0$ for very thin films.

For lack of a comprehensive theory incorporating both wall effects and hindered settling, we simply assume the effects are multiplicative, obtaining the settling velocity

$$\mathbf{v_{rel}} = f(\phi)w(h)\mathbf{v_s} \tag{2.17}$$

relative to the fluid which we interpret as a depth average. We assume f refers to f_{RZ} until section 6, when we consider another settling function. The velocities $\mathbf{v_{rel}}$ in (2.17) and $\mathbf{v_{av}}$ in (2.11) complete the one-fluid description (2.3). The evolution equations

$$\frac{\partial h}{\partial t} + \nabla\cdot(h\mathbf{v_{av}}) = 0, \qquad \frac{\partial \phi h}{\partial t} + \nabla\cdot(\phi h\mathbf{v_p}) = 0 \tag{2.18}$$

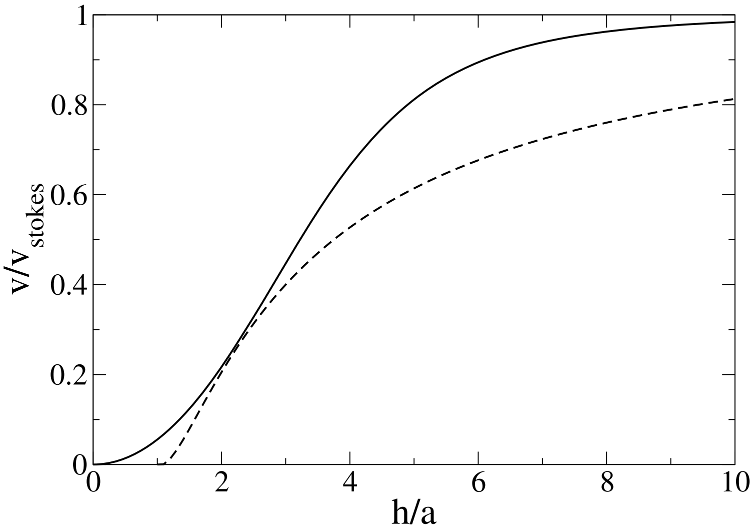


FIG. 2.2. *Our correction representing the impeding effect of the solid boundary on a single particle's settling velocity (solid), and the depth average of* (2.15) *(dashed).*

follow from conservation of volume in both the mixture as a whole and, using (2.4), in the particulate phase. Note that (2.18) differs from the model proposed by Zhou et al., which incorrectly used conservation of mass. Identifying $\mathbf{v_{av}}$ in (2.18) with $\mathbf{v}$ in (2.4) and inserting (2.11) and (2.17) into (2.18) then gives the complete system

$$\text{(2.19a)}\quad \frac{\partial h}{\partial t} + \nabla\cdot\left(\frac{h^3}{3\mu(\phi)}\left[\gamma\nabla\nabla^2 h - g_\perp\left(\nabla(\rho(\phi)h) - \frac{5}{8}h\nabla\rho(\phi)\right) + \rho(\phi)\mathbf{g}_\parallel\right]\right) = 0,$$

$$\frac{\partial(\phi h)}{\partial t} + \nabla\cdot\left(\frac{\phi h^3}{3\mu(\phi)}\left[\gamma\nabla\nabla^2 h - g_\perp\left(\nabla(\rho(\phi)h) - \frac{5}{8}h\nabla\rho(\phi)\right) + \rho(\phi)\mathbf{g}_\parallel\right]\right.$$

$$\text{(2.19b)}\quad \left. + \phi h(1-\phi)f(\phi)w(h)\mathbf{v_s}\right) = 0.$$

Next we nondimensionalize the equations for the constant flow rate problem, with the rescaling used in [3] for a clear fluid. If the upstream gate height h_0 represents a typical film thickness, then the first- and fourth-order terms in (2.19) are comparable at a length scale $x_0 = (\ell^2 h_0)^{1/3}$, where $\ell = \sqrt{\gamma/\rho_f g_\parallel}$ is the capillary length. The time derivative is on the same scale as well if $t \sim t_0 = (3\mu_f/\gamma)x_0\ell^2/h_0^2$, and the corresponding capillary number is $Ca \equiv \mu_f x_0/\gamma t_0 = h_0^2/3\ell^2$. Defining $\tilde{h} = h/h_0$, $\tilde{x} = x/x_0$, $\tilde{t} = t/t_0$, $\tilde{\rho}(\phi) = 1 + \Delta\phi$, $\tilde{\mu}(\phi) = (1-\phi/\phi_m)^{-2}$, $\tilde{w}(\tilde{h}) = w(h)$, and dropping the tildes, and replacing ∇ with $\partial/\partial x$ in anticipation of a y-independent solution, we obtain the dimensionless system

$$\text{(2.20a)}\quad \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}\left(\frac{h^3}{\mu(\phi)}\left[h_{xxx} - D(\theta)\left((\rho(\phi)h)_x - \frac{5}{8}h\rho(\phi)_x\right) + \rho(\phi)\right]\right) = 0,$$

$$\frac{\partial(\phi h)}{\partial t} + \frac{\partial}{\partial x}\left(\frac{\phi h^3}{\mu(\phi)}\left[h_{xxx} - D(\theta)\left((\rho(\phi)h)_x - \frac{5}{8}h\rho(\phi)_x\right) + \rho(\phi)\right]\right.$$

$$\text{(2.20b)}\quad \left. + v_s\phi h(1-\phi)f(\phi)w(h)\right) = 0.$$

Note that the inclination angle θ has been scaled out, and now appears only in the parameter $D(\theta) = (3Ca)^{1/3} \cot \theta$ measuring the relative importance of the second-order terms.

Before introducing the boundary conditions, a discussion is necessary of the microscopic contact-line physics, for which we rely on literature dealing with pure fluids. It has been shown [14] that the no-slip boundary condition we have employed above requires infinite viscous energy dissipation in the vicinity of a moving contact line. This singularity is removed if the fluid-solid boundary condition is modified to allow finite slip [20], which generally takes the form

$$\mathbf{v}|_{z=0} = b \left.\frac{\partial \mathbf{v}}{\partial z}\right|_{z=0}, \tag{2.21}$$

where b is a length on the order of the molecular size. This slip length has been observed experimentally and is known to be particularly large (on the order of microns [33]) for polymer liquids such as the PDMS used in [59]. Another technique used to model the contact line derives from attractive Van der Waals forces between the fluid and solid, which for many wetting films (again including PDMS on acrylic) causes a precursor film of microscopic thickness to extend ahead of the apparent contact line [1, 10]. Modeling this precursor explicitly is a complex thermodynamic problem at a microscopic length scale; however, it has been shown that the effect of the precursor on the macroscopic fluid problem can be approximated by incorporating this length scale into the fluid boundary condition [13]. Models that simply impose a thickness b for the precursor have been used successfully [3, 55] and been seen to give similar predictions to the slip model with the same value of b [41, 51]. In both cases b is difficult to know precisely and is often treated as an unknown parameter. Meaningful values range from perhaps 100 μm for a prewet surface down to 1 nm for a smooth, dry surface.

In this work, we choose the precursor model because it preserves the symmetry of the Riemann problem, discussed in section 3 below, and will present results for a range of thicknesses. With an inflow at concentration ϕ_L and a precursor of nondimensional thickness $b \ll 1$ and concentration ϕ_R, the initial conditions for the constant flow rate problem are

$$(h, \phi)|_{t=0} = \begin{cases} (1, \phi_L) & \text{if} \quad x < 0, \\ (b, \phi_R) & \text{if} \quad x > 0. \end{cases} \tag{2.22}$$

In addition to b, ϕ_R is also a model parameter not determined by the bulk flow, and must be specified. The appropriate value of ϕ_R may vary: for a prewet surface it may be equal to ϕ_L, while in a microscopic precursor it is probably zero. We mainly consider $\phi_R = \phi_L$ for definiteness, but also discuss $\phi_R = 0$.

The large-scale behavior of lubrication equations such as (2.20) is often well described by the corresponding first-order system, obtained by simply dropping all higher-order terms. This reduced system,

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}\left(h^3 \rho(\phi)/\mu(\phi)\right) = 0, \tag{2.23a}$$

$$\frac{\partial(\phi h)}{\partial t} + \frac{\partial}{\partial x}\left(\phi h^3 \rho(\phi)/\mu(\phi) + v_s \phi h(1-\phi) f(\phi) w(h)\right) = 0, \tag{2.23b}$$

corresponds to a rescaling of (2.19) with $x \gg x_0$ and $t \gg t_0$; however, we study it below because it allows solutions to be understood as simple shocks and rarefactions, while retaining the essential convective dynamics.

Zhou et al. [59] presented numerical evidence that first-order and fourth-order models agree well for this problem. This agreement is also seen in the homogeneous case ($\phi \equiv 0$, or $\phi \equiv \phi_0 > 0$ with $a = 0$), where (2.23) reduces to (1.1). The Riemann problem for that equation features simple shock solutions; however, it was first studied by Huppert with Dirac mass initial data, which leads to rarefaction-shock solutions (defined in section 3) that compare favorably to a constant-volume experiment [21]. Such correspondence between full and reduced systems is not guaranteed, however: Bertozzi, Münch, and Shearer [4] studied a variant of (1.1) in which Marangoni forcing competes with gravity, and they described examples of more complex shock structures for which the first-order and fourth-order solutions do not agree. Similar lubrication models have given rise to pairs of equations describing a thin film containing surfactant [22, 35]. Also related are models for sedimenting mixtures in which the particle concentration exhibits kinematic shocks [31].

3. The Riemann problem for systems of conservation laws. This section reviews the theory of systems of nonlinear conservation laws in one dimension, of which (2.23) is an example. This class contains equations of the form

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x} F(U) = 0, \qquad U, F(U) \in \Omega \subset \mathbb{R}^n. \tag{3.1a}$$

Although initial-value problems for (3.1a) are not in general well-posed, there is a large body of analytical techniques for finding and characterizing solutions when they exist [32]. The analysis is especially simplified for the Riemann problem, in which the initial data is a step function

$$U(x,0) = \begin{cases} U_L & \text{if} \quad x < 0, \\ U_R & \text{if} \quad x > 0, \end{cases} \tag{3.1b}$$

such as (2.22) with uniform concentration.

Both the equation and initial data of the Riemann problem can be expressed in terms of the single variable $\xi = x/t$, and this symmetry extends to solutions as well. Imposing this form on the solution reduces the problem to finding a heteroclinic orbit for the autonomous system

$$\left[J\big(U(\xi)\big) - \xi I\right] \dot{U}(\xi) = 0, \tag{3.2a}$$

$$U(-\infty) = U_L, \qquad U(+\infty) = U_R, \tag{3.2b}$$

where $J(U)$ is the Jacobian derivative of the flux function F. Smooth solutions of (3.1a), known as rarefactions, are therefore either constant or vary along integral curves R_i of a Jacobian eigenvector r_i. For this reason, most existence results apply to strictly hyperbolic systems, in which the eigenvalues are real and distinct.

Equation (3.2a) also requires that rarefaction solutions be parametrized by the corresponding eigenvalue λ_i, which is possible only if λ_i is strictly increasing on R_i between U_L and U_R. We discuss here the simplified case when F satisfies the genuine nonlinearity condition, which states that λ_i varies strictly monotonically along R_i for all i and R_i; we consider the more general case in the appendix.

In a genuinely nonlinear system, $R_i(U)$ consists of two connected curves $R_i^+(U) = \{U' \in R_i(U) \mid \lambda_i(U') > \lambda_i(U)\}$ and $R_i^-(U) = \{U' \in R_i(U) \mid \lambda_i(U') < \lambda_i(U)\}$, and a connecting orbit exists when $U_L = U$ and $U_R \in R_i^+(U)$, or $U_R = U$ and $U_L \in R_i^-(U)$. Consequently smooth solutions do not exist for general data, and solutions are generally sought from the larger class of weak solutions.

A weak solution to the conservation law (3.1a) is an L^∞ function $U(x,t)$ that in addition to the initial condition satisfies

$$\int_{x_1}^{x_2} \big(U(x,t_2) - U(x,t_1)\big)dx + \int_{t_1}^{t_2} \big(F(U(x_2,t)) - F(U(x_1,t))\big)dt = 0 \tag{3.3}$$

for all $x_2 > x_1$ and $t_2 > t_1 > 0$. This includes all smooth solutions to (3.1a), but also allows discontinuities along a curve $x = st$ that satisfies the vector Rankine–Hugoniot condition

$$F(U^+) - F(U^-) = s\big(U^+ - U^-\big), \tag{3.4}$$

where U^- and U^+ are the values of U on either side of the discontinuity. The Hugoniot locus $H(U^-)$ is defined as the set of U^+ that satisfy (3.4) for some s. (Note that while the symmetry of (3.4) implies $U_2 \in H(U_1)$ is equivalent to $U_1 \in H(U_2)$, it does not follow that $H(U_1) = H(U_2)$.)

Such weak solutions are not unique, however, and a method must be chosen to select a single solution. Various criteria, known as entropy conditions, have been proposed in order to distinguish the shock, or admissible discontinuity, from any other weak solutions. One condition, the method of viscous profiles, is motivated by the fact that conservation laws often appear physically as approximations to higher-order regularized equations such as

$$\frac{\partial}{\partial t}U^\epsilon + \frac{\partial}{\partial x}F(U^\epsilon) = \epsilon\frac{\partial^2}{\partial x^2}U^\epsilon, \tag{3.5}$$

which are well-posed for $\epsilon > 0$. A solution to (3.1a), according to this method, should be stable in the sense that it appears as the pointwise limit in x, t of solutions U_ϵ to (3.5) as $\epsilon \to 0$. This condition has the advantage of a clearly desirable physical interpretation that assures shock solutions are unique; however, it has the drawback of being difficult to verify.

A simpler method from the analytical perspective is the Lax entropy condition, which is equivalent to the viscous profile condition for a certain class of scalar conservation laws. This method relies on strict hyperbolicity to index the eigenvalues λ_i of $J(U)$ in increasing order for each U. These eigenvalues represent the characteristic speeds at which the equation propagates information, as can be seen in rarefaction solutions to the Riemann problem in the persistence of the left state U_L for $x \le \lambda_i(U_L)t$ and the right state U_R for $x \ge \lambda_i(U_R)t$. The Lax entropy condition requires the discontinuity to be continually reinforced by conflicting information from a single characteristic field, i.e., it moves with a speed s that satisfies

$$\lambda_i(U_L) > s > \lambda_i(U_R) \tag{3.6}$$

for exactly one i. That characteristic is emphasized by calling the discontinuity an i-shock.

In a neighborhood of any U the Hugoniot locus $H(U)$ consists of two smooth curves intersecting at U, and the four branches leaving U correspond to the four cases of 1- or 2-shocks with U as the right or left state. We denote the continuations of these branches by U_i^+ if U is the left state and U_i^- if U is the right state. The allowable connections $C_i^+(U_L) = R_i^+(U_L) \cup S_i^+(U_L)$ through the ith characteristic also locally form a smooth curve for each i. The variation of an i-shock or i-rarefaction solution is confined to the interval $\{\xi : \min(\lambda_i(U_L), \lambda_i(U_R)) < \xi < \max(\lambda_i(U_L), \lambda_i(U_R))\}$, so compound connections can be generated by stringing together waves of different characteristics as long as ξ increases with i. In fact, $\{C_i^+\}_{i=1}^n$ locally generate a smooth coordinate system, so if U_R is sufficiently close to U_L, the Riemann problem is well-posed.

Existence of solutions for large data depends on the topology of $H(U)$. A famous example of a system with no solutions for certain Riemann data is the Keyfitz–Kranzer equation (5.1) [29], in which $H(U)$ is compact. A bounded Hugoniot locus implies a bound on the strength of a shock, and consequently some large-data Riemann problems have no weak solutions. Section 5 describes a theory for such systems relating the regularized profiles to a Dirac mass; however, this theory is far from complete.

A final complication to the selection of weak solutions is the nature of the regularization actually present in the physical system. The Lax and Oleinik conditions are intended to admit those shocks that appear as viscous limits under the simplest possible regularization. If the actual regularization is different, the viscous profiles could converge to a weak solution other than that selected by the entropy criteria. This possibility is indeed relevant to conservation laws describing thin films, which are generally regularized by nonlinear fourth-order capillary terms such as in (2.20). In fact, a scalar thin film equation with similar regularization is known to select an entropy-violating double-shock solution, rather than the single-shock entropy solution [4].

4. Particular solutions. The system (2.23) is physically meaningful for $h > 0$ and $0 \leq \phi < \phi_m$, or equivalently $0 \leq v < \phi_m u$ in terms of the conserved quantities $u \equiv h$ and $v \equiv \phi h$. While the above theory depends on the latter parameterization, the equations are most simply expressed in terms of the physical variables h and ϕ, which we will use to present our results.

As shown in Figure 4.1, the equations using (2.14) are not strictly hyperbolic near the maximum concentration, where the eigenvalues become complex and the equations become elliptic. It is not clear whether this feature is desired in a model of the thin film. Change of type certainly complicates the mathematical question of well-posedness for such a system, but the parabolic system (2.20) is well-posed regardless of the first-order approximation. Also models proposed for dry granular materials result variously in hyperbolic, parabolic, and elliptic equations, so physically the change of type does not seem altogether unreasonable. Equations (2.23) are also not genuinely nonlinear on the entire domain; the significance of this is discussed in the appendix.

Since h has been rescaled to unity and θ appears only in the time scale, solutions to (2.23), (2.22) depend on ϕ_L, ϕ_R, b, and a. Although the relative values of ϕ_L and ϕ_R appear to be important, we consider only the cases $\phi_R = \phi_L$ and $\phi_R = 0$, which are most likely to occur in experiments. The value of ϕ_L itself appears to have only qualitative significance. The particle radius a has two effects: the time scale is proportional to a^2, and the film thickness at which the wall effect cutoff occurs (the inflection point of $w(h)$) is proportional to a. The appropriate range for a is fairly small, however; as for $a > 0.2$ discrete particle effects may be important, and for small a the relative velocity vanishes as a^2, so we use $a = 0.1$ for all calculations. The precursor thickness b is the most important parameter, but before discussing its effects we describe a typical solution.

We choose $(h_L, \phi_L) = (1.0, 0.3)$ as a representative left (upstream) state, and display in Figure 4.1 the four connection curves (1-shock, 1-rarefaction, 2-shock, 2-rarefaction) containing points that can be reached directly from this state. The rar-

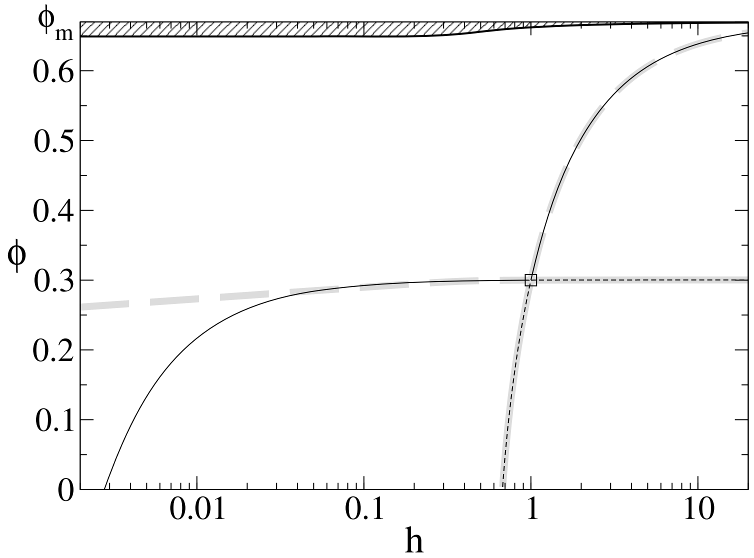


FIG. 4.1. *The phase space of the reduced model, and the connections from* $(h_0, \phi_0) = (1.0, 0.3)$, □. *The system is hyperbolic except in the shaded region. Black lines represent shock connections and gray represents rarefactions. Solid lines are connections to the right, i.e., the* (h_0, ϕ_0) *is the left state, and dashed lines are connections to the left. 1-waves and 2-waves can be distinguished by their slope at* (h_0, ϕ_0)*: 2-waves are nearly horizontal at this scale. Except at very small h, the shocks and rarefactions nearly coincide.*

efaction curves have been integrated from (3.2a) by a Runge–Kutta method, and $H(U_L)$ has been calculated by eliminating s from (3.4) at each point and solving the resulting equations for u and v. For a given shock connection, the shock speed can be recovered by substituting u and v back into (3.4).

For a specified right state (b, ϕ_R) representing the precursor, a solution can be determined by finding an intersection between the two connection diagrams, since the intersection represents an intermediate state that connects to both the left and right boundary conditions through shocks and/or rarefactions. In Figure 4.2 we have plotted the possible shock-shock connections for four values of b with $\phi_R = \phi_L$. At $b = 0.1$ there is a solution with a 1-shock from the upstream state to an intermediate height and concentration slightly larger, and a 2-shock from this intermediate state to the precursor. As the precursor becomes thinner, the height and concentration of this intermediate state increase. For $b = 0.01$ the intermediate state is approximately $(h, \phi) = (1.1757, 0.3663)$. In Figure 4.3 we compare this connection with a numerical solution with the same initial data and find that both shock speeds and the height and concentration of the ridge are in agreement. The numerical solution was calculated using the Lax–Friedrichs finite difference method with grid spacing 3.3×10^{-7} and time step 3.3×10^{-7}.

At $b = 0.008$ the Hugoniot locus has undergone a bifurcation such that the 1- and 2-shock curves are no longer distinct, and an additional connected component has appeared. Inspection of the shock speed and characteristic speeds along these curves reveals that various sections correspond to 1-shocks, 2-shocks, or are not admissible at all. There is still a shock-shock connection for $b = 0.008$ that satisfies the Lax entropy condition; however, at $b = 0.0015$ there are no longer any intersections, and therefore no solution. We discuss this last case in section 5, and in section 6 describe

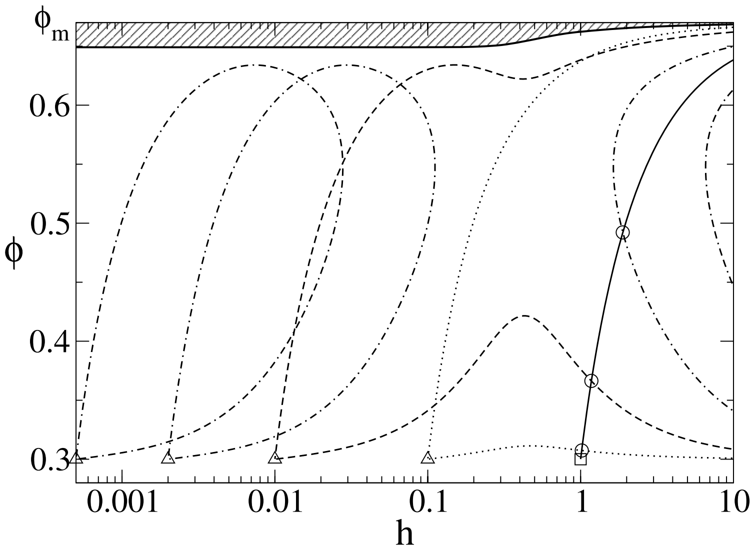


FIG. 4.2. *1-shock connections (solid line) from an upstream state* $(h_L, \phi_L) = (1.0, 0.3)$ (□) *and 1- and 2-shock connections from four precursor states* $(h_R, \phi_R) = (b, 0.3)$ (△)*, where* $b = 0.1$ *(dot),* 0.01 *(dash),* 0.002 *(dot-dash), and* 0.0005 *(dot-dash-dash). The solutions involve an intermediate state between the two shocks, marked by* ○*. As b becomes small, the Hugoniot locus undergoes a bifurcation and ultimately fails to produce a shock solution.*

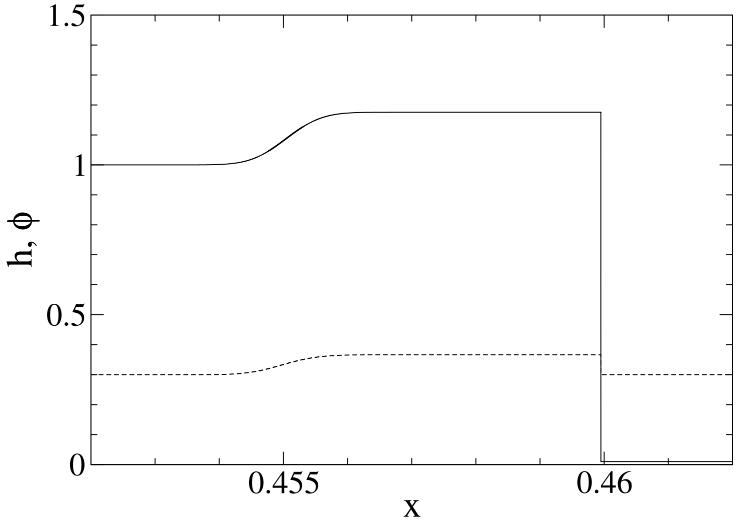


FIG. 4.3. *Film thickness (solid) and concentration (dashed) of a numerical solution of the conservation laws at* $t = 1$*, with* $(h_L, \phi_L) = (1.0, 0.3)$ *and* $(h_R, \phi_R) = (0.01, 0.3)$*. The intermediate state (between the shocks) is* $(h, \phi) = (1.1757, 0.3663)$*, as calculated in Figure* 4.2*. The speed of the (trailing)* 1*-shock is nearly equal to one of the characteristic speeds, making this shock especially susceptible to numerical diffusion.*

a change to the hindered settling function that ensures a solution does exist.

If the concentration in the precursor is taken to be 0 rather than ϕ_L, double-shock solutions still occur for moderately small b, and again no solution exists for smaller b. At larger b, another type of solution occurs consisting of a 1-rarefaction and a

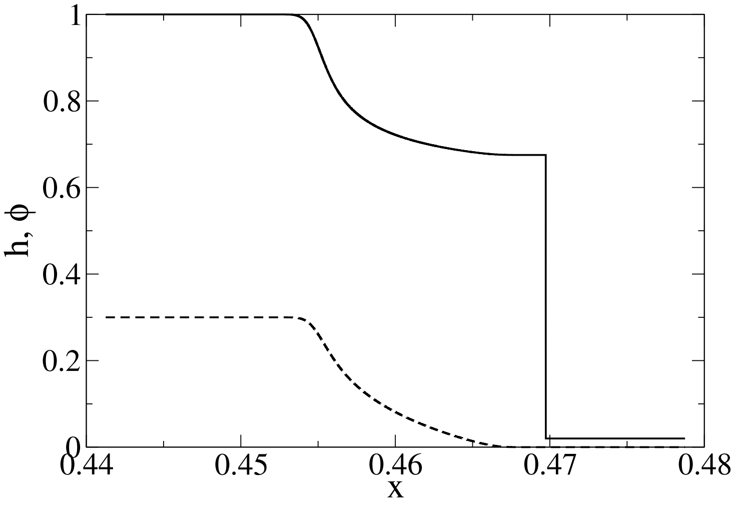


FIG. 4.4. *Numerical solution of the conservation laws at* $t = 1$*, with* $(h_L, \phi_L) = (1.0, 0.3)$ *and* $(h_R, \phi_R) = (0.02, 0)$*, corresponding to a* 1*-rarefaction and* 2*-shock. While some of the smoothness is due to numerical diffusivity, the* 1*-rarefaction can also be distinguished from a* 1*-shock by the fact that both* h *and* ϕ *are less than their values on the left.*

2-shock, with both h and ϕ in the intermediate state less than their values at the left. A numerical solution for this case is shown in Figure 4.4, again computed using the Lax–Friedrichs method in a moving frame.

As found by Zhou et al., the double-shock solutions agree qualitatively with the particle-rich ridge seen in experiments. The consistent trend is toward a thicker and more concentrated ridge as the b becomes smaller, until the solution ceases to exist. While it is difficult to know what value to use for b for a given experiment, this trend does indicate that a prewet surface or a more strongly wetting fluid-solid combination will result in a relatively smaller and less concentrated ridge. Very small values of b for which there is no solution are harder to interpret, since it is possible that solutions to the full system (2.20) display behavior that cannot be approximated by first-order equations. The problem of nonexistence is avoided, however, in the modified equations introduced below in section 6, which indicate a simple continuation of the trend toward thicker ridges. In contrast the rarefaction-shock solutions obtained with a fairly thick and particle-free precursor are unlike anything seen in experiments. These solutions are characterized by a thinner, particle-depleted region near the contact line which is created as the advancing film is diluted by the clear fluid in the precursor, and this effect is enhanced as the resulting drop in viscosity causes the depleted region to spread downstream. Perhaps this behavior may be observable with a sufficiently thick and particle-free prewet surface.

5. Singular shocks. The problem of nonexistence due to nontrivial Hugoniot topology has been studied before, and a weaker form of solution known as a singular shock has been described. An illustrative example is the Keyfitz–Kranzer equation [29]

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} u^2 - v \\ \frac{1}{3}u^3 - u \end{pmatrix} = 0, \tag{5.1}$$

which is everywhere both strictly hyperbolic and genuinely nonlinear, but for all $U = (u, v)$ the Hugoniot locus is compact, specifically figure-eight shaped. Thus shocks can only connect states that are sufficiently close, and certain Riemann problems have no classical solution.

In [27], Kranzer and Keyfitz present three sequences of functions $U^\epsilon(\xi = x/t)$ to (5.1) that approximately solve (5.1) as $\epsilon \to 0$ but are also singular in this limit. The first sequence results from an asymptotic expansion of the solution to the regularized equation

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x} F(U) = \epsilon t \frac{\partial^2 U}{\partial x^2} \tag{5.2}$$

in ϵ, and the second and third are explicitly constructed from C^∞ functions and piecewise constant functions. They introduce a space of measures in which these sequences converge to a limit involving Dirac-like masses superimposed on a classical shock. They also propose overcompression as an admissibility requirement for singular shocks, i.e., (3.6) must hold for both characteristics; if singular shocks are accepted under this restriction, (5.1) is well-posed for all Riemann data. However, these conclusions are restricted to (5.1). Also, Kranzer and Keyfitz emphasize that while the limiting measures appear as limits of approximate solutions, no well-defined criterion has been proposed by which the limits themselves can be called solutions.

Sever discusses the selection mechanism for singular shocks in a more general context in [48]. For a distribution solution

$$U(x,t) = M(t)\delta(x - st) + \begin{cases} U_L & \text{if } x < st, \\ U_R & \text{if } x > st \end{cases} \tag{5.3}$$

characterized by a point mass $M(t)$ located at $x = st$, conservation implies the singular mass must satisfy

$$\frac{dM}{dt} = s(U_R - U_L) - \big[F(U_R) - F(U_L)\big]. \tag{5.4}$$

Since the speed s is unknown, this is an undetermined system for the $n+1$ parameters dM/dt, s. For (5.1), Kranzer and Keyfitz determined unique solutions by requiring the first component of M to vanish, justified by an argument specific to that system. Sever writes that this last constraint generally comes from properties of the system such as symmetry groups or a convex entropy function. The proper constraint for system (2.23) is not yet apparent.

Equations (2.23) with regularization (3.5) also show behavior consistent with a singular shock. In order to investigate this, numerical solutions were generated with a fully implicit centered difference scheme on a moving nonuniform grid. The number of grid points at each mesh size was fixed; however, every 10 time steps the grids were rearranged using cubic interpolation as necessary to center the area of maximum resolution around the singularity. Meanwhile the entire computational domain moved at a constant speed chosen to approximately match the speed of the discontinuity. The scaling of the regularized solution satisfies $U^\epsilon(x,t) = U^1(\epsilon x, \epsilon t)$, so rather than take $\epsilon \to 0$ we fixed $\epsilon = 1$ and evaluated the solution at long times.

Figure 5.1 contains the results of this calculation. Both components of the singular mass increase linearly in time, as required by (5.4), and the singularity is overcompressive. As the singularity evolves in time the maximum height and concentration grow, and at $t \approx 3 \times 10^8$ the concentration exceeds the packing fraction. Clearly this solution does not describe the physical problem. While the nonlinear fourth-order diffusion in (2.20) may behave differently than the linear second-order diffusion studied here, possibly resulting in realistic solutions for the full model, the modification to

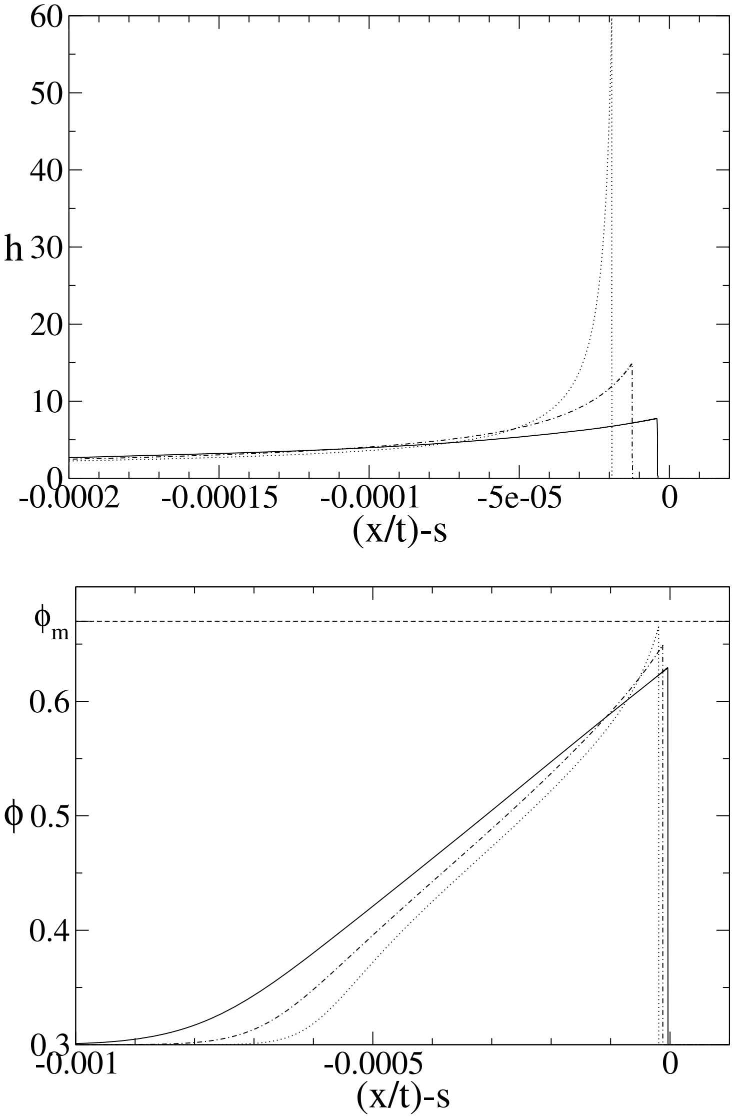


FIG. 5.1. *Film thickness (top) and particle concentration (bottom), from numerical solutions of the regularized system* (3.5) *in the singular shock regime, with* $b = 0.001$, $\phi_0 = 0.3$, *and* $\epsilon = 1$, *calculated on a grid moving at speed* $s = 0.45547$ *and evaluated at times* 5×10^7 *(solid),* 1×10^8 *(dot-dash), and* 2×10^8 *(dot).*

the model introduced in section 6 suggests the crucial issue is the high-concentration physics, rather than any divergence between the first- and fourth-order equations.

6. Alternative settling function. In this section we propose a modification to the unregularized system (2.23) that prevents the concentration from exceeding ϕ_m. We begin with a heuristic explanation of how (2.14) may be incompatible with (2.1) in the limit $\phi \to \phi_m$. The volume-averaged velocity is controlled by $\mu(\phi)^{-1}$, which vanishes in this limit, while $f_{RZ}(\phi)$, and hence the relative flux, is nonzero. This imposes a forward flux of particles with no net volume flux, requiring fluid

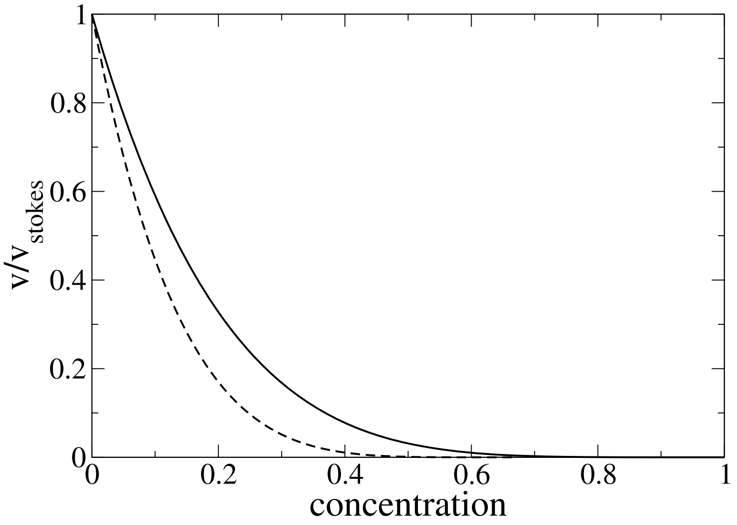


Fig. 6.1. *Two forms of the hindered settling function. The Richardson–Zaki function (solid line) given by* (2.14) *vanishes at concentration* 1.0*; the Buscall et al. function (dashed line) given by* (6.1) *vanishes at the packing fraction* $\phi_m = 0.67$.

therefore to move backward. This situation is probably unrealistic, because the limit $\mu(\phi) \to \infty$ is intended to model the case when the particles are packed tightly enough to prevent any shear flow. In that case, it seems more appropriate to model the particles as an immobile porous medium, with a Darcy's law flux of pure fluid and $\mathbf{v_{rel}} < 0$. Incorporating such a transition into the current model presents challenges, as the particle velocity must be specified relative to the laboratory frame rather than the fluid, essentially changing to a two-fluid model at high concentrations. A much simpler alternative is to simply let $\mathbf{v_{rel}}$ vanish along with $\mathbf{v}$ at $\phi = \phi_m$; this is readily accomplished by using the hindered settling function proposed by Buscall et al. [7],

$$f_B(\phi) = (1 - \phi/\phi_m)^5, \tag{6.1}$$

instead of (2.14). The two settling functions are plotted in Figure 6.1.

With this modification, solving the Riemann problem is simplified in two significant ways: the equations are strictly hyperbolic throughout the relevant domain Ω, and the bifurcation causing shock solutions to break down does not occur. In Figure 6.2 we have plotted shock-shock connections for four values of b. These solutions exist even for very small precursors, so the system appears to be well-posed regardless of b. Figure 6.3 summarizes the manner in which the type of solution depends on the settling function and the Riemann data.

In Figures 6.4–6.5, we compare the shock solutions to the two systems and their dependence on the precursor b. The behavior of the Hugoniot curves in the $f_{RZ}(\phi)$ system, shown in Figure 4.2, implies the intermediate height and concentration approach a maximum value at a critical precursor thickness $b = b_* \approx 9 \times 10^{-4}$, below which there is no meaningful solution. As $b \to 0$ in the $f_B(\phi)$ system, the intermediate height increases apparently without bound and the concentration approaches ϕ_m. We also observed in both limits that the speeds of the 1- and 2-shocks become approximately equal, indicating that the ridge, located between the shocks, is compressed horizontally while growing vertically.

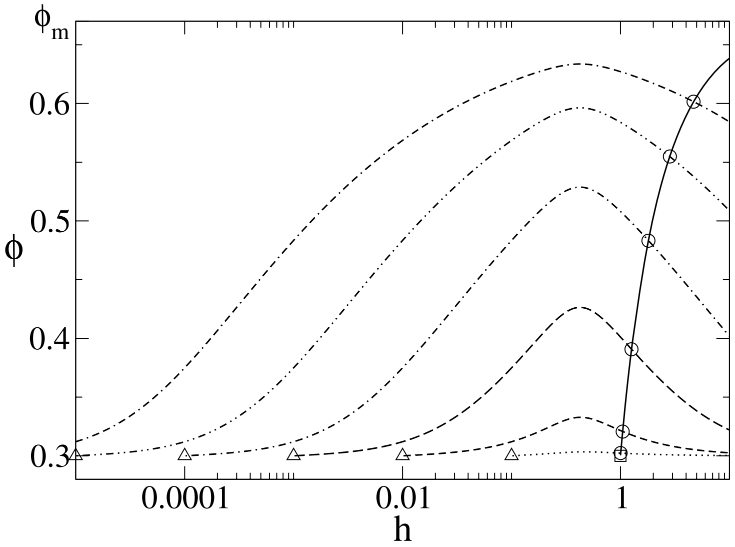


FIG. 6.2. *Shock connections using the settling function $f_B(\phi)$ instead of $f_{RZ}(\phi)$. The bifurcation that caused some initial data to have no solution no longer occurs. The solid line is the* 1*-shock connection from* (h_L, ϕ_L) (□), *and the* 2*-shocks are plotted from various precursors* (△) *given by* $b = 10^{-1}$ *(dot),* 10^{-2} *(short dash),* 10^{-3} *(long dash),* 10^{-4} *(dot-dash),* 10^{-5} *(dot-dot-dash), and* 10^{-6} *(dot-dash-dash). Each solution involves an intermediate state marked by* ○.

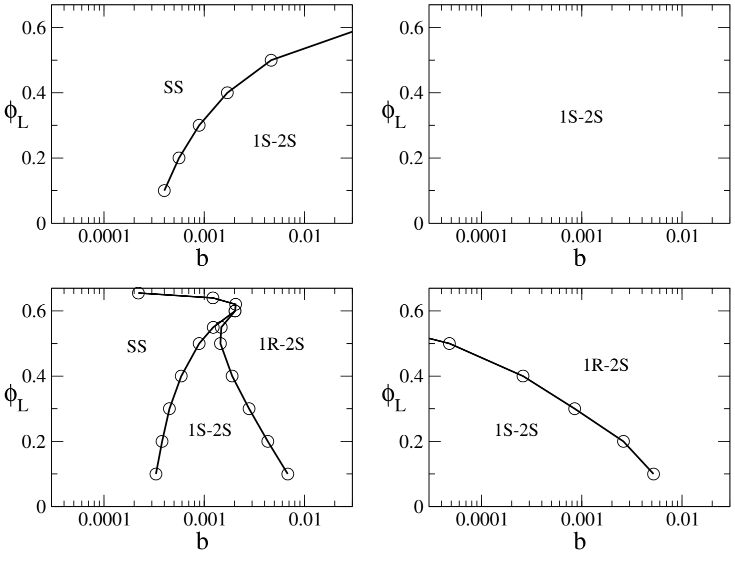


FIG. 6.3. *Type of solution (*1*-rarefaction and* 2*-shock,* 1*-shock and* 2*-shock, or singular shock) as determined by b and* ϕ_L *(assuming* $h_L = 1$ *and either* $\phi_R = \phi_L$ *or* $\phi_R = 0$*) for both hindered settling functions. Richardson–Zaki settling and* $\phi_R = \phi_L$ *(upper left), Richardson–Zaki settling and* $\phi_R = 0$ *(lower left), Buscall et al. settling and* $\phi_R = \phi_L$ *(upper right), Buscall et al. settling and* $\phi_R = 0$ *(lower right).*

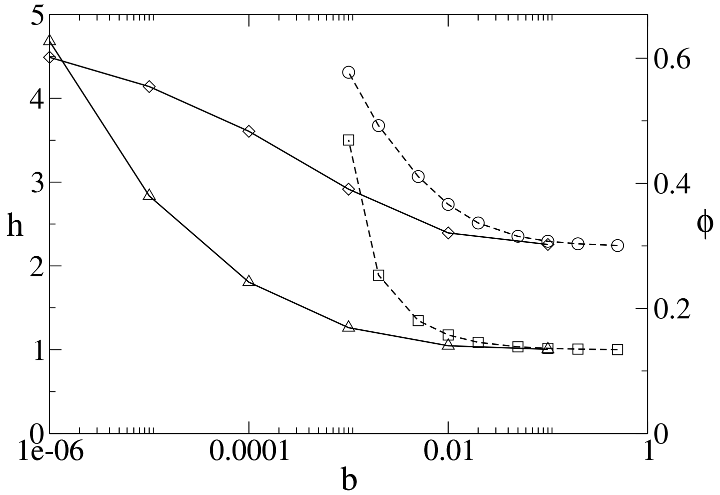


Fig. 6.4. *Height and concentration of the intermediate state vs. the precursor thickness b. Squares and circles are the height and concentration of solutions using the hindered settling function $f_{RZ}(\phi)$, triangles and diamonds are the height and concentration of solutions using $f_B(\phi)$.*

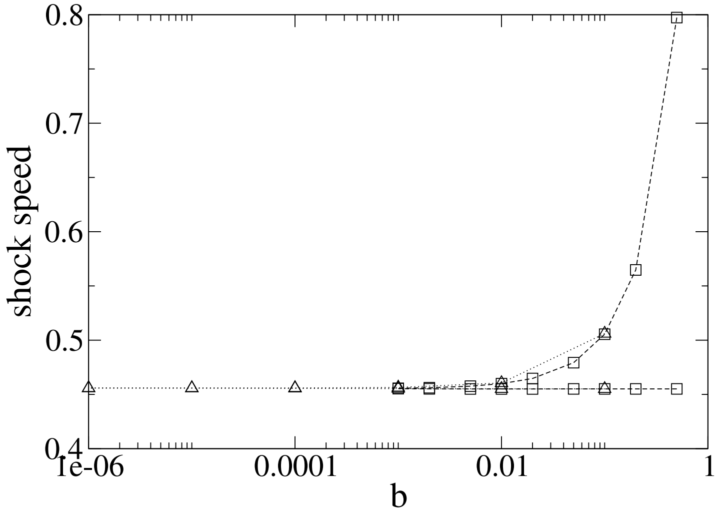


Fig. 6.5. *The speed of the shocks that make up the solutions to the connection problem for various precursors. Squares are solutions using the hindered settling function $f_{RZ}(\phi)$, and triangles with $f_B(\phi)$.*

7. Conclusion. In section 2, we derived a lubrication model for particle-laden films in the case where particle settling occurs only in the direction of flow. We did not analyze this fourth-order system, but rather the associated first-order reduced model; analogies with similar problems suggest this may be a reasonable approximation to the full system. While establishing correspondence between the reduced and full models is beyond the scope of this paper, the potential correspondence motivates our main result, a complete characterization of the first-order problem and a discussion of its possible connections to experiments.

The most important parameter in the reduced system is the precursor thickness b. When b is large enough, this system has a double-shock solution in qualitative agreement with the experimentally observed particle-rich ridge, while for smaller b, there is no classical solution. If the concentration in the precursor is the same as in the upstream source, these are the only two cases; setting the precursor concentration to zero allows a third possibility of a rarefaction-shock solution that has not been seen in experiments. We have confirmed the converging shock speeds that Zhou et al. reported in their preliminary discussion of the double-shock solutions, and we find that the speeds appear to become equal precisely (at the same value of b) when the classical shock solution breaks down.

At precursor thicknesses for which classical solutions do not exist, we have investigated a simple regularization of the equations for which the solution resembles a singular shock. These solutions are not at all realistic, partly because the growing delta mass at the shock location means the height is unbounded as $t \to \infty$, and partly because the close packing concentration ϕ_m is eventually exceeded.

A heuristic explanation was offered in section 6 for this exotic behavior: inspecting the limiting fluxes as $\phi \to \phi_m$ suggests the relative velocity should also vanish in this limit. This can be achieved by substituting the hindered settling function (6.1) of Buscall et al. for that of Richardson and Zaki, and the resulting Riemann problem appears to be well-posed for all precursor thicknesses. Thus physical arguments and the expectation of a well-posed first-order system both suggest that functions such as $f_B(\phi)$ that vanish at ϕ_m are most appropriate for this problem.

Many interesting questions remain unanswered regarding this model. More work is needed to determine how well the present results concerning the first-order system (2.23) approximate the full fourth-order system (2.20). Also of interest is the stability of the two-dimensional model (2.19) with respect to fingering patterns, as the experiments of Zhou et al. found the instability to be suppressed when a particle-rich ridge develops [59]. Other questions arise from the limitations of the current model. Explaining the three distinct settling behaviors observed by Zhou et al. requires a more general model considering particle settling in the normal direction, perhaps balanced by a shear-induced particle flux as in [47]. In addition to explaining the phase diagram, such a model could help determine whether the assumption in the current model—that particle concentration is constant across the film depth—is realistic. Changes to the model may also be needed to describe very high concentrations, as suggested in section 6, because contact forces between particles can be expected to become important.

Appendix. Genuine nonlinearity. While most physical systems are strictly hyperbolic, systems arising naturally are often not genuinely nonlinear. In the Euler equations of compressible flow, one characteristic field is linearly degenerate: $r_i \cdot \nabla\lambda_i \equiv 0$. For this characteristic, $R_i(U)$ and $S_i(U)$ coincide and connections take the form of contact discontinuities, which satisfy (3.4) with the inequalities in (3.6) replaced by equality. More generally, when the variation of λ_i along R_i changes sign, the strict inequality in (3.6) becomes too restrictive and a more general entropy condition is needed to select which contact discontinuities are admissible solutions.

For a scalar conservation law, genuine nonlinearity is simply the strict convexity (or concavity) of the flux function F. If the function changes concavity, contact discontinuities are chosen by the Oleinik condition [43], which states that the shock speed $s(U_L, U_R)$ satisfies

$$s(U_L, U_R) \le s(U_L, U) \tag{A.1}$$

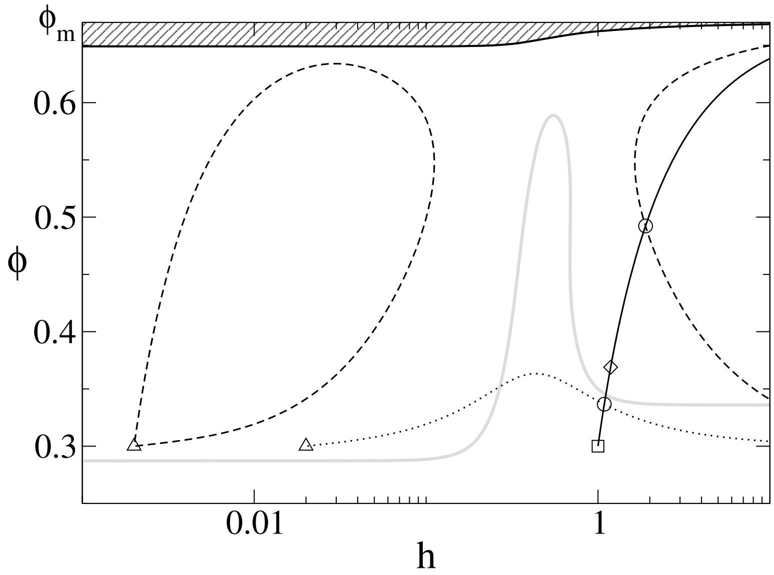


FIG. A.1. *Failure of genuine nonlinearity for* (2.23)*:* $\nabla\lambda_1 \cdot r_1 = 0$ *on the gray line. Connections from* $(h_L, \phi_L) = (1.0, 0.3)$ (□) *are plotted on the dashed line, which include shocks up to* $(h_*, \phi_*) \approx (1.18, 0.369)$ (◇) *or a compound shock to* (h_*, ϕ_*) *followed by a rarefaction.* 2*-shocks are plotted from right states* (△) *for one case* ($b = 0.02$*, dotted line) with a simple* 1*-shock,* 2*-shock solution, and another case* ($b = 0.002$*, dashed line) with a compound* 1*-shock,* 1*-rarefaction wave and a* 2*-shock. The equations are elliptic in the shaded region.*

for every U between U_L and U_R. Liu has generalized the Oleinik condition to 2×2 [36] and $n \times n$ systems [37] by requiring (A.1) to hold for all $U \in H(U_L)$ between U_L and U_R. Both Liu's and Oleinik's conditions reduce to (3.6) for a genuinely nonlinear system. While potentially only a bounded segment of $H(U_L)$ could be available for discontinuous waves, relaxing condition (3.6) provides more solutions by allowing both continuous and discontinuous waves in the same characteristic. Liu provides an existence proof by constructing such a compound wave. This connection involves a shock to the first point U_* satisfying $s(U_L, U_*) = \lambda_i(U_*)$, followed by a rarefaction from U_* to $U_R \in R_i^+(U_*)$. The point U_* is both the first local minimum of s along $H(U_L)$, hence the last point for which Liu's entropy condition is satisfied, and the first point for which $\lambda_i \geq s$, necessary for a continuing rarefaction wave.

In (2.23), $r_1 \cdot \nabla\lambda_1 = 0$ holds along the curve shown in Figure A.1. For $(h_L, \phi_L) = (1, 0.3)$ the branches S_1^+ and R_1^- nearly coincide, so this branch represents to good approximation the states accessible through a 1-shock, 1-rarefaction compound wave as well. In Figure A.2 the eigenvalue and shock speed are plotted on this curve as a function of ϕ. For $\phi < \phi_L$, both speeds increase away from U_L, indicating a simple rarefaction. With $\phi_L < \phi_* \approx 0.369$, the shock speed is strictly decreasing with ϕ so the connection is a shock satisfying the Liu–Oleinik condition. This case includes the solutions described in section 4 for $b = 0.1$ and $b = 0.01$. For $\phi > \phi_*$ neither simple wave is feasible, but a contact discontinuity from ϕ_L to ϕ_* can connect with a rarefaction from ϕ_* to ϕ because λ_1 is now both increasing and greater than the shock speed.

This compound wave is in practice difficult to distinguish from a simple shock. As noted above, the states accessible to a compound wave are nearly the same states lying

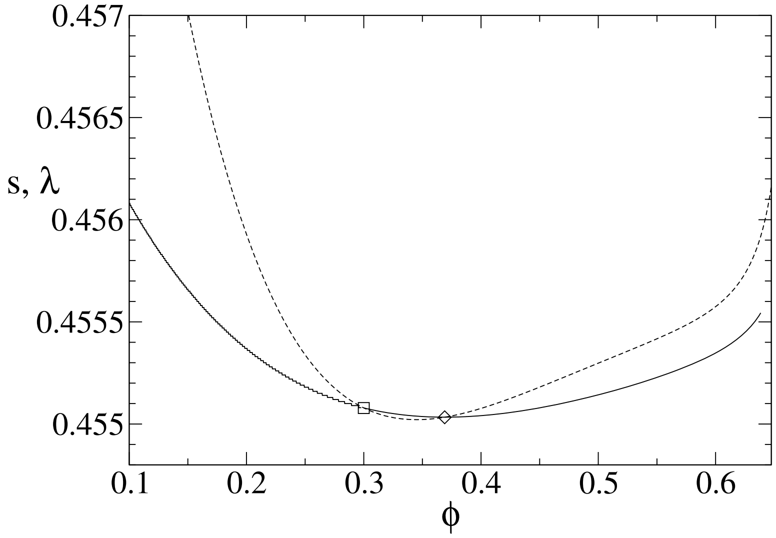


FIG. A.2. *Rarefaction speeds (dashed line) and shock speeds (solid) for the connections along the first characteristic from a left state* $(h_L, \phi_L) = (1.0, 0.3)$ (□) *(corresponding to Figure* 4.1*), plotted as a function of the concentration* ϕ_R *at the right state. The linear degeneracy curve in Figure* A.1 *indicates the location of the minimum characteristic speed. If* $\phi_R > \phi_* \approx 0.37$ (◇)*, a single shock solution is not admissible and the solution consists of a hybrid shock-rarefaction wave.*

on R_1 or S_1, so the constant state U_I appearing between 1-waves and 2-waves cannot easily be used to identify the compound wave. Additionally, Figure A.2 demonstrates that λ_1 changes very slowly along its characteristic at intermediate concentrations, so, for instance, in the presence of numerical diffusion, the rarefaction appears indistinguishable from a shock. Thus although some solutions are necessarily compound waves, their observable properties (other than failing to satisfy the Lax condition) are similar to those of a simple shock.

Acknowledgments. The authors would like to thank Professors Barbara Keyfitz and Michael Shearer for insightful comments and suggestions.

REFERENCES

[1] D. Ausserré, A. M. Picard, and L. Léger, *Existence and role of the precursor film in the spreading of polymer liquids*, Phys. Rev. Lett., 57 (1986), pp. 2671–2674.

[2] R. Benzi, L. Biferale, M. Sbragaglia, S. Succi, and F. Toschi, *Mesoscopic modeling of a two-phase flow in the presence of boundaries: The contact angle*, Phys. Rev. E, 74 (2006), article 021509.

[3] A. L. Bertozzi and M. P. Brenner, *Linear stability and transient growth in driven contact lines*, Phys. Fluids, 9 (1997), pp. 530–539.

[4] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.

[5] M. P. Brenner and P. J. Mucha, *That sinking feeling*, Nature, 409 (2001), pp. 568–570.

[6] J. P. Burelbach, S. G. Bankoff, and S. H. Davis, *Nonlinear stability of evaporating/condensing liquid films*, J. Fluid Mech., 195 (1988), pp. 463–494.

[7] R. Buscall, J. W. Goodwin, R. H. Ottewill, and T. F. Tadros, *The settling of particles through Newtonian and non-Newtonian media*, J. Colloid Interface Sci., 85 (1982), pp. 78–86.

[8] R. E. Caflisch and J. H. C. Luke, *Variance in the sedimentation speed of a suspension*, Phys. Fluids, 28 (1985), pp. 759–760.

[9] I. C. Carpen and J. F. Brady, *Gravitational instability in suspension flow*, J. Fluid Mech., 472 (2002), pp. 201–210.

[10] A. M. Cazabat, *Wetting films*, Adv. Colloid Interface Sci., 34 (1991), pp. 72–88.

[11] R. H. Davis and A. Acrivos, *Sedimentation of noncolloidal particles at low Reynolds numbers*, Annu. Rev. Fluid Mech. 17 (1985), pp. 91–118.

[12] J. R. de Bruyn, *Growth of fingers at a driven three-phase contact line*, Phys. Rev. A, 46 (1992), pp. R4500–R4503.

[13] P. G. de Gennes, *Wetting: Statics and dynamics*, Rev. Modern Phys., 57 (1985), pp. 827–863.

[14] E. B. Dussan V. and S. H. Davis, *On the motion of fluid-fluid interface along a solid surface*, J. Fluid Mech., 65 (1974), pp. 71–95.

[15] Z. W. Fang, A. A. Mammoli, J. F. Brady, M. S. Ingber, L. A. Mondy, and A. L. Graham, *Flow-aligned tensor models for suspension flows*, Int. J. Multiphase Flow, 28 (2002), pp. 137–166.

[16] R. Goodwin and G. M. Homsy, *Viscous flow down a slope in the vicinity of a contact line*, Phys. Fluids A, 3 (1991), pp. 515–528.

[17] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.

[18] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics, with Special Applications to Particulate Media*, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[19] L. M. Hocking, *Spreading and instability of a viscous fluid sheet*, J. Fluid Mech., 211 (1990), pp. 373–392.

[20] C. Huh and L. E. Scriven, *Hydrodynamic model of steady movement of a solid/liquid/fluid contact line*, J. Colloid Interface Sci., 35 (1971), pp. 85–101.

[21] H. E. Huppert, *Flow and instability of a viscous current down a slope*, Nature, 300 (1982), pp. 427–429.

[22] M. P. Ida and M. J. Miksis, *The dynamics of thin films* I: *General theory*, SIAM J. Appl. Math., 58 (1998), pp. 456–473.

[23] A. Indekina, I. Veretennikov, and H.-C. Chang, *Front dynamics and fingering of a driven contact line*, J. Fluid Mech., 373 (1998), pp. 81–110.

[24] J. M. Jerrett and J. R. de Bruyn, *Fingering instability of a gravitationally driven contact line*, Phys. Fluids A, 4 (1992), pp. 234–242.

[25] M. F. G. Johnson, R. A. Schluter, M. J. Miksis, and S. G. Bankoff, *Experimental study of rivulet formation on an inclined plate by fluorescent imaging*, J. Fluid Mech., 394 (1999), pp. 339–354.

[26] P. Jop, Y. Forterre, and O. Pouliquen, *A constitutive law for dense granular flows*, Nature, 441 (2006), pp. 727–730.

[27] B. L. Keyfitz and H. C. Kranzer, *Spaces of weighted measures for conservation laws with singular shock solutions*, J. Differential Equations, 118 (1995), pp. 420–451.

[28] L. Kondic and J. Diez, *Pattern formation in the flow of thin films down an incline: Constant flux configuration*, Phys. Fluids, 13 (2001), pp. 3168–3184.

[29] H. C. Kranzer and B. L. Keyfitz, *A strictly hyperbolic system of conservation laws admitting singular shocks*, in Nonlinear Evolution Equations That Change Type, IMA Vol. Math. Appl. 27, Springer, New York, 1990, pp. 107–125.

[30] I. M. Krieger, *Rheology of monodisperse latices*, Adv. Colloid Interface Sci., 3 (1972), pp. 111–136.

[31] G. J. Kynch, *A theory of sedimentation*, Trans. Faraday Soc., 48 (1952), pp. 166–176.

[32] P. D. Lax, *Hyperbolic Systems of Conservation Laws and Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.

[33] L. Léger, H. Hervet, G. Massey, and E. Durliat, *Wall slip in polymer melts*, J. Phys. Condens. Matter, 9 (1997), pp. 7719–7740.

[34] D. Leighton and A. Acrivos, *The shear-induced migration of particles in concentrated suspensions*, J. Fluid Mech., 181 (1987), pp. 415–439.

[35] R. Levy and M. Shearer, *The motion of a thin liquid film driven by surfactant and gravity*, SIAM J. Appl. Math., 66 (2006), pp. 1588–1609.

[36] T.-P. Liu, *The Riemann problem for general* 2×2 *conservation laws*, Trans. Amer. Math. Soc., 199 (1974), pp. 89–112.

[37] T.-P. Liu, *Riemann problem for general systems of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.

[38] V. Ludviksson and E. N. Lightfoot, *The dynamics of thin liquid films in the presence of surface-tension gradients*, Am. Inst. Chem. Engrs. J., 17 (1971), pp. 1166–1173.

[39] J. A. Moriarty and L. W. Schwartz, *Dynamic considerations in the closing and opening of holes in thin liquid films*, J. Colloid Interface Sci., 161 (1993), pp. 335–342.

[40] J. F. Morris and J. F. Brady, *Pressure-driven flow of a suspension: Buoyancy effects*, Int.

J. Multiphase Flow, 24 (1998), pp. 105–130.

[41] A. Münch and B. Wagner, *Numerical and asymptotic results on the linear stability of a thin film spreading down a slope of small inclination*, European J. Appl. Math., 10 (1999), pp. 297–318.

[42] P. R. Nott and J. F. Brady, *Pressure-driven flow of suspensions: Simulation and theory*, J. Fluid Mech., 275 (1994), pp. 157–199.

[43] O. A. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl., 26 (1963), pp. 95–172.

[44] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[45] R. J. Phillips, R. C. Armstrong, R. A. Brown, A. L. Graham, and J. R. Abbott, *A constitutive equation for concentrated suspensions that accounts for shear-induced particle migration*, Phys. Fluids A, 4 (1992), pp. 30–40.

[46] J. F. Richardson and W. N. Zaki, *Sedimentation and fluidization: Part* I, Trans. Inst. Chem. Eng., 32 (1954), pp. 35–53.

[47] U. Schaflinger, A. Acrivos, and K. Zhang, *Viscous resuspension of a sediment within a laminar and stratified flow*, Int. J. Multiphase Flow, 16 (1990), pp. 567–578.

[48] M. Sever, *Distribution solutions of nonlinear systems of conservation laws*, Mem. Amer. Math. Soc., 889 (2007).

[49] A. Sierou and J. F. Brady, *Rheology and microstructure in concentrated noncolloidal suspensions*, J. Rheology, 46 (2002), pp. 1031–1056.

[50] N. Silvi and E. B. Dussan V., *On the rewetting of an inclined solid surface by a liquid*, Phys. Fluids, 28 (1985), pp. 5–7.

[51] M. A. Spaid and G. M. Homsy, *Stability of Newtonian and viscoelastic dynamic contact lines*, Phys. Fluids, 8 (1995), pp. 460–478.

[52] J. J. Stickel and R. L. Powell, *Fluid mechanics and rheology of dense suspensions*, Annu. Rev. Fluid Mech., 37 (2005), pp. 129–149.

[53] H. Tang, W. Grivas, D. Homentcovschi, J. Geer, and T. Singler, *Stability considerations associated with the meniscoid particle band at advancing interfaces in Hele-Shaw suspension flows*, Phys. Rev. Lett., 85 (2000), pp. 2112–2115.

[54] B. D. Timberlake and J. F. Morris, *Particle migration and free-surface topography in inclined plane flow of a suspension*, J. Fluid Mech., 538 (2005), pp. 309–341.

[55] S. M. Troian, E. Herbolzheimer, S. A. Safran, and J. F. Joanny, *Fingering instabilities of driven spreading films*, Europhys. Lett., 10 (1989), pp. 25–30.

[56] M. Ungarish, *Physico-mathematical formulation*, in Hydrodynamics of Suspensions, Springer-Verlag, New York, 1993, pp. 7–36.

[57] T. P. Witelski and A. J. Bernoff, *Dynamics of three-dimensional thin film rupture*, Phys. D, 147 (2000), pp. 155–176.

[58] L. Xianfan and C. Pozrikidis, *Film flow of a suspension down an inclined plane*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 361 (2003), pp. 847–869

[59] J. Zhou, B. Dupuy, A. L. Bertozzi, and A. E. Hosoi, *Theory for shock dynamics in particle-laden thin films*, Phys. Rev. Lett., 94 (2005), article 117803.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 784–805

IDENTIFICATION AND CHARACTERIZATION OF A MOBILE SOURCE IN A GENERAL PARABOLIC DIFFERENTIAL EQUATION WITH CONSTANT COEFFICIENTS*

STEVEN KUSIAK† AND JOHN WEATHERWAX‡

Abstract. We discuss an inverse source problem for a general parabolic differential equation in $\mathbb{R}^n \times \mathbb{R}_+$ with constant coefficients and a source whose strength and support may vary with time. We demonstrate that a knowledge of the solution on any bounded open set $\mathcal{M}$ in $\mathbb{R}^n$ located away from the source for any fixed time $T \geq 0$ determines the so-called carrier support (originally defined in the article "Notions of support for far fields" [J. Sylvester, *Inverse Problems*, 22 (2006), pp. 1273–1288] as a nontrivial subset of the support of the true source) at that coincident time. Additionally, we provide a reconstruction algorithm which can locate the time-varying position of the carrier support of the assumed unknown source with extremely few discrete (possibly nonuniform) measurements taken on such an open set over a wide range of regularity classes of the source. Finally, we provide a few numerical examples which illustrate the efficacy and robustness of this location and tracking method.

Key words. inverse problems, diffusion, partial differential equations

AMS subject classifications. 49N30, 58J35, 35K10

DOI. 10.1137/070696970

1. Introduction.

Remote sensing endeavors, especially those in connection with modern defense and industrial quality control applications, have evolved significantly in recent years in both their technologies and the realistic scenarios they address. In particular, the release of toxic or impure substances into an environment of interest, urban or otherwise, by either intentional or unintentional means has become an outstanding contemporary problem of considerable and immediate importance and consequence. Recently, many promising technologies, such as those presented in [12, 7, 4, 6, 14, 13], have been created which can remotely detect the presence of foreign materials in a region of interest and estimate their concentration as a function of position and time, provided one has ample knowledge of the diffusion field on either large spatial measurement sets or over long periods of time, or both. Clearly, however, we can never hope to simultaneously monitor vast regions of space in many real-world settings, nor can we tolerate the need for long periods of measurement time, or indefinite ones for that matter. Hence, there is an immediate need to possess the capability to quickly detect and determine the location(s) and output strength(s) of life-threatening, or otherwise destructive, sources with extremely limited measurements in space and time of such diffusing substances.

An immediate extension of the work developed in [10] and more recently in [16] is made in this article that efficiently treats the problem of determining the location

*Received by the editors July 11, 2007; accepted for publication (in revised form) September 26, 2007; published electronically January 4, 2008. This research was partially supported by the Centre Nationale de la Recherche Scientifique (CNRS), France.

http://www.siam.org/journals/siap/68-3/69697.html

†New Frontier Advisors, LLC, Boston, MA 02110 (skusiak@newfrontieradvisors.com). Part of this research was done while this author was a research fellow at the Laboratoire POEMS, UMR 2706, CNRS/ENSTA/INRIA, Ecole Nationale Supérière de Techniques Avancées (ENSTA), 32 Boulevard Victor, 75739 Paris cedex 15, France.

‡Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02420 (wax@alum.mit.edu).

of a (potentially mobile) source in a generalized advection-diffusion environment with extremely limited real-world resources. For instance, the technology lends itself extremely well to the problem of the airborne release of a life-threatening substance aboveground or in an underground subway system. Specifically, in n-dimensional space, if the concentration of the diffusing substance can be measured on a small, coarsely sampled or highly distributed, n-dimensional array of sensors, at a single snapshot in time, then we can robustly, and expeditiously, determine the location of the source with considerable accuracy in the presence of considerable measurement noise. Hence, with a collection of such time snapshots of measurement data, we further show that we are able to locate the time-varying position of a moving source and track its current location. Such a capability is of tremendous value to time-sustained source release problems across a variety of scenarios.

We begin the analytical treatment of this problem by considering the general nonhomogeneous second order parabolic partial differential equation

$$(\partial_t - L_x)u(x,t) = f(x,t), \quad u(x,0) = 0, \quad (x,t) \in \mathbb{R}^n \times \mathbb{R}_+, \quad n \geq 1, \tag{1.1}$$

where we define the elliptic operator L_x as

$$L_x := \sum_{i,j=1}^{n} a_{i,j}\partial_{x_i}\partial_{x_j} + \sum_{j=1}^{n} b_j\partial_{x_j} + c,$$

which governs such things as the molecular diffusion of gases, or particulates, generated by the autonomous source f throughout $x \in \mathbb{R}^n$ and over time $t \in \mathbb{R}_+$. For the purposes of clarity we will limit ourselves to the treatment of the case where L_x has constant coefficients. We note, however, that much of the following analysis and framework suits the more general case of coefficients which at least vary with position. This is in fact the aim of future work.

In the treatment to follow we will assume that the source may be decomposed into the product of a temporally dependent function $s \geq 0$ with a potentially spatially dependent and mobile one $g \geq 0$, such that

$$f(x,t) = g(x - \gamma(t))s(t) \geq 0 \quad \forall\ (x,t) \in \mathbb{R}^n \times \mathbb{R}_+,$$

where, a priori, g is assumed to be compactly supported for each time t within the domain $B_R(p)$, i.e., the ball of radius R and center p, and $\gamma : [0,T] \to \mathbb{R}^n$, $T \geq 0$. Moreover, we assume that the structure of the strength function s is such that s identically vanishes for values of $t < 0$ and takes the form of regular or possibly singular distribution for values of $t \geq 0$, e.g., the Dirac-delta distribution or the Heaviside function.

Characterization of the source amounts to determining the (possibly time-varying) carrier support—which we will define in detail shortly, and for more detail refer the reader to [16]—of the source f, which we will assume to be strictly positive for this article. The concept of the carrier support generalizes that of the so-called scattering support, which was originally defined and analyzed in much detail in [10]. In short, for any differential operator, such as $P = \partial_t - L_x$, which admits the concept of the unique continuation principle (UCP),[1] the carrier support of the measured field u on $\mathcal{M}$ at time T in the differential equation $Pu = f$, where f is compactly supported,

[1] Suppose $Pu = 0$ in some domain $\mathcal{V}$. Then if u restricted to an open subset $\mathcal{M} \subset \mathcal{V}$ vanishes, the UCP implies that u vanishes throughout the larger domain $\mathcal{V}$.

is that subset of the support of f such that there exists an equivalent source $\tilde{f} = \tilde{g}\tilde{s}$ (with $\tilde{f}$ residing in the same regularity class as f, and where $\tilde{f} \neq f$ in the sense of distributions) in the sense that $Pu = \tilde{f}$ everywhere on the complement of the support of f. We summarize this concept with the following definition.

Definition 1.1 (carrier support). *Let P be a differential operator which admits the UCP such that $Pu = f$ on $\mathcal{V}$ and* supp $f \subset \Omega \subsetneq \mathcal{V}$*, and let P possess the fundamental solution E_P. Then, for some open set $\mathcal{M} \subset \mathcal{V}$, where $\overline{\mathcal{M}} \cap \overline{\Omega} = \emptyset$, and $\tau \leq T$,*

$$\text{carr supp } u(\cdot, T)|_{\mathcal{M}} := \bigcap_{E_P * \tilde{f} = u(\cdot, T)|_{\mathcal{M}}} \text{ch supp } \tilde{f}(\cdot, \tau), \quad 0 \leq \tau \leq T,$$

where ch *denotes the convex hull.*

In summary, this definition states that the carrier support of the solution u restricted to the set $\mathcal{M}$ is that common set over all possible sets such that there exists a compactly supported source away from $\mathcal{M}$ such that u on $\mathcal{M}$ may be generated by such a candidate source $\tilde{f}$, i.e., $P\tilde{u} = \tilde{f}$ on $\mathcal{V}$ and $\tilde{u}$ agrees with u on $\mathcal{V} \backslash \text{ch supp } f$.

We should also remark that this definition implies that there exists the possibility that the source generating the data u on $\mathcal{M}$ could have existed previous to time T, i.e., $T \geq \tau$. This means that the solution as observed on $\mathcal{M}$ is that of a now-extinct source that last existed at time τ whose support was last on supp $\tilde{g}(\cdot, \tau)$. This may be summarized through the following (identity) example. Let $H_\tau(t)$ denote the Heaviside function,[2] and suppose for $0 \leq \tau_1 < \tau_2$ that

$$f(x, t) = \delta_{\gamma(t)}(x)\left(H_{\tau_1}(t) - H_{\tau_2}(t)\right)$$

moves along the trajectory $\gamma : [0, T] \to \mathbb{R}^n$. Then

$$\text{carr supp } u(\cdot, t)|_{\mathcal{M}} = \begin{cases} \emptyset, & t < \tau_1, \\ \text{supp } \delta_{\gamma(t)}(x), & \tau_1 \leq t \leq \tau_2, \\ \text{supp } \delta_{\gamma(\tau_2)}(x), & t \geq \tau_2. \end{cases}$$

It should be mentioned that this is a very desirable phenomenon, in that detection of the location of an impulse-like source released at time $\tau \geq 0$ can be made with the observed data u on $\mathcal{M}$ at time $t \geq \tau$. This amounts to the ability to track mobile time-sustained sources and determine the point of detonation of impulse-like ones, both of which are invaluable capabilities for a variety of modern and future endeavors across many disciplines and industries.

We note that even though—as we shall come to prove in the following section—we can determine the current location, or that final one when the source's strength was last nonzero, with a snapshot of the current diffusion field on $\mathcal{M}$, we cannot reconstruct the entirety of γ over all times subsequent to t. In particular this means we may analytically extend the solution u of the homogeneous equation $(\partial_t - L_x)u = 0$ back to time t^* at which minimal time u was the solution of a nonhomogeneous equation of the form $(\partial_t - L_x)u = \tilde{f} \neq 0$. Again, we shall revisit this concept in further detail and provide all the necessary technical arguments which support this notion in the following section and in the proof of our main result (Theorem 3.2) in section 3.

[2] For a test function $\phi \in C^\infty(\mathbb{R}^n \times \mathbb{R})$, for $\delta_{\gamma(t)} H_\tau \in \mathcal{E}'(\mathbb{R}^n) \otimes \mathcal{D}'(\mathbb{R})$, we define the action of the distributional pairing $\langle \delta_{\gamma(t)} H_\tau, \phi \rangle = \int_\tau^\infty \int_{\mathbb{R}_x} \phi(x, t)\delta(x - \gamma(t))dxdt$.

Several results have been presented to date which can determine the actual support of the source and its strength function s as a function of time; see, for example, [12, 7, 4, 6]. The fundamental difference between these methods and that to follow in this article is that we require knowledge of the scalar diffusion field only at single snapshots in time on some (possibly small) open subset of the ambient space $\mathbb{R}^n$ which we will assume to be disjointly located from the source in question. Since we require far less information than that required in these previous works, we might expect to fail to fully characterize the source in all its attributes. We shall come to see that this is indeed the case, and that what we can estimate with this limited information is that of the carrier support of the instantaneous support of the true source f at each point $T \geq 0$ in time. We accomplish this by following a strategy similar in nature to the one presented in the articles [10, 11]. Essentially, we employ a unique continuation-like strategy for the assumed positive solution which, together with the Picard theorem, gives us a way to uniquely determine the carrier support at any time T. We will describe these ideas in much further detail in a subsequent section.

We again stress the importance of having measurements on some limited finite domain located away from the source, and which need not surround the source, which serves such applications as source-release problems in complex urban environments, and atmospheric or reservoir problems. More important, if the source is moving, we wish to determine its trajectory, and current location, for purposes of perhaps its rapid neutralization, i.e., the source is emitting toxic or impure substances into a domain of interest.

We will make several assumptions on the nature of the coefficients appearing in L_x and of that of the bounded open set $\mathcal{M} \subset \mathbb{R}^n$, where will assume to know the scalar values of the time-varying field $u(x,t)$; for instance, we require $\mathcal{M}$ to have a smooth boundary for the purposes of maintaining well-behaved norms of the solution on such sets of interest. Additionally, we will define a few function spaces of interest in which our solutions of the main problem will uniquely exist.

Remark 1. *In the analysis to follow in the upcoming section, we will assume that the following conditions hold:*

- *$\partial_t - L_x$ is parabolic on $\mathbb{R}^n \times \mathbb{R}$, i.e.,*

$$\sum_{i,j=1}^{n} a_{i,j}\xi_i\xi_j > 0, \quad \mathbb{R} \ni \xi_i, \xi_j \neq 0.$$

- *$c \leq 0$.*
- *The matrix of coefficients $a_{i,j}$ is positive definite and invertible.*

Remark 2. *Additionally, for $-n/2 < \sigma_1 \leq 0$ and $-1/2 < \sigma_2 \leq 0$, we define the function spaces*[3]

$$\overset{o}{H}{}^{\sigma_1}_{+}(\overline{\Omega}) = \{g \in H^{\sigma_1}(\mathbb{R}^n) \ : \ g \geq 0, \quad \operatorname{supp} g \subset \overline{\Omega}\},$$

where H^σ is the usual Sobolev space of regularity $\sigma \in \mathbb{R}$, and similarly

$$H^{\sigma_1}_{+}(\mathbb{R}) = \{g \in H^{\sigma_1}(\mathbb{R}) \ : \ g \geq 0\}.$$

Using these conventions we define the positive space of sources, i.e., distributions,

$$f \in \overset{o}{H}{}^{\sigma_1}_{+}(\overline{\Omega}) \bigotimes \overset{o}{H}{}^{\sigma_2}_{+}([0,T])$$

[3] We wish to include singular temporal and spatial distributions such as the Dirac-delta distribution in the larger collection of positive sources; hence we are interested in taking $0 \geq \sigma_1 > -n/2$ and $0 \geq \sigma_2 > -1/2$.

and the space of restricted solutions

$$u|_{\mathcal{M}} \in L^2_+\left([0,T], L^2_+(\mathcal{M})\right),$$

where

$$L^2_+(\mathbb{R}^n) = \{g \in L^2(\mathbb{R}^n) \ : \ g \geq 0\}.$$

In order to minimize symbolic clutter and ease notation a bit, we make the identification

$$X_f^{\sigma_1,\sigma_2}(\Omega,T) = \overset{o}{H}{}^{\sigma_1}_+(\overline{\Omega}) \bigotimes \overset{o}{H}{}^{\sigma_2}_+([0,T])$$

and note that its dual space admits the representation

$$\left(X_f^{\sigma_1,\sigma_2}\right)' = H^{-\sigma_1}_+(\mathbb{R}^n) \bigotimes H^{-\sigma_2}_+(\mathbb{R}_+)$$

for any finite T and bounded $\mathcal{M}$. Finally, we note that since $0 \leq -\sigma_1 < n/2$ and $0 \leq -\sigma_2 < 1/2$, by (complex) interpolation, i.e., see, for instance, [17, *p.* 277],

$$[L^2(\mathbb{R}^n), H^k(\mathbb{R}^n)]_\theta = H^{k\theta}(\mathbb{R}^n), \quad k = 0, 1, 2, \ldots, \quad 0 \leq \theta \leq 1,$$

we have the inclusion

$$H^n_+(\mathbb{R}^n) \bigotimes H^1_+(\mathbb{R}_+) \subset \left(X_f^{\sigma_1,\sigma_2}\right)' \subset L^2_+(\mathbb{R}^n) \bigotimes L^2_+(\mathbb{R}_+).$$

This last fact will be important for us in characterizing the behavior of the solution in the following section on the forward problem.

The plan of the remainder of the paper is as follows. In section 2 we discuss the forward problem. Namely, we describe how, given a well-defined source, f generates the solution u and develop the appropriate mapping characterizations between the two. In section 3, we focus on the inverse source problem, which again is to locate the support of f from measurements of the diffusion field u taken on some open region which is disjoint and distant from the assumed unknown source f. In this section we prove uniqueness results for the time-varying reconstruction of the carrier support of f as well as develop a viable reconstruction method which estimates it with little, sparse, and possibly nonuniformly sampled data. Section 4 considers a few numerical examples which robustly illustrate the simplicity and effectiveness of this reconstruction algorithm for a spatially stationary, impulse-like point-source release in one dimension and a time-sustained, moving point source in a convective two-dimensional environment.

2. The forward problem. We begin this section by noting that the solution of (1.1) is well known and may be constructed with the aid of a fundamental solution which we will call Z; see [5, 3] for the original details of this parametrix-based method. Specifically, let $A = \det a_{i,j}$ and $a^{i,j}$ be the determinant and matrix inverse of the positive definite matrix a, respectively. Then

$$u(x,t) = \int_0^t \int_{\mathbb{R}^n_y} Z(x,y,t,\tau) f(y,\tau) dy d\tau, \tag{2.1}$$

where

$$Z(x,y,t,\tau) = W(x,y,t,\tau) + \int_{\tau}^{t} \int_{\mathbb{R}_z^n} W(x,z,t,s)\Phi(z,y,s,\tau)dzds$$

and where we define

$$W(x,y,t,\tau) = [(4\pi(t-\tau))^n A]^{-1/2} \exp\left(- \sum_{i,j=1}^{n} \frac{a^{i,j}(x_i - y_i)(x_j - y_j)}{4(t-\tau)} \right)$$

and lastly require that Φ satisfy

$$\Phi(x,y,t,\tau) = (L_x - \partial_t)W(x,y,t,\tau) + \int_{\tau}^{t} \int_{\mathbb{R}_z^n} (L_x - \partial_t)W(x,z,t,s)\Phi(z,y,s,\tau)dzds.$$

For convenience we will denote the action of Z acting on f as simply $\mathcal{Z}f$. Furthermore, we will denote the restriction of $\mathcal{Z}$ to observations on $\mathcal{M}$ and limited to sources f having compact support on Ω as $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ in what is to follow. Additionally, we note that we will write $Z(x-y,t-\tau)$ in the place of $Z(x,y,t,\tau)$ when it appears in the context of the kernel of the convolution integral which maps the source f to the solution u.

We recall some important properties established in [5]. We use them to prove the following boundedness and denseness result.

PROPOSITION 2.1 (local boundedness of the solution on $\mathcal{M}$). *Given the assumptions detailed earlier in Remarks* 1 *and* 2, *then for each* $T \geq 0$, $\mathcal{Z} : X_f^{\sigma_1,\sigma_2}(\Omega,T) \to L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$ *and has dense range in the latter space.*

Proof. First, for $x \in \mathbb{R}^n \backslash \overline{\Omega}$,

$$\begin{aligned} |u(\cdot,T)|^2 &= \left| \int_0^T \int_{\Omega} Z(x-y,T-\tau)f(y,\tau)dyd\tau \right|^2 \\ &= \left| \langle Z, f \rangle_{L^2(\Omega\times[0,T])} \right|^2 \\ &\leq \|Z(x-\cdot,T-\cdot)\|^2_{\left(X_f^{\sigma_1,\sigma_2}\right)'} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} \\ &\leq \|Z(x-\cdot,T-\cdot)\|^2_{\left(X_f^{n/2,1/2}\right)'} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} \\ &\leq \|Z(x-\cdot,T-\cdot)\|^2_{\left(X_f^{n,1}\right)'} \|f\|^2_{X_f^{\sigma_1\cdot\sigma_2}(\Omega,T)}. \end{aligned}$$

Then, according to the Malgrange–Ehrenpreis theorem, the remainder of the proof follows immediately from the fact that $Z(x-\cdot,T-\cdot)$ is smooth on $\mathcal{M}$, i.e., that

$$\int_K \|Z(x-\cdot,T-\cdot)\|^2_{\left(X_f^{n,1}\right)'} dx < \infty$$

for all compact subsets K in $\mathbb{R}^n \backslash \overline{\Omega}$.

We now employ some supporting facts of interest discussed and proven in Chapter 1, and Theorems 1 and 15, of [5] which help us to establish the claim that $\mathcal{Z}$ as a map from $X_f^{\sigma_1,\sigma_2}(\Omega,T)$ into $L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$ has dense range in $L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$. We first note, as given on page 28 of [5], that Z and its adjoint are related through the identity

$$Z(x,y,t,\tau) = Z^*(y,x,\tau,t), \quad t > \tau.$$

We also remark that, again under the assumptions made in Remark 1, Z is a positive kernel, in the sense that the action of Z on any nonnegative source f must be greater than or equal to zero. Next, we examine the homogeneous integral equation for the unknown function $v \in L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$,

$$(\mathcal{Z}^* v)(x,t) = 0. \tag{2.2}$$

Since v is nonnegative, in addition to the action $\mathcal{Z}^*$ on any nonnegative function in its domain, then it follows that (2.2) admits only the trivial solution $v = 0$. Hence, according to Proposition 2.3 on page 46 of [9], $\mathcal{Z}$ as a map from $X_f^{\sigma_1,\sigma_2}(\Omega, T)$ to $L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$ has dense range in $L^2_{+,loc}(\mathbb{R}^n \backslash \overline{\Omega})$. □

We now consider a unique continuation principle for general parabolic differential equations. Friedman [5] has shown the following proposition for $p_0 = (x_0, t_0) \in \mathbb{R}^n \times \mathbb{R}_+$ and

$$\begin{aligned} N(p_0) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R}_+ \,:\, 0 \le t \le t_0, \text{ and the cylinder } \{x\} \times [0,t] \\ \text{centered at } p_0 \text{ is simply connected as } t \text{ increases}\}. \end{aligned}$$

Proposition 2.2. *If $(\partial_t - L_x)u \le 0$ ($(\partial_t - L_x)u \ge 0$) in $\mathbb{R}^n \times \mathbb{R}_+$ and if u has a positive maximum (negative minimum) which is attained at $p_0 = (x_0, t_0)$, then $u(p) = u(p_0)$ for all $p \in N(p_0)$.*

This proposition implies that should u vanish on any open, simply connected subset of $\mathbb{R}^n$ for any $T \ge 0$, then u must vanish everywhere such that u remains a (homogeneous) solution of $(\partial_t - L_x)u = 0$. This property is essential in the range characterization to follow and to our estimation of the carrier support of $u(\cdot, T)|_{\mathcal{M}}$.

We now turn our attention to a few fundamental properties of the restriction of $\mathcal{Z}$ to the sets $\mathcal{M}$ and Ω, which we call $\mathcal{Z}|_{(\mathcal{M},\Omega)}$, and note that it is especially important for our inverse problem of determining the location of the unknown source f given measurements away from, but not necessarily surrounding, it. In what follows, $\mathcal{R}(\mathcal{Z}|_{(\mathcal{M},\Omega)})$ denotes the range of $\mathcal{Z}|_{(\mathcal{M},\Omega)}$.

Proposition 2.3. *Let $\Omega_{1,2} \in \mathbb{R}^n$ be two convex open sets whose closures have empty intersection, and suppose $\overline{\mathcal{M}} \cap (\overline{\Omega}_1 \cup \overline{\Omega}_2) = \emptyset$. Then*

$$\mathcal{R}\left(\mathcal{Z}|_{(\mathcal{M},\Omega_1)}\right) \cap \mathcal{R}\left(\mathcal{Z}|_{(\mathcal{M},\Omega_2)}\right) = \{0\}.$$

Proof. Let $\overline{\Omega}_1 \cap \overline{\Omega}_2 = \emptyset$ and let

$$\begin{aligned} (\partial_t - L_x)u_1 &= f_1, \quad \operatorname{supp} f_1(\cdot,t) \subset \Omega_1, \\ (\partial_t - L_x)u_2 &= f_2, \quad \operatorname{supp} f_2(\cdot,t) \subset \Omega_2 \end{aligned}$$

such that u_1 and u_2 do not vanish on $\mathcal{M}$. Next, let $v = u_1 - u_2$. Then v satisfies

$$\begin{aligned} (\partial_t - L_x)v &= f_1 \quad \text{on } \Omega_1, \\ (L_x - \partial_t)v &= f_2 \quad \text{on } \Omega_2, \end{aligned}$$

so that $v = u_1$ on Ω_1 and $v = -u_2$ on Ω_2. Hence, $u_{1,2}$ vanish on $\Omega_{2,1}$, respectively. Then, by Proposition 2.2, u_1 and u_2 must also vanish on $\mathbb{R}^n \backslash \Omega_{2,1}$. Hence, $u_{1,2} \equiv 0$ on $\mathcal{M}$. This is a contradiction. □

We now consider the remainder of the fundamental properties of the restricted mapping $\mathcal{Z}|_{(\mathcal{M},\Omega)}$.

PROPOSITION 2.4. *Let $\mathcal{M}$ and Ω be open subsets of $\mathbb{R}^n$ such that $\Omega \supset \operatorname{supp} f(\cdot, t)$ for all time $t \in [0,T]$ and assume $\overline{\mathcal{M}} \cap \overline{\Omega} = \emptyset$. Then, for all $T \geq 0$, $\mathcal{Z}|_{(\mathcal{M},\Omega)} : X_f^{\sigma_1,\sigma_2}(\Omega, T) \to L^2_+(\mathcal{M})$ is a compact linear map and has dense range in the latter space.*

Proof. Let $f \in X_f^{\sigma_1,\sigma_2}(\Omega, T)$. Since $Z(x - \cdot, t - \cdot) \in C^\infty_+(\mathbb{R}^n \backslash \overline{\Omega} \times [0,T])$ by the Malgrange–Ehrenpreis theorem, for each $T \geq 0$ we have

$$\begin{aligned}
\|(\mathcal{Z}|_{(\mathcal{M},\Omega)} f)(\cdot, T)\|^2_{L^2_+(\mathcal{M})} &= \int_{\mathcal{M}} \left| \int_0^T \int_\Omega Z(x-y, T-\tau) f(y,\tau) dy d\tau \right|^2 dx \\
&= \int_{\mathcal{M}} \left| \langle Z, f \rangle_{L^2_+(\mathbb{R}^n \times [0,T])} \right|^2 dx \\
&\leq \int_{\mathcal{M}} \|Z(x - \cdot, T - \cdot)\|^2_{\left(X_f^{\sigma_1,\sigma_2}\right)'} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} dx \\
&\leq \int_{\mathcal{M}} \|Z(x - \cdot, T - \cdot)\|^2_{\left(X_f^{n/2,1/2}\right)'} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} dx \\
&\leq \int_{\mathcal{M}} \|Z(x - \cdot, T - \cdot)\|^2_{\left(X_f^{n,1}\right)'} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} dx \\
&\leq C_{1,n,\Omega,\mathcal{M}} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)},
\end{aligned}$$

where $C_{1,n,\Omega,\mathcal{M}}$ is a constant given by

$$C_{1,n,\Omega,\mathcal{M}} = n\mu(\mathcal{M}) \max_{\mathcal{M}} \|Z(x - \cdot, T - \cdot)\|^2_{\left(X_f^{n,1}\right)'} < \infty.$$

Now, since

$$\|Z(x - \cdot, T - \cdot) f\|^2_{H^n_+(\mathcal{M})} = \|Z(x - \cdot, T - \cdot) f\|^2_{L^2_+(\mathcal{M})} + \left\| \sum_{k,i=1}^n \partial^k_{x_i} Z(x - \cdot, T - \cdot) f \right\|^2_{L^2_+(\mathcal{M})},$$

Dirichlet's theorem and the same arguments above imply that

$$\left\| \sum_{k,i=1}^n \partial^k_{x_i} Z(x - \cdot, T - \cdot) f \right\|^2_{L^2_+(\mathcal{M})} \leq C_{2,n,\Omega,\mathcal{M}} \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)},$$

where

$$C_{2,n,\Omega,\mathcal{M}} = n\mu(\mathcal{M}) \max_{\substack{\mathcal{M} \\ 1 \leq k \leq n}} \|Z(x - \cdot, T - \cdot)\|^2_{\left(X_f^{k,1}\right)'} < \infty.$$

Hence, $\mathcal{Z}$ is bounded between $\overset{o}{H}{}^{\sigma_1}_+(\overline{\Omega}) \bigotimes \overset{o}{H}{}^{\sigma_2}_+([0,T])$ and $H^n_+(\mathcal{M})$. Finally, since the inclusion maps

$$H^n_+(\mathcal{M}) \overset{i}{\hookrightarrow} H^{n-1}_+(\mathcal{M}) \overset{i}{\hookrightarrow} \cdots \overset{i}{\hookrightarrow} H^1_+(\mathcal{M}) \overset{i}{\hookrightarrow} L^2_+(\mathcal{M})$$

are compact (see, for example, Theorem 6.98 in [15]), so must it be that $\mathcal{Z}$ is compact from the source space $\overset{o}{H}{}^{\sigma_1}_+(\overline{\Omega}) \bigotimes \overset{o}{H}{}^{\sigma_2}_+([0,T])$ into $L^2_+(\mathcal{M})$ for each T.

Lastly, $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ has dense range in $L^2_+(\mathcal{M})$ since, according to Proposition 2.1, the adjoint equation posed on $\mathbb{R}^n \times \mathbb{R}_+$,

$$\mathcal{Z}^* v(x,t) = 0,$$

implies $v(x,t)$ vanishes throughout $\mathbb{R}^n \times \mathbb{R}$, and again according to Proposition 2.3 on page 46 of [9], while linearity of $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ is clear. □

Of interest to us in the following section on the inverse problem is the existence of the Hilbert adjoint of this restricted operator. This result follows as a corollary to the previous proposition. That is, we obtain the following.

COROLLARY 2.5 (Hilbert adjoint). *Let $\mathcal{M}$ and Ω be open subsets of $\mathbb{R}^n$ such that $\Omega \supset \operatorname{supp} f(\cdot,T)$ and assume $\overline{\mathcal{M}} \cap \overline{\Omega} = \emptyset$. Then, for all $T \in \mathbb{R}_+$, $\mathcal{Z}|^*_{(\mathcal{M},\Omega)} : L^2_+(\mathcal{M}) \to X^{\sigma_1,\sigma_2}_f(\Omega,T)$ exists as a bounded linear map. Moreover,*

$$\left(\mathcal{Z}|^*_{(\mathcal{M},\Omega)} u\right)(x,T) = \int_{\mathcal{M}} Z(z-x,T)u(z,T)dz.$$

Proof. Let $f \in X^{\sigma_1,\sigma_2}_f(\Omega,T)$ and suppose $u(\cdot,T) \in L^2_+(\mathcal{M})$. Since $\mathcal{M}$, Ω, and T are all bounded, then each of the spaces on which they integrate is sigma finite, and hence we may interchange all orders of integration. We arrive at the formula for $\mathcal{Z}|^*_{(\mathcal{M},\Omega)}$ by noting that

$$\begin{aligned}
\langle u(\cdot,T)|_{\mathcal{M}}, \mathcal{Z}|_{(\mathcal{M},\Omega)} f\rangle_{L^2_+(\mathcal{M})} &= \int_{\mathcal{M}} u(x,T) \int_0^T \int_{\Omega} Z(x-y,T-\tau) f(y,\tau)dyd\tau dx \\
&= \int_0^T \int_{\Omega} \int_{\mathcal{M}} u(x,T) Z(x-y,T-\tau)dxf(y,\tau)dyd\tau \\
&= \langle \mathcal{Z}|^*_{(\mathcal{M},\Omega)} u(\cdot,T)|_{\mathcal{M}}, f\rangle_{L^2_+(\mathbb{R}^n\times[0,T])}.
\end{aligned}$$

Hence,

$$\left(\mathcal{Z}|^*_{(\mathcal{M},\Omega)} u\right)(x,T) = \int_{\mathcal{M}} Z(z-x,T)u(z,T)dz.$$

Similarly, we complete the proof by noting that

$$\begin{aligned}
\left|\langle u(\cdot,T)|_{\mathcal{M}}, \mathcal{Z}|_{(\mathcal{M},\Omega)} f\rangle_{L^2_+(\mathcal{M})}\right|^2 &= \left|\int_{\mathcal{M}} u(x,T) \int_0^T \int_{\Omega} Z(x-y,T-\tau) f(y,\tau)dyd\tau dx\right|^2 \\
&= \left|\int_0^T \int_{\Omega} \int_{\mathcal{M}} u(x,T) Z(x-y,T-\tau)dxf(y,\tau)dyd\tau\right|^2 \\
&= \left|\langle \mathcal{Z}|^*_{(\mathcal{M},\Omega)} u(\cdot,T)|_{\mathcal{M}}, f\rangle_{L^2_+(\mathbb{R}^n\times[0,T])}\right|^2 \\
&\le \|\mathcal{Z}|^*_{(\mathcal{M},\Omega)} u(\cdot,T)|_{\mathcal{M}}\|^2_{\left(X^{\sigma_1,n/2}_f(\Omega,T)\right)'} \|f\|^2_{X^{\sigma_1,\sigma_2}_f(\Omega,T)} \\
&< \infty. \quad \square
\end{aligned}$$

To ease notation, in what follows $\mathcal{R}\left(\mathcal{Z}|_{X^{\sigma_1,\sigma_2}_f(\Omega,T)}\right)$ denotes the range of $\mathcal{Z}$ restricted to positive sources supported on the set Ω and observations limited to $\mathcal{M}$. Additionally, we will use the shorthand $\mathcal{Z}$ to denote $\mathcal{Z}|_{(\mathcal{M},\Omega)}$. Finally, if Ω is a set in

$\mathbb{R}^n$, then $N_\epsilon(\Omega)$ denotes the union of the set Ω and a neighborhood of its boundary, so that Ω is strictly contained in $N_\epsilon(\Omega)$ for each $\epsilon > 0$.

We are now able to fully characterize the range of $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ acting on distributions in the space $X_f^{\sigma_1,\sigma_2}(\Omega, T)$ for various sets Ω, without specifying their regularity parameter $\sigma = (\sigma_1, \sigma_2) \in \mathbb{R}^2$. This will prove of much use in the numerical implementation and of the forthcoming result. That is, we have the following.

PROPOSITION 2.6. *Let $\Omega_{1,2}$ be bounded convex subsets of $\mathbb{R}^n$ with smooth boundaries. Then*

$$\mathcal{R}(\mathcal{Z}|_{X_f^{\sigma_1,\sigma_2}(\overline{\Omega}_1\cap\overline{\Omega}_2,T)}) \subset \mathcal{R}(\mathcal{Z}|_{X_f^{\sigma_1,\sigma_2}(\overline{\Omega}_1,T)})\cap\mathcal{R}(\mathcal{Z}|_{X_f^{\sigma_1,\sigma_2}(\overline{\Omega}_2,T)}) \subset \mathcal{R}(\mathcal{Z}|_{X_f^{0,0}(N_\epsilon(\Omega_1\cup\Omega_2),T)}).$$

Proof. Let $t \in [0, T]$ and let $\Omega_{1,2}$ be as stated above. The left lower containment follows from the fact that for $f_{1,2} \in X_f^{\sigma_1,\sigma_2}(\Omega_{1,2}, T)$, the trivial extension of the form

$$\tilde{f}_1(x,t) = \begin{cases} f_1(x,t), & (x,t) \in \overline{\Omega}_1 \times [0,T], \\ 0, & x \notin \overline{\Omega}_1, \end{cases}$$

ensures the containment.

Next, in the spirit of the proof of Lemma 3.6 in [10] let $f_{1,2} \in X_f^{\sigma_1,\sigma_2}(\Omega_{1,2}, T)$ such that for each T, $\mathcal{Z}f_1 = \mathcal{Z}f_2 = u|_\mathcal{M}$. Then, by the unique continuation principle given in Proposition 2.2, we have

$$(\mathcal{Z}f_1)(x,T) = (\mathcal{Z}f_2)(x,T), \quad x \in \mathbb{R}^n\backslash(\Omega_1 \cup \Omega_2).$$

Let $\phi \in C^\infty(\mathbb{R}^n \times [0, T])$ be a smooth cut-off function satisfying

$$\phi(x,t) = \begin{cases} 1, & (x,t) \in \mathbb{R}^n\backslash N_\epsilon(\Omega_1 \cap \Omega_2) \times [0,T], \\ 0, & (x,t) \in N_{\epsilon/2}(\Omega_1 \cap \Omega_2) \times [0,T]. \end{cases}$$

Then, for

$$v(x,t) = \begin{cases} \phi(x,t)u_1(x,t), & (x,t) \in \mathbb{R}^n\backslash\Omega_1 \times [0,T], \\ \phi(x,t)u_2(x,t), & (x,t) \in \mathbb{R}^n\backslash\Omega_2 \times [0,T], \\ 0, & (x,t) \in \Omega_1 \cap \Omega_2 \times [0,T], \end{cases}$$

it follows that $v \in C_+^\infty(\mathbb{R}^n \times [0, T])$ and that $(\partial_t - L_x)v = f_3 \in C_+^\infty(\mathbb{R}^n \times [0, T])$ such that

$$(\mathcal{Z}f_3)(x,T) = u_1(x,T) = u_2(x,T), \quad x \notin \Omega_1 \cap \Omega_2.$$

More importantly,

$$(\mathcal{Z}f_3)(x,T) = u_1(x,T) = u_2(x,T), \quad x \in \mathcal{M},$$

where $f_3 \in X_f^{0,0}(N_\epsilon(\Omega_1 \cap \Omega_2), T)$. □

3. The inverse source problem. We present our main result (theorem) concerning the estimation of the time-varying carrier support in this section. The main result presented here owes itself in part to Picard's theorem. This theorem essentially provides a denumerable representation of a compact linear operator A between two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ in terms of the operator's singular system, as well as a means to assess whether a given element of the second space $\mathcal{H}_2$ is also an element of the closure of the range of A. We take a moment to state the theorem and refer to [1] for its proof and further commentary.

Theorem 3.1 (Picard). *Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact linear operator from the Hilbert space $\mathcal{H}_1$ into the Hilbert space $\mathcal{H}_2$ with singular system $\{\lambda_n, \varphi_n, \psi_n\}_{n=1}^{\infty}$, i.e.,*

$$A\varphi_n = \lambda_n \psi_n$$

and

$$A^* \psi_n = \lambda_n \varphi_n,$$

and let $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathcal{H}_2$. Then the equation $Af = g$ is solvable if and only if $g \in N(A^*)^{\perp}$ and

$$\sum_{n=1}^{\infty} \frac{|\langle g, \psi_n \rangle|^2}{\lambda_n^2} < \infty.$$

Moreover, any $\tilde{f}$ of the form

$$\tilde{f} = \sum_{n=1}^{\infty} \frac{\langle g, \psi_n \rangle}{\lambda_n} \varphi_n$$

solves $A\tilde{f} = g$.

Given Picard's theorem and our previous range characterizations of the operator $\mathcal{Z}|_{(\mathcal{M},\Omega)}$, presented in the previous section, we now have a test which can determine whether the carrier support of the field u on $\mathcal{M}$ at time T is fully within some set of interest Ω by means of testing the convergence of the sum

$$\|\tilde{f}\|^2_{X_f^{\sigma}(\Omega,T)} = \sum_{n=1}^{\infty} \left| \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T) \rangle}{\lambda_n^{(\sigma)}} \right|^2,$$

where

$$\tilde{f}(x,t) = (\mathcal{Z}|^*_{(\mathcal{M},\Omega)} \mathcal{Z}|_{(\mathcal{M},\Omega)})^{-1} \mathcal{Z}|^*_{(\mathcal{M},\Omega)} u(\cdot,T)|_{\mathcal{M}}, \quad (x,t) \in \Omega \times [0,T],$$

and where the functions $\psi_n^{(\sigma)}(\cdot,T)$, $\varphi_n^{(\sigma)}(\cdot,T)$, and $\lambda_n^{(\sigma)}$—which depend on the regularity parameter $\sigma = (\sigma_1, \sigma_2)$ and the sets of interest $\mathcal{M}$ and Ω—are defined through the relationships

$$\left(\mathcal{Z}|^*_{(\mathcal{M},\Omega)} \psi_n^{(\sigma)} \right)(x,T) = \lambda_n^{(\sigma)} \varphi_n^{(\sigma)}(x,t), \quad 0 \le t \le T,$$

and

$$\left(\mathcal{Z}|_{(\mathcal{M},\Omega)} \varphi_n^{(\sigma)} \right)(x,T) = \lambda_n^{(\sigma)} \psi_n^{(\sigma)}(x,T).$$

If the sum does not converge, then we are able to conclude that the carrier support of $u(\cdot,T)|_{\mathcal{M}}$ at time T is not fully within the test region Ω.

We formalize this statement with the main theorem of this section.

THEOREM 3.2 (carrier support). *Let $f \in X_f^{\sigma_1,\sigma_2}(\Omega, T)$. Suppose further that $\Omega \subset \mathbb{R}^n$ is bounded with a smooth boundary and let $\mathcal{Z}|^*_{(\mathcal{M},\Omega)}$ denote the Hilbert adjoint of $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ such that $\mathcal{Z}|^*_{(\mathcal{M},\Omega)} : L^2_+(\mathcal{M}) \to X_f^{\sigma_1,\sigma_2}(\Omega, T)$. Suppose further that*

$$\left(\mathcal{Z}|_{(\mathcal{M},\Omega)}\varphi_n^{(\sigma)}\right)(x,T) = \lambda_n^{(\sigma)}\psi_n^{(\sigma)}(x,T).$$

Then, for each fixed T,

$$\text{carr supp } u(\cdot,T)|_{\mathcal{M}} \subset \Omega \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} \left| \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle}{\lambda_n^{(\sigma)}} \right|^2 < \infty.$$

Proof. Suppose carr supp $u(\cdot,T)|_{\mathcal{M}} \subset \Omega$. Then, by definition of the carrier support, there exists a source $f \in X_f^{\sigma_1,\sigma_2}(\Omega,T)$ supported on a subset of Ω such that for each T, $(\mathcal{Z}|_{(\mathcal{M},\Omega)}f)(x,T) = u(\cdot,T)|_{\mathcal{M}}$, where $u(\cdot,T)|_{\mathcal{M}} \in \mathcal{R}\left(\mathcal{Z}|_{X_f^{\sigma_1,\sigma_2}(\Omega,T)}\right)$. Since $\mathcal{Z}$ is a compact linear operator between the two Hilbert spaces exhibited in the previous proposition, it admits the representation

$$\mathcal{Z} = \sum_{n=1}^{\infty} \lambda_n^{(\sigma)}\psi_n^{(\sigma)} \otimes \varphi_n^{(\sigma)}$$

such that the action of $\mathcal{Z}$ on f may be written as

$$(\mathcal{Z}f)(x,T) = \sum_{n=1}^{\infty} \lambda_n^{(\sigma)}\langle f, \varphi_n^{(\sigma)}\rangle \psi_n^{(\sigma)}(x,T), \quad x \in \mathcal{M}.$$

Since the left and right eigenfunctions, $\psi_n^{(\sigma)}$ and $\varphi_n^{(\sigma)}$, satisfy

$$\mathcal{Z}^*\psi_n^{(\sigma)} = \lambda_n^{(\sigma)}\varphi_n^{(\sigma)}$$

we note that

$$\begin{aligned}
\langle f, \varphi_n^{(\sigma)}\rangle &= \int_0^T \int_\Omega f(x,\tau)\varphi^{(\sigma)}(x,\tau)dxd\tau \\
&= \langle f, \mathcal{Z}^*\psi_n^{(\sigma)}/\lambda_n^{(\sigma)}\rangle \\
&= \overline{1/\lambda_n^{(\sigma)}}\langle \mathcal{Z}f, \psi_n^{(\sigma)}\rangle \\
&= \overline{1/\lambda_n^{(\sigma)}}\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle.
\end{aligned}$$

Bessel's inequality states

$$\sum_{n=1}^{\infty} |\langle f, \varphi_n^{(\sigma)}\rangle|^2 \leq \|f\|^2_{X_f^{\sigma_1,\sigma_2}(\Omega,T)} < \infty.$$

Hence,

$$\sum_{n=1}^{\infty} \left| \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle}{\lambda_n^{(\sigma)}} \right|^2 < \infty.$$

Now, suppose $u(\cdot,T)|_{\mathcal{M}} \in \mathcal{R}(\mathcal{Z}|_{X_f^{\sigma_1,\sigma_2}(\Omega,T)})$, where the closure of the latter space is $L_+^2(\mathcal{M})$, which we established in Proposition 2.1, and suppose the Picard sum

$$\sum_{n=1}^{\infty} \left| \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle}{\lambda_n^{(\sigma)}} \right|^2$$

converges. Let f be any source of the form

$$f(x,t) = \sum_{n=1}^{\infty} \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle}{\lambda_n^{(\sigma)}} \varphi_n^{(\sigma)}(x,t), \quad (x,t) \in \Omega \times [0,T].$$

Then, since each $\varphi_n^{(\sigma)}$ is supported on Ω, any such source will have similar such support. Also, its nontrivial image, a fact from Proposition 2.3, under $\mathcal{Z}$ is in $L_+^2(\mathcal{M})$ for each T. Finally,

$$\|\mathcal{Z}f\|_{L_+^2(\mathcal{M})}^2 = \sum_{n=1}^{\infty} |\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle|^2 \le \|u(\cdot,T)\|_{L_+^2(\mathcal{M})}^2 < \infty.$$

Hence, carr supp $u(\cdot,T)|_{\mathcal{M}} \subset \Omega$. □

Our main result provides us with a reconstruction algorithm which can determine the time-dependent carrier support of the field $u|_{\mathcal{M}}$. We describe this algorithm in the form of the following

Corollary 3.3. *Let Ω be an open bounded convex subset of $\mathbb{R}^n$, and let $\{\lambda_n^{(\sigma)}, \psi_n^{(\sigma)}, \varphi_n^{(\sigma)}\}$ be the singular system for $\mathcal{Z}|_{(\mathcal{M},\Omega)}$. Then*

$$\text{carr supp } u(\cdot,T)|_{\mathcal{M}} = \bigcap \Omega \quad \textit{such that} \quad \sum_{n=1}^{\infty} \left| \frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_n^{(\sigma)}(\cdot,T)\rangle}{\lambda_n^{(\sigma)}} \right|^2 < \infty.$$

Proof. Let Ω be an open bounded convex set and suppose that the infinite series in Corollary 3.3 converges. Then there exists a source $g^{(\Omega)}$, depending on Ω, in $X_f^{\sigma_1,\sigma_2}(\Omega,T)$ such that $\mathcal{Z}|_{(\mathcal{M},\Omega)} g^{(\Omega)} = u(\cdot,T)|_{\mathcal{M}}$. Taking the intersection of the supports of all such g's then yields

$$\bigcap \Omega = \bigcap_{\mathcal{Z}|_{(\mathcal{M},\Omega)} g^{(\Omega)} = u|_{\mathcal{M}}} \text{supp ch } g^{(\Omega)} = \text{carr supp } u(\cdot,T)|_{\mathcal{M}}. \qquad \square$$

This summability test offers a theoretical and computational basis for the determination and ultimate reconstruction of the carrier support of $u(\cdot,T)|_{\mathcal{M}}$. Numerically speaking, however, there is some issue of how to actually sum the infinite series. The fact that the operator $\mathcal{Z}|_{(\mathcal{M},\Omega)}$ is compact and smoothing tells us that zero is either an eigenvalue or a point of accumulation. It turns out that zero is a point of accumulation and so the singular values $\lambda_n^{(\sigma)}$ rapidly converge to zero, thereby creating a computational instability problem if we seek to sum the series numerically. Hence, either the summability test needs to be regularized in some manner or we need to pursue another approach which is linked to the infinite series.

4. Numerical examples. We now consider two numerical examples which demonstrate the ability of the proposed theorem to locate the positions of two localized sources. In the first case, we consider the delta-function impulse source $f(x,t) =$

$\delta_p(x)\delta_0(t) \in \mathcal{E}'(\mathbb{R}) \otimes \mathcal{E}'(\mathbb{R})$. Here, we study the two problems of having knowledge of the scalar diffusion field away from the source with a nontrivial convection field flowing in the downstream sense; i.e., the measurement set is, say, to the left of the source, while the flow field moves from the left to the right. We examine this problem in the two cases where, in the first, we have values of u at some fixed time larger than zero sampled uniformly on our measurement interval, and where, in the second, these values are measured in a nonuniform fashion. In short, in both situations, the minimum of the logarithm of the truncated moving Picard sum is evidently observable and location of the impulse point source is readily accomplished.

In the second case, we study another convective problem, specifically in two-dimensional space $\mathbb{R}^2$. In this case we allow the source to be a sustained one in time—once it has "turned on"—and allow it to move through space along a T^*-period track or orbit $\gamma(t)$, of lifetime T^*, i.e.,

$$f(x,t) = \delta_{\gamma(t)}(x)(H_0(t) - H_{T^*}(t)).$$

For this problem we present only the case with uniform measurements on $\mathcal{M}$, now in $\mathbb{R}^2$, yet we are able to clearly demonstrate that the instantaneous tracking, or location, of the localized source may be done in a very robust fashion, as evidenced by the comparison of the exhibited true and reconstructed source trajectories.

For simplicity in the numerical generation of the data and the numerical implementation of the main results of this article we assume the coefficients in parabolic operator L_x take the form

$$\begin{aligned} &a = I \text{ on } \mathbb{R}^n, \quad n = 1,2,\\ &b = \begin{cases} -1 & \text{on } \mathbb{R}^1,\\ -(1,0) & \text{on } \mathbb{R}^2, \end{cases}\\ &c = 0. \end{aligned}$$

This means our governing equations of interest in the one- and two-dimensional cases are

$$\left(\partial_t + \partial_x - \partial_x^2\right) u(x,t) = f(x,t), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+,$$

and

$$\left(\partial_t + \partial_x - \partial_x^2 - \partial_y^2\right) u(x,y,t) = f(x,y,t), \quad (x,y,t) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+.$$

REMARK 3. *In the following numerical experiments the variable c will now, instead, denote a coordinate value in either $\mathbb{R}$ or $\mathbb{R}^2$, depending on whether we are addressing the inverse problem on the real line or on the real plane, which will be evidently clear from the associated context. This coordinate value represents the center, and hence the preferable letter c, of certain test domains of interest (Ω_c), which will be moved around the larger space of interest and will be centered at the various points c to follow.*

Location, and tracking of the moving source, is accomplished by covering the totality of the search space of interest Ω_s with candidate test domains of the forms

$$\Omega_c^{(1,2)} = \Omega_0^{(1,2)} + \{c_j\},$$

where in $\mathbb{R} \ni c_j$ we have

$$\Omega_0^{(1)} = [0, 1/10]$$

and in $\mathbb{R}^2 \ni c$ we define

$$\Omega_0^{(2)} = [0, 1/10] \times [0, 1/10],$$

and by seeking to minimize the objective function

$$J(\Omega_c) = \|\mathcal{Z}|^{\dagger}_{(\mathcal{M},\Omega_c)} u(\cdot, T)|_{\mathcal{M}}\|^2.$$

Here, † denotes the pseudoinverse.

In summary, when the test set Ω_c fully contains the carrier support of the source, then the objective function should take on small values; contrarily, when such a test domain does not fully contain the source it should be singular. We acknowledge that the phrasing "small values" is indeed rather ambiguous; however, in the numerical examples to follow, we observe a global minimum value of the objective function when the test domain is exactly centered on the support of the true source. Furthermore, we add that the test domain position centers are taken as

$$\mathbb{R} \ni c_j = \{-10 + j/20\}, \quad j = 0, 1, 2, \ldots, 400,$$

and

$$\mathbb{R}^2 \ni c_j = (-2 + j/20, 0 + j/20), \quad j = 0, 1, 2, \ldots, 80.$$

We end this section, which discusses the overall strategy pursued in the generation of the numerical evaluation of the results presented in this article, with a few words on the simulated data used in this evaluation. In summary, we discretize the standard action of $\mathcal{Z}f$ and employ a Newton–Cotes-type numerical integration scheme to generate its approximation on discretely sampled points of $\mathcal{M}$. More importantly, so as to not perpetrate an inverse crime, we corrupt these approximations with considerable white noise in the levels of approximately 5%, 10%, and 30%.

4.1. Locating a stationary impulse source in $\mathbb{R}$. In this section the scalar field is generated by the discretization scheme (via the standard rectangle rule) of the integral

$$(4.1)\quad u(x,T) = \int_0^T \int_{\mathbb{R}} \frac{e^{-\frac{|x-(T-\tau)-y|^2}{4(T-\tau)}}}{\sqrt{4\pi(T-\tau)}} \delta_p(y)\delta_0(\tau)dyd\tau = \frac{1}{\sqrt{4\pi T}} \int_{\mathbb{R}} e^{-\frac{|x-T-y|^2}{4T}} \delta_p(y)dy$$

and again subsequently corrupted by various levels of white noise to avoid committing the standard inverse crime. We examine the data taken over the discretely sampled uniform domain,

$$\mathcal{M}_1 = \{-1.0, -0.8, -0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8, 1.0\} \subset \mathbb{R},$$

and the nonuniform domain,

$$\mathcal{M}_2 = \{-0.14, -0.178, -.21, -.23, -.49, -0.8, -0.01, 0.62, 0.81, 0.90, 1.0\} \subset \mathbb{R}.$$

We examine three instances of noise-corrupted data, in the amounts 10, 20, and 30 dB—which amounts to 32.0%, 10.0%, and 3.2% relative signal-to-noise ratio. For relative time $T = 10$ after impulse release, for the discrete samplings taken on $\mathcal{M}_1$ and $\mathcal{M}_2$, we observe that the running truncated Picard series are each minimized

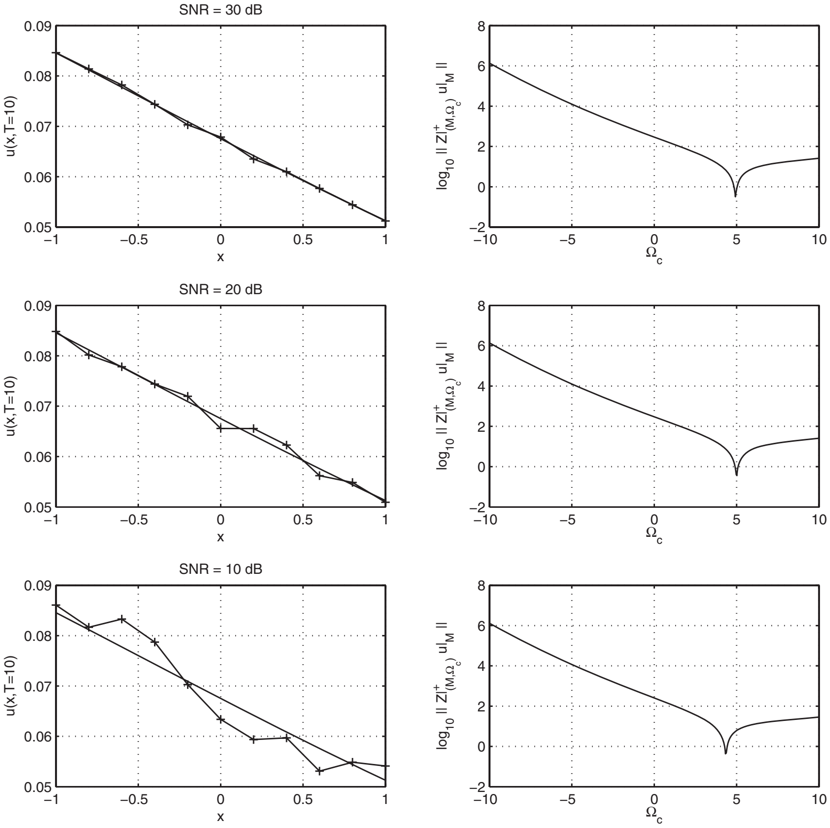


FIG. 1. *The exact and measured (noise-corrupted) diffusion fields and the norm of the associated truncated Picard test series on a uniform measurement grid.*

at, or very near, the point of the impulse release, which is $p = 5.0$ in all cases. This observation is very helpful in establishing the fact that the discrete samplings taken on general sets of observation $\mathcal{M}$ are not restricted to such things as equally spaced gridded points. Rather, any collection of discrete points which are coplanar in $\mathbb{R}^n$ will suffice. Moreover, numerical investigations have shown that larger random samplings yield better conditioned systems than for the analogous case of equally separated points, when the number of samplings is the same in both instances. For commercial purposes, this is both necessary and highly advantageous. Figures 1 and 2 demonstrate the efficacy of locating the localized source, again at $p = 5$, when we have uniform measurements on $\mathcal{M}_1$ and nonuniform ones on $\mathcal{M}_2$.

4.2. Locating a sustained moving source in $\mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. As in the previous subsection, we begin with the description of the integral which we discretize, and corrupt with ample random noise, which provides our simulated data. Here, we present the numerical method used to construct the numerical solution to the two-dimensional forward problem. We begin by considering the two-dimensional heat equation with a time-dependent source $g((x,y) - \gamma(t))s(t)$ located at $(x_s(t), y_s(t)) = \gamma(t)$ and given by

$$\left(\partial_t + \partial_x - \partial_x^2 - \partial_y^2\right) u(x,y,t) = g((x,y) - \gamma(t))s(t).$$

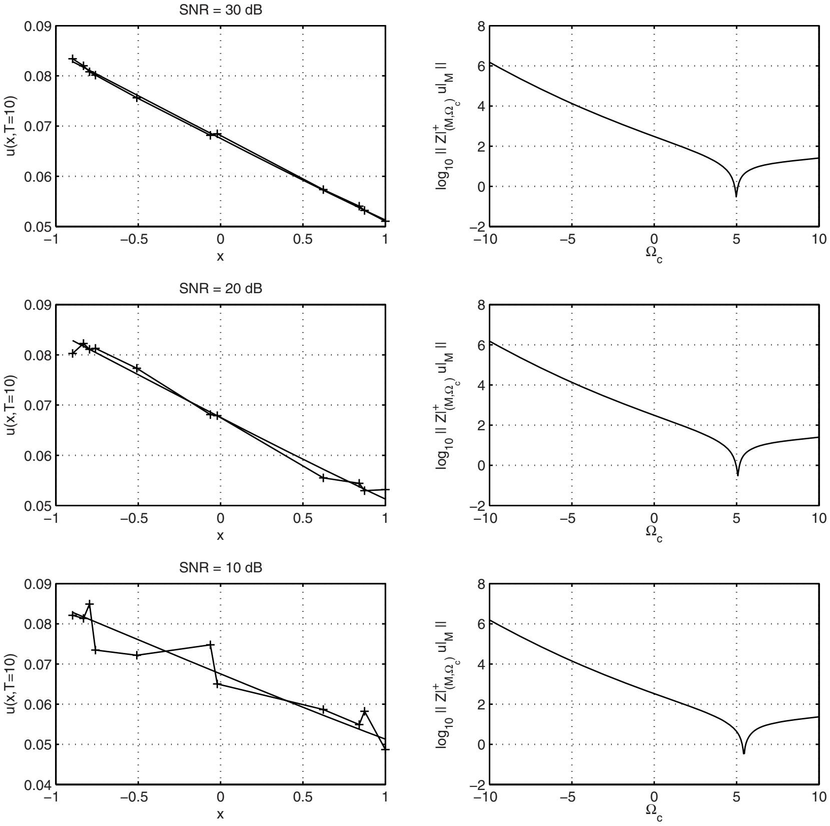


FIG. 2. *The exact and measured (noise-corrupted) diffusion fields and the norm of the associated truncated Picard test series on a nonuniform (random) measurement grid.*

Following [2] the fundamental solution for the above operator (defined over all space) is given by

$$Z(x-\xi_1, y-\xi_2, t-\tau) = \frac{1}{4\pi(t-\tau)} e^{-\frac{((x-\xi_1)-(t-\tau)+(y-\xi_2))^2}{4(t-\tau)}}.$$

In our two-dimensional simulation problem we are interested in solving the inverse source problem for the convective-diffusion equation on the half space

$$\Pi_+ = [-\infty, \infty] \times [0, \infty],$$

where we assume there is no flux of the field over the boundary $\{y = 0\}$, i.e., $\partial_y u(x, 0, t) = 0$ for all x and t. This problem then resembles (in one less dimension) the problem of some form of elevated source release in, say, a large but local neighborhood of the atmosphere and its contact with the ground. Then, in this case, using the method of images (see [8, 2]), we form the Green's function

$$\tilde{Z}(x-\xi_1, y-\xi_2, t-\tau) = Z(x-\xi_1, y-\xi_2, t-\tau) + Z(x-\xi_1, y+\xi_2, t-\tau),$$

which satisfies our boundary condition.

To solve the diffusion equation with a time- and spatially dependent source term one can integrate it against such a Green's function to obtain

$$u(x,y,T) = \int_0^T \int_{\mathbb{R}_{\xi_1}} \int_{\mathbb{R}_{\xi_2}} \tilde{Z}(x-\xi_1, y-\xi_2, T-\tau) f(\xi_1,\xi_2,\tau) d\xi_1 d\xi_2 d\tau. \tag{4.2}$$

Due to the the weak singularity at $\tau = T$, the numerical integration requires special treatment. To perform the integration we used a member of the semiopen quadrature rules [18] that do not explicitly evaluate their integrand at the limit of the integration range where the singularity exists. The explicit scheme selected for this investigation is given by (for an integrand $f(x)$ that is singular at the left endpoint x_1)

$$\int_{x_1}^{x_N} f(x)dx = h\left[\frac{23}{12}f_2 + \frac{7}{12}f_3 + f_4 + f_5 + \cdots + f_{N-2} + \frac{13}{12}f_{N-1} + \frac{5}{12}f_N\right]. \tag{4.3}$$

Note that substituting the expression $v = T - \tau$ in (4.2) results in the singular limit at the left-hand endpoint, to which (4.3) can be applied.

We employed the source moving along the figure-eight-like lemniscate over one period (of 10 units), which then goes extinct, i.e., in this case

$$f(x,y,t) = \delta_{(\cos\frac{2\pi t}{10}, \sin\frac{4\pi t}{10})}(x,y)(H_0(t) - H_{10}(t)), \quad (x,y) \in \Pi_+.$$

Moreover, our data was then sampled on the grid of points

$$\mathcal{M} = \{0, 1/2, 1\} \times \{0, 1/2, 1\}$$

at snapshots in time corresponding to every 1/10 of unitless time. We demonstrate the outcome of this numerical experiment in Figures 3, 4, 5, and 6. We use a signal-to-noise level of 25 dB, which again corresponds to 5.6% relative error between the signal strength and that of the additive white noise. The truncation value N which defines the dimension of the singular system used in the truncated Picard series test is chosen so that the singular values obey $\lambda_n^{(0,0)} < 10^{-4}$ for each $n > N$, that is, our TSVD tolerance is 0.0001. We used this value as we found the singular values $\lambda_n^{(0,0)}$ rapidly approach zero after this value. Hence, our method of regularization is based on the philosophy of principle component analysis. That is, we found the tolerance criterion $\lambda_n^{(0,0)} \geq 10^{-4}$ yielded the principle (dominant) components of the (pseudoinverted) operator in question.

Figure 3 shows the entire ensemble of the reconstructed truncated Picard series over all the test domains at time $T = 1.0$. The colored surface is the value of the logarithm of the truncated series

$$\|\tilde{f}_j\|_{X_f^0(\Omega,T)} = \left(\sum_{n=1}^{N}\left|\frac{\langle u(\cdot,T)|_{\mathcal{M}}, \psi_{n,j}^{(0)}(\cdot,T)\rangle}{\lambda_{n,j}^{(0)}}\right|^2\right)^{1/2}.$$

Figure 4 shows the same objective-type function as the previous one in Figure 3, only here we look from beneath to better observe the minimum located near the coordinate pair $(2\pi/10, 4\pi/10)$.

After obtaining the global minimum of the objective function which locates the (potentially mobile) source, we then use smaller moving, time-adaptive test domains of interest to locate the source, knowing with 100% certainty that the source and its

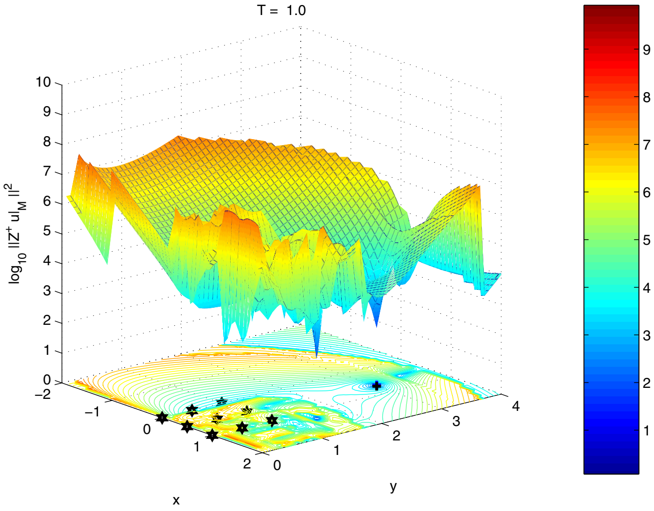


FIG. 3. *Numerical truncated Picard tests at time $T = 1.0$ for the localized two-dimensional moving source over the fully tessellated domain.*

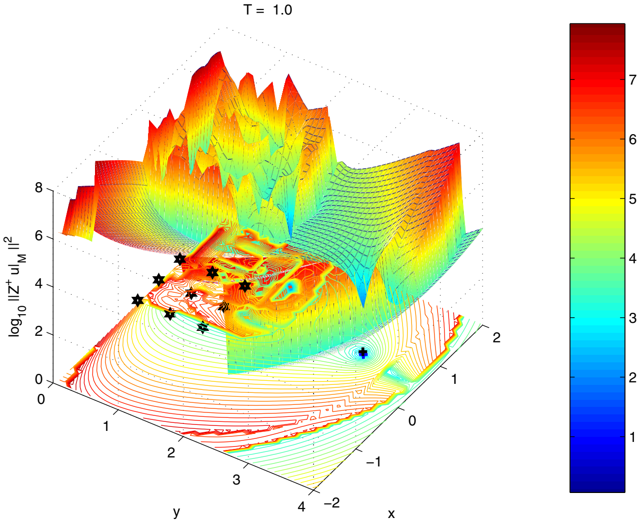


FIG. 4. *Inverted view of the former numerical truncated Picard tests at time $T = 1.0$ for the localized two-dimensional moving source over the fully tessellated domain.*

carrier support reside with each such test domain. Using the most previous estimates of the coordinates of the autonomous source, $(\hat{x}_{j-1}, \hat{y}_{j-1})$, this adaptive search domain is formed as

$$\Omega_j = [\hat{x}_{j-1} - 1/4, \hat{x}_{j-1} + 1/4] \times [\hat{y}_{j-1} - 1/4, \hat{y}_{j-1} + 1/4], \quad j \geq 1.$$

Clearly, we do this to make the computations as efficient as possible and avoid unnecessary searching. This idea and its results are presented in Figure 5, which shows the localized inversions at the time snapshots $T = 2.0$, $T = 4.0$, $T = 6.0$, $T = 8.0$, $T = 10.0$, and $T = 12.0$.

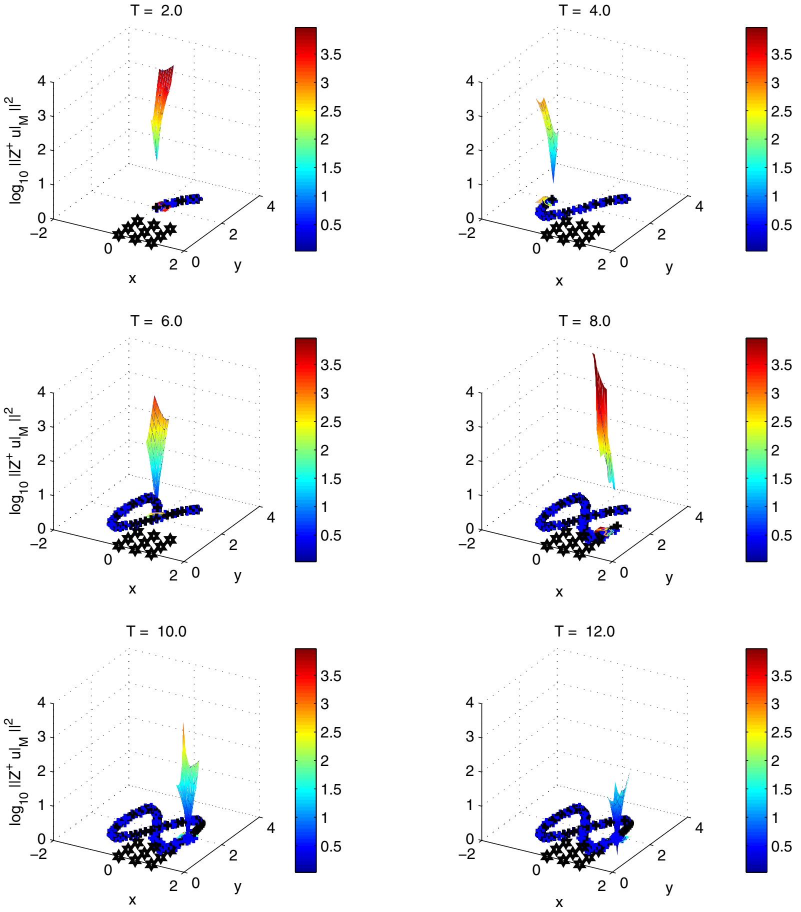


FIG. 5. *Localized numerical truncated Picard tests at the time snapshots* $T = 2.0$, $T = 4.0$, $T = 6.0$, $T = 8.0$, $T = 10.0$, *and* $T = 12.0$.

Finally, in Figure 6 we show the true track of the source, and its locations at the discrete snapshots in time, along with the estimated or reconstructed track and its instantaneous estimated positions. We remark that at time $T = 10.0$, when the source turns off, i.e., $s(10) = 0$, the carrier support's location remains constant, up to the noise and ill-posedness of the problem. That is, in the event of perfect data and a very well conditioned linear system, the source estimate would remain fixed at point $(1, 2)$.

5. Summary and conclusions. We have demonstrated that a simple knowledge of the instantaneous scalar field u on any bounded open set $\mathcal{M}$ located away from the (possibly time-varying) support of a source f is sufficient to estimate a nontrivial subset of the actual convex hull of the support of the source which we have called the carrier support of $u|_{\mathcal{M}}$. Additionally, we have provided and examined a viable numerical implementation of this result which can estimate, to essentially arbitrary precision, the trajectory of the carrier support over time, and hence track the moving source in real time, without a priori assumptions on the regularity of the source.

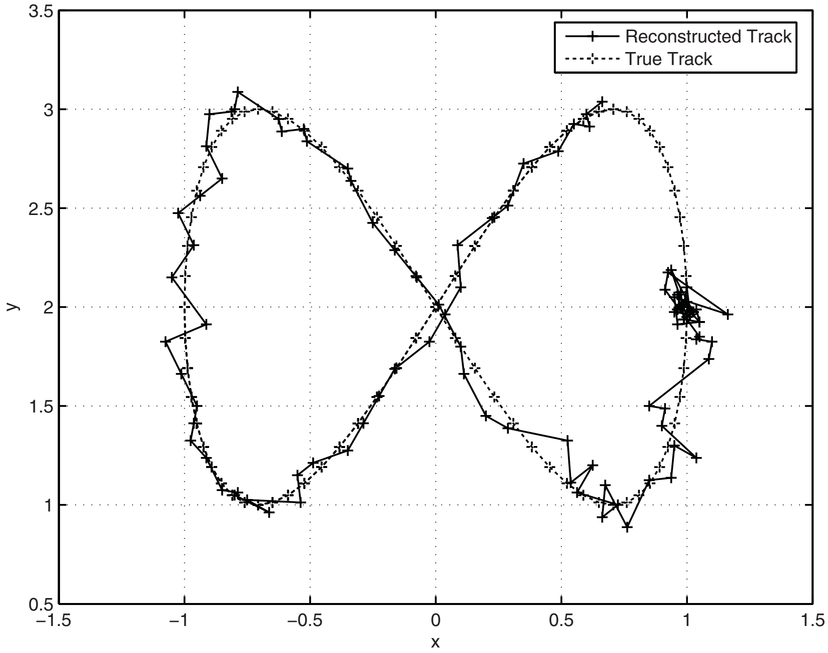


FIG. 6. *The discretely sampled true two-dimensional source track and the reconstructed (estimated) source track for the one-period sustained source.*

Moreover, we have shown that nonuniform sampling of the bounded and open measurement set $\mathcal{M}$ works as effectively as a uniform one. This result is important as it suggests that a wide collection of point samples distributed over a large domain of interest constitutes a robust methodology to locate and track sources of interest in a variety of applied problems, such as complex convective urban environments or large (aquatic) reservoir-like problems. In each of these, the robust and timely location of the effluent source is critical in nature, and may be accomplished with the few, sparse, and possibly nonuniformly sampled data assumed known in the analysis in this article.

In brief we mention that the concept of the carrier support does not provide us with a direct method which allows us to estimate the source strength as a function of time, and that this is certainly a significant problem of interest. In some simple cases, such as for constant coefficients of L_x, it may well be the case that by simply examining the local behavior of the objective function near the source, and observing it diminish in size over time, we may conclude that the source is no longer emitting into the system of interest and has become extinct. However, when the coefficients of L_x become more complex, such a simple observation may not persist. In such cases we propose an additional (forthcoming) technique which can estimate the strength of s over the time interval of measurements, $[0, T]$, which is based on a combination of the results provided here and some analysis of the Laplace transform of the governing equations of our main problem. Additionally, the current framework developed here accommodates the location and tracking of a source having but one component. The extension of the result to the problem of sources having multiple disjoint components is of much interest and is underway at the time of this writing.

Clearly, many future paths of research exist and warrant pursuit, foremost among them being a sensitivity analysis of how this method performs when the coefficients are known only to within some specified tolerance of their true values. Additional work consists of treating the problem in a stochastic setting, where these coefficients are not known instantaneously, as we have assumed throughout this article; rather they are known to possess certain distributional moments and belong to certain distributional families. Again, the possibilities for interesting and valuable future work are many indeed.

REFERENCES

[1] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1998.

[2] D. G. Duffy, *Green's Functions with Applications*, Stud. Adv. Math., Chapman & Hall/CRC, Boca Raton, FL, 2001.

[3] Yu. V. Egorov and M. A. Shubin, eds., *Partial Differential Equations* I, Encyclopaedia Math. Sci. 30, Springer-Verlag, Berlin, 1991.

[4] A. Elayyan and V. Isakov, *On an inverse diffusion problem*, SIAM J. Appl. Math., 57 (1997), pp. 1737–1748.

[5] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[6] F. Hettlich and W. Rundell, *Identifications of a discontinuous source in the heat equation*, Inverse Problems, 17 (2001), pp. 1465–1482.

[7] M. Ikehata, *An inverse source problem for the heat equation and the enclosure method*, Inverse Problems, 23 (2007), pp. 183–202.

[8] J. Kevorkian, *Partial Differential Equations. Analytical Solution Techniques*, 2nd ed., Springer-Verlag, New York, 2000.

[9] S. Krantz, *A Panorama of Harmonic Analysis*, Mathematical Association of America, Washington, DC, 1999.

[10] S. Kusiak and J. Sylvester, *The scattering support*, Comm. Pure Appl. Math., 56 (2003), pp. 1525–1548.

[11] S. Kusiak and J. Sylvester, *The convex scattering support in a background medium*, SIAM J. Math. Anal., 36 (2005), pp. 1142–1158.

[12] L. Ling, M. Yamamoto, Y. C. Hon, and T. Takeuchi, *Identification of source locations in two dimensional heat equations*, Inverse Problems, 22 (2006), pp. 1289–1305.

[13] A. Rap, L. Elliot, D. B. Ingham, D. Lesnic, and X. Wen, *The inverse source problem for the variable coefficients convection-diffusion equation*, Inverse Probl. Sci. Eng., 15 (2007), pp. 413–440.

[14] A. Rap, L. Elliot, D. B. Ingham, D. Lesnic, and X. Wen, *An inverse source problem for the convection-diffusion equation*, Internat. J. Numer. Methods Heat Fluid Flow, 16 (2006), pp. 125–150.

[15] M. Renardy and R. C. Rodgers, *An Introduction to Partial Differential Equations*, Texts Appl. Math. 13, Springer-Verlag, New York, 1992.

[16] J. Sylvester, *Notions of support for far fields*, Inverse Problems, 22 (2006), pp. 1273–1288.

[17] M. E. Taylor, *Partial Differential Equations* I: *Basic Theory*, Appl. Math. Sci. 116, Springer-Verlag, New York, 1996.

[18] W. Vettering, W. Press, S. Teukolsky, and B. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1992.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 806–824

DIFFEOMORPHIC SURFACE FLOWS: A NOVEL METHOD OF SURFACE EVOLUTION*

SIRONG ZHANG†, LAURENT YOUNES‡, JOHN ZWECK§, AND J. TILAK RATNANATHER¶

Abstract. We describe a new class of surface flows, diffeomorphic surface flows, induced by restricting diffeomorphic flows of the ambient Euclidean space to a surface. Different from classical surface PDE flows such as mean curvature flow, diffeomorphic surface flows are solutions of integro-differential equations in a group of diffeomorphisms. They have the potential advantage of being both topology-invariant and singularity free, which can be useful in computational anatomy and computer graphics. We first derive the Euler–Lagrange equation of the elastic energy for general diffeomorphic surface flows, which can be regarded as a smoothed version of the corresponding classical surface flows. Then we focus on diffeomorphic mean curvature flow. We prove the short-time existence and uniqueness of the flow, and study the long-time existence of the flow for surfaces of revolution. We present numerical experiments on synthetic and cortical surfaces from neuroimaging studies in schizophrenia and auditory disorders. Finally we discuss unresolved issues and potential applications.

Key words. elastic energy, diffeomorphisms, mean curvature flow, computational anatomy

AMS subject classifications. 53C44, 58D25

DOI. 10.1137/060664707

1. Introduction.

Surface evolution is both an important tool and an intriguing focus of mathematical research in geometric analysis, e.g., [5], and geometric PDEs, e.g., [37]. It also has been extensively applied in image processing, e.g., [3], and computer vision and interface modeling, e.g., [38]. In this paper, we develop and study a novel method of surface evolution under the action of the diffeomorphisms of the ambient Euclidean space.

We are interested in flows that can minimize surface area or mean curvature of a surface without inducing changes in topology or creating singularities. Therefore, it is natural to consider surface flows that are described by diffeomorphisms of the ambient Euclidean space. Motivated by the general framework of deformable template theory [17], we study transformations acting on objects, rather than the objects themselves. More specifically, we analyze flows on a group of diffeomorphisms of Euclidean space, rather than studying flows on the surfaces themselves. The foundations of this general framework have been rigorously established and have enabled comparisons to be made between deformable objects; see [11, 32] and the references therein. The theory has been successfully applied to image matching problems in which landmarks [24],

*Received by the editors July 11, 2006; accepted for publication (in revised form) September 7, 2007; published electronically January 9, 2008. This work was supported by NSF (DMS-0456253), NIH (R01-MH064838, P41-RR15241, R01-MH60883, P20-MH071616, P01-AG003991, P01-AG026276, R01-EB00975-01), and the Pacific Alzheimer Research Foundation.

http://www.siam.org/journals/siap/68-3/66470.html

†Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218-2686 (sirongzhang@jhu.edu).

‡Center for Imaging Science, Institute for Computational Medicine, and Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2686 (laurent.younes@jhu.edu).

§Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250 (zweck@math.umbc.edu).

¶Center for Imaging Science and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218-2686 (tilak@cis.jhu.edu).

curves [14], or surface patches [41] evolve under diffeomorphisms of Euclidean space. Here we use this framework in a variational setting, in an approach that is reminiscent of shape optimization methods [10].

In this paper, we transform variational formulations of classical surface flows (that minimize geometrical properties such as area, elastic energy, or total curvature) into optimization problems on a group of diffeomorphisms. This approach leads to solving the corresponding Euler–Lagrange equations as evolutions in the group that induce, via the group action, stable evolutions of surfaces. We call such flows *diffeomorphic surface flows*.

In the last decade, there have been studies on geometric flows such as the celebrated mean curvature flow, e.g., [12] and references therein, surface diffusion flow, e.g., [13], and the Willmore flow, e.g., [27, 26]. The equations for these surface flows are second or fourth order parabolic PDEs which require sophisticated numerical methods, e.g., [9]. Moreover, these flows can change surface topology and introduce singularities, e.g., [12, 29, 30].

We show that diffeomorphic surface flows can be regarded as smooth versions of the corresponding classical surface flows. They flow to minimize the energy while preserving surface topology and do not break down due to finite-time singularities since they are induced from the evolution of diffeomorphisms. Moreover, these flows are solutions of integro-differential equations on the diffeomorphism group, which are somewhat easier to discretize than the PDEs that govern the classical surface flows.

A major motivation for this work came from a desire to smooth triangulated cortical surfaces that are generated by marching cubes or tetrahedra isosurface algorithms [28, 18] based on a threshold derived from the segmentation of volumetric images of the brain, e.g., [35, 36]. In addition, topological defects generated by isosurface algorithms can be corrected by multiscale and morphological operations, e.g., [20, 21]. The end result is a triangulated surface that may contain several anomalous protrusions which may distort the true curvature of the surface and thus confound the interpretation of possible biological processes in disease such as neuronal migration of tissues, e.g., [1, 4]. Smoothing flows could be used to minimize these distortions. However, it is important that the topological properties of the surfaces be preserved to reflect the inherent biology at the scale of the voxel resolution of 0.5mm^3 or 1mm^3 and to not generate additional artifacts.

Algorithms for smoothing "noisy" surfaces have been the focus of intense efforts in the computer vision field, e.g., [23, 45]. Unfortunately little effort has been made to apply these sophisticated algorithms to cortical surfaces without losing accuracy and simplification. Among the earliest such algorithms, Joshi et al. [25] used the approach of Hamann [19] for generating local quadratic approximations to a discrete surface in order to locally smooth triangulated meshes and thus curvature. More sophisticated algorithms have since been developed. Among the most recent such methods are PDE algorithms based on the powerful level set method, e.g., [39, 43, 7], which, however, may lead to topological changes or singularities that may confound biological inference.

In this paper, we will present some encouraging preliminary results in which cortical surfaces are smoothed using diffeomorphic surface flows. In future work we will further develop the method and apply it in statistical analysis of cortical surfaces.

The organization of the paper is as follows. In section 2 we describe the mathematical background on flows of diffeomorphisms, classical surface energies, and the variational formulation for the surface flow. In section 2.3 we derive the Euler–Lagrange equations for diffeomorphic surface flows for a general elastic energy before focusing our attention on diffeomorphic mean curvature flow. In section 3 we prove the short-time existence of the solution of the diffeomorphic mean curvature flow equation and discuss the long-time existence for surfaces of revolution. In section 4 we describe a numerical implementation of the method, and in section 5 we present the results of our numerical simulations. Finally, unresolved issues and future directions are discussed in section 6.

2. Mathematical background.

2.1. Flows of diffeomorphisms. The set of diffeomorphisms, φ, of $\mathbb{R}^3$ forms a group under the operation of composition of mappings. Following the theory of flows of diffeomorphisms [11, 40], we introduce a Hilbert space V of smooth vector fields on $\mathbb{R}^3$, which is assumed to be continuously included in $\mathcal{X}_0^1(\mathbb{R}^3)$, the set of all C^1 vector fields that converge to zero (with their first derivatives) at infinity, equipped with the supremum norm. Any time-dependent vector field, $\nu_t : \mathbb{R} \to V$, generates a trajectory, φ_t, in the group of diffeomorphisms by

$$\frac{\partial \varphi_t}{\partial t} \circ \varphi_t^{-1} = \nu_t, \tag{2.1}$$

with initial condition $\varphi_0 = \mathrm{Id}$. We let $\mathcal{G}$ denote the group generated by all solutions φ_t of (2.1) with $\nu_s \in V$ for all $s \leq t$ and $\max_{s \leq t} \|\nu_s\|_V < \infty$. (The fact that this set forms a group is proved, for example, in [40].) We shall also use the notation $\mathcal{G}_V$ to make explicit the dependence of this group on the Hilbert space V.

We want to implement gradient descent algorithms in the group of diffeomorphisms, which is an issue often referred to as *shape optimization* [10]. A basic notion in this context is the one of *shape differential.* Given a scalar function F defined on $\mathcal{G}_V$ and an element $\varphi \in \mathcal{G}_V$, the shape differential of F at φ, denoted $\partial F(\varphi)$, is (if it exists) the linear form on V ($\partial F(\varphi) \in V^*$) defined by

$$\partial F(\varphi).v = \frac{d}{d\varepsilon} F((\mathrm{id} + \varepsilon v) \circ \varphi).$$

Assume that, for each φ, a dot product $\langle\, . \,\rangle_\varphi$ is defined on V^*. The gradient of F at φ with respect to this dot product, denoted $\nabla F(\varphi)$, is then defined by the following identity: for all $m \in V^*$,

$$\langle \partial F(\varphi), m \rangle_\varphi = m.\nabla F(\varphi).$$

The associated gradient descent algorithm is the flow defined by

$$\frac{d\varphi}{dt} = -\nabla F(\varphi) \circ \varphi. \tag{2.2}$$

By the chain rule, we can write

$$\frac{d}{dt} F(\varphi) = \frac{d}{d\varepsilon} F((\mathrm{id} - \varepsilon \nabla F(\varphi)) \circ \varphi) = -\|\partial F(\varphi)\|_\varphi^2, \tag{2.3}$$

which shows that the algorithm does indeed minimize F. If $\nabla F(\varphi)$ is a smooth vector field over a time interval $[0, T]$, then φ in (2.2) is the flow associated to an ODE and therefore a diffeomorphism.

We denote by K the duality operator between V^* and V, defined by $m.v = \langle Km, v \rangle_V$ for $m \in V^*$ and $v \in V$. The assumption that V is continuously included in

$\mathcal{X}_0^1(\mathbb{R}^3)$ implies that K is a kernel operator, making V a reproducing kernel Hilbert space [2, 42]. Indeed, for $a \in \mathbb{R}^3$, the linear form $m = a \otimes \delta_x$ defined by $m.v = a^T v(x)$ is continuous on V, so that $K(a \otimes x) \in V$ is well defined and obviously linear in a. This defines a mapping (also denoted K) from $\mathbb{R}^3 \times \mathbb{R}^3$ to $GL_3(\mathbb{R})$ by

$$K(x, y)a = K(a \otimes \delta_x)(y). \tag{2.4}$$

A key point here is that V can be specified by the definition of its kernel. In our case, K will be chosen as a scalar Gaussian kernel (or more precisely by a Gaussian kernel multiplied by the identity matrix). The corresponding Hilbert space (at scale σ) is defined by

$$V_\sigma = \left\{ v = K^{1/2}u = \int_{\mathbb{R}^3} e^{-\frac{2\|x-y\|^2}{\sigma^2}} u(y)dy,\ u \in L^2(\mathbb{R}^3) \right\}, \tag{2.5}$$

where the inner product on V_σ is defined by $\langle K^{1/2}u, K^{1/2}u' \rangle_V = \langle u, u' \rangle_{L^2}$. The associated kernel is $K = (K^{1/2})^2$, which is proportional to $\exp(-\|x - y\|^2/(2\sigma^2))$.

The dual dot product on V^* comes straightforwardly from the fact that K is a duality operator, yielding

$$\langle m, \tilde{m} \rangle_{V^*} = m.(K\tilde{m}).$$

The φ-dependent dot product $\langle\, .\, ,\, .\, \rangle_\varphi$ used in this paper will be weighted versions of this dual product, taking the form

$$\langle m, \tilde{m} \rangle_\varphi = \langle \rho_\varphi m, \rho_\varphi \tilde{m} \rangle_{V^*} = m.(\rho_\varphi K(\rho_\varphi \tilde{m})),$$

where ρ_φ is a nonnegative scalar function and $(\rho_\varphi m).v := m(\rho_\varphi v)$. The associated gradient descent algorithm becomes

$$\frac{d\varphi}{dt} = -(\rho_\varphi K(\rho_\varphi \partial F(\varphi))) \circ \varphi. \tag{2.6}$$

We now describe how this is implemented, with a suitable choice for ρ_φ, for surface evolution.

2.2. Surface energy.

We consider the general surface energy functional [34, 22]

$$E(\Sigma) = \int_\Sigma (\alpha + \beta H^2 - \gamma G)d\sigma, \quad \text{where} \quad \alpha \geq 0,\ \beta \geq \gamma \geq 0, \tag{2.7}$$

where H and G are the mean and Gauss curvature of Σ, respectively. This elastic energy functional is a linear combination of three basic energy functionals:

- area: $U(\Sigma) = \int_\Sigma d\sigma$,
- Willmore energy: $U(\Sigma) = \int_\Sigma 4H^2 d\sigma$, and
- total curvature: $U(\Sigma) = \int_\Sigma (4H^2 - 2G)d\sigma = \int_\Sigma (k_1^2 + k_2^2)d\sigma$, where k_1, k_2 are principal curvatures.

These energy functionals can be locally minimized using the classical surface flows known as mean curvature (area-minimizing), Willmore, and total curvature flow, respectively.

We generate diffeomorphic surface flows as follows. If Σ_0 is the initial surface, we can define $F(\varphi) = E(\varphi(\Sigma_0))$. We then let $\Sigma_t = \varphi_t(\Sigma_0)$, where φ_t is given by (2.6).

2.3. Euler–Lagrange equation. In this section, we derive the gradient flow equation for the diffeomorphic surface flow minimizing the elastic energy. From section 2.2, we first compute the variation of the energy. The following lemma is due to Nitsche [34]. For simplicity, we suppose that all surfaces are *oriented and closed.* Let the surface be $\Sigma(p)$ and its variation be $\Sigma_\varepsilon(p) = \Sigma(p) + \varepsilon\nu(p)$.

LEMMA 2.1 (variation of elastic energy). *For surface elastic energy, $E(\Sigma) = \int_\Sigma (\alpha + \beta H^2 - \gamma G) d\sigma$, the energy variation is*

$$\frac{\partial}{\partial\varepsilon} E(\Sigma_\varepsilon)_{|\varepsilon=0} = \int_\Sigma \alpha\nu^\perp H - \beta\nu^\perp(\Delta_\Sigma H + 2H(H^2 - G))d\sigma, \tag{2.8}$$

where $\nu^\perp = \langle \nu, N\rangle$ is the normal component of the vector field ν, Δ_Σ is the intrinsic Laplace operator, and N is the surface normal.

A proof may be found in Willmore [44, pp. 279–282].

Remark 1. For a closed surface, the term with γ is absent since by the Gauss–Bonnet theorem, the integral of the Gauss curvature is a constant.

This lemma directly provides the expression of the shape derivative of F at φ, since, for $\Sigma = \varphi(\Sigma_0)$, $((\mathrm{id} + \varepsilon v) \circ \varphi)(\Sigma_0) = \Sigma + \varepsilon v(\Sigma)$, yielding

$$\partial F(\varphi).\nu = \int_\Sigma \alpha\nu^\perp H - \beta\nu^\perp(\Delta_\Sigma H + 2H(H^2 - G))d\sigma.$$

Now, from (2.6), we obtain the diffeomorphic evolution equations, in which we assume that ρ_φ depends on φ only via the deformed surface Σ, hence employing the notation $\rho_\varphi = \rho_\Sigma$ and $\Sigma_t = \varphi_t \circ \Sigma_0$:

$$\begin{aligned} \frac{\partial\varphi_t(y)}{\partial t} &= -\rho_{\Sigma_t}(\varphi_t(y)) \\ &\quad \cdot \int_{q\in\Sigma_t} \big(\alpha H - \beta(\Delta_{\Sigma_t} H + 2H(H^2 - G))\big) K(\varphi_t(y), q)\rho_{\Sigma_t}(q) N(q) d\sigma_t(q). \end{aligned} \tag{2.9}$$

We define ρ_Σ as an area normalization factor as follows. Define, for a surface Σ, the local area function

$$a_\Sigma(p) = \int_{q\in\Sigma} K(p, q) d\sigma. \tag{2.10}$$

We then set

$$\rho_\Sigma(p) = a_\Sigma(p)^{-1/2}. \tag{2.11}$$

Choosing this normalization ensures that the right-hand sides in the diffeomorphic flows have the same dimensions as the corresponding classical flows (e.g., 1/length for the mean curvature flow). Doing so, the large-scale behavior (relative to the width of the kernel) is expected to be similar for both flows.

Although all quantities introduced so far are defined on the whole space, we are primarily interested in the evolution of the surface $\Sigma_t = \varphi_t \circ \Sigma_0$. Hence, for the area and Willmore energy functionals, the equations that govern the flow of each point $p = \varphi_t(y)$ on the closed surface Σ_t are given by the following integro-differential equations:

- diffeomorphic mean curvature flow ($\alpha = 1$ and $\beta = 0$):

$$\frac{\partial p}{\partial t} = -a_{\Sigma_t}^{-1/2}(p) \int_{q\in\Sigma_t} K(p, q) H(q) a_{\Sigma_t}^{-1/2}(q) N(q) d\sigma_t, \tag{2.12}$$

- diffeomorphic Willmore flow ($\alpha = 0$ and $\beta = 1$):

$$(2.13)\quad \frac{\partial p}{\partial t} = a_{\Sigma_t}^{-1/2}(p) \cdot \int_{q\in\Sigma_t} K(p,q)(\Delta_{\Sigma_t} H(q) + 2H(q)(H^2(q) - G(q)))a_{\Sigma_t}^{-1/2}(q)N(q)d\sigma_t.$$

We will use these formulae in the numerical implementation in section 4. Notice that they are similar to the corresponding equations for the classical mean curvature and Willmore flows. The diffeomorphic surface flows have the same energy minimizing property as their classical counterparts, but since they are diffeomorphisms, they preserve the topology of the surface. In the next section, we focus only on diffeomorphic mean curvature flow.

3. Diffeomorphic mean curvature flow. From now on, we use the Gaussian kernel function

$$(3.1)\qquad K(p,q) = e^{-\frac{\|p-q\|^2}{2\sigma^2}}.$$

Here σ is the kernel size, which corresponds (up to a change in the normalization factor) to the reproducing kernel of V_σ given by (2.5).

The flow equation is therefore given by

$$(3.2)\qquad \frac{\partial \varphi_t(y)}{\partial t} = -a_{\Sigma_t}(\varphi_t(y))^{-1/2} \int_{q\in\Sigma_t} e^{-\frac{\|\varphi_t(y)-q\|^2}{2\sigma^2}} H(q)a_{\Sigma_t}(q)^{-1/2}N(q)d\sigma_t,$$

with the initial condition $\varphi_0 = \mathrm{id}$. As indicated by (2.3), this is an area-minimizing flow for the surface Σ_t, with the explicit formula

$$\frac{d|\Sigma_t|}{dt} = -\int_{\Sigma_t^2} e^{-\frac{\|p-q\|^2}{2\sigma^2}} \left(\frac{H(p)N(p)}{\sqrt{a_{\Sigma_t}(p)}}\right)^T \left(\frac{H(q)N(q)}{\sqrt{a_{\Sigma_t}(q)}}\right) d\sigma_t(p)d\sigma_t(q).$$

3.1. Short-time existence of solution. The classical flows are local flows and by PDE theory, there are short-time solutions for smooth initial data. Because of the integro-differential form of diffeomorphic surface flows, short-time existence follows from standard ODE arguments on Banach spaces.

Theorem 1 (short-time existence and uniqueness). *For any initial compact smooth surface, there exists a unique solution for the flow equation* (3.2) *in a small time interval* $[0, t_0]$.

Proof. Consider the space $A = \mathbb{R}^3 \times \mathrm{GL}_3(\mathbb{R}) \times \mathrm{Bil}(\mathbb{R}^3, \mathbb{R}^3)$, where the last factor is the set of bilinear functions from $\mathbb{R}^3 \times \mathbb{R}^3$ to $\mathbb{R}^3$. A generic element of A will be denoted $Q = (\varepsilon, F, S)$, and we will consider the Banach space B of continuous functions $Q(\cdot) : \mathbb{R}^3 \to A$, with the supremum norm

$$(3.3)\qquad \|Q\| = \|\varepsilon\|_\infty + \|F\|_\infty + \|S\|_\infty.$$

Here $\varepsilon(y)$ is a C^2 vector field that converges to zero (with its first and second derivatives) at infinity, and $F(y)$ and $S(y)$ are first and second derivatives of $\varepsilon(y)$.

Letting $\varphi(y) = y + \varepsilon(y)$, we first embed (3.2) in an ODE on B. We rewrite the right-hand side of (3.2) using integrals over Σ_0. Covering Σ_0 with local charts $f : U \to \Sigma_0$, we have that $\Sigma_t(y) = \varphi_t \circ f(y)$. Suppose that $\{f_u, f_v, N_0\}$ is an orthonormal basis at the point y. Then, using $N(t,\varphi(y)) = d\varphi^{-T}N_0/\|d\varphi^{-T}N_0\|$ [6] and the expression for the mean curvature in coordinates (omitting the subscript t), we have

$$\begin{aligned}H(t,\varphi) = \frac{1}{\|d\varphi f_u \times d\varphi f_v\|^2\,\|d\varphi^{-T}N_0\|^2}\Big(&\langle d\varphi^{-1}d^2\varphi(f_u,f_u)+f_{uu},N_0\rangle\|d\varphi f_v\|^2\\ &+\langle d\varphi^{-1}d^2\varphi(f_v,f_v)+f_{vv},N_0\rangle\|d\varphi f_u\|^2\\ &-2\langle d\varphi^{-1}d^2\varphi(f_u,f_v)+f_{uv},N_0\rangle\langle d\varphi f_u, d\varphi f_v\rangle\Big).\end{aligned}$$

Defining the quadratic forms on $T\Sigma = d\varphi T\Sigma_0$,

$$A_\varphi(g,h) = \langle d\varphi^{-1}d^2\varphi(d\varphi^{-1}g, d\varphi^{-1}h), N_0\rangle \tag{3.4}$$

and

$$II_\varphi(g,h) = -\langle d\varphi^{-1}g, dN_0 d\varphi^{-1}h\rangle, \tag{3.5}$$

and using the fact that $\|d\varphi f_u \times d\varphi f_v\| = |\det D\varphi|\,\|d\varphi^{-T}N_0\|$, we can rewrite

$$H(t,\varphi) = \frac{\operatorname{trace}(A_\varphi)+\operatorname{trace}(II_\varphi)}{|\det D\varphi|^2\,\|d\varphi^{-T}N_0\|^4}. \tag{3.6}$$

So, the evolution equation can be written as

$$\begin{aligned}\frac{\partial\varphi_t(y)}{\partial t} &= -a_\Sigma(\varphi_t(y))^{-1/2}\\ &\cdot\int_{\Sigma_0} e^{-\frac{\|\varphi_t(x)-\varphi_t(y)\|^2}{2\sigma^2}}\frac{\operatorname{trace}(A_{\varphi_t})+\operatorname{trace}(II_{\varphi_t})}{|\det D\varphi_t|\,\|d\varphi_t^{-T}N_0\|^4}a_\Sigma(\varphi_t(x))^{-1/2}d\varphi_t^{-T}N_0(x)d\sigma_0,\end{aligned} \tag{3.7}$$

with

$$a_\Sigma(\varphi(y)) = \int_{\Sigma_0} e^{-\frac{\|\varphi(x)-\varphi(y)\|^2}{2\sigma^2}}|\det D\varphi|\,\|d\varphi^{-T}N_0\|d\sigma_0(x).$$

Introducing $d\varphi = \mathrm{Id} + d\varepsilon = \mathrm{Id} + F$ and $d^2\varphi = d^2\varepsilon = S$, (3.7) takes the form

$$\frac{d\varepsilon}{dt} = J_0(\varepsilon, F, S). \tag{3.8}$$

The time evolution of F is obtained by computing the differential of this equation with respect to the space variable. Since in (3.7), the variable y appears only in evaluations of φ (not its derivatives), it is clear that the evolution of F will also take the form $dF/dt = J_1(\varepsilon, F, S)$. Since the same argument can be made for S, we obtain the fact that (ε, F, S) follows an ODE in B of the form

$$\frac{d\varepsilon}{dt} = J_0(\varepsilon,F,S),\quad \frac{dF}{dt} = J_1(\varepsilon,F,S),\quad \frac{dS}{dt} = J_2(\varepsilon,F,S). \tag{3.9}$$

It is clear now that J_0, J_1, J_2 are integrals of rational functions of ε, F, S which are well defined in a neighborhood of $(\varepsilon, F, S) = (0,0,0)$. This ensures short-term existence of solutions of the system in B.

It remains to show that, if (ε, F, S) is a solution of this system, the first component, ε, is in fact a solution of (3.8) with initial condition $\varepsilon_0 = 0$. For this it suffices to

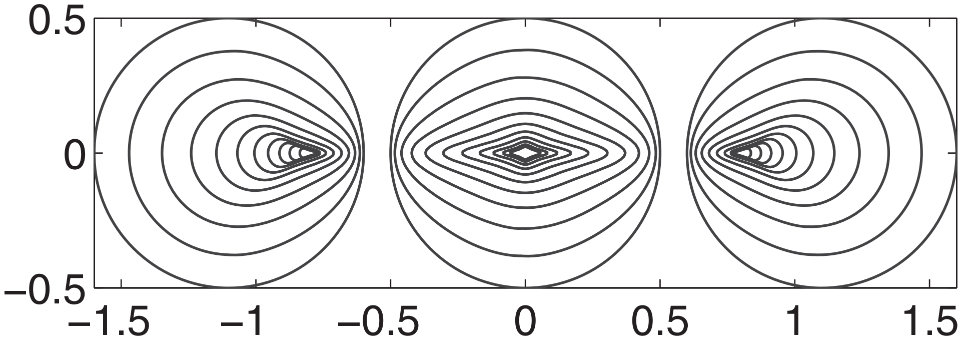


FIG. 1. *Interaction of three nearby spheres during the diffeomorphic mean curvature flow. The initial radius of all spheres is* 0.5*, and the minimum distance among them is* 0.1*. The kernel size is* $\sigma = 0.3$*, and the stopping time is* $T = 1$*.*

prove that ε is twice differentiable, with its first and second differentials given by F and S. This follows from standard arguments for ODEs, and we omit the details here. Finally, it is easy to see that we have a unique solution $\varphi(y) = y + \varepsilon(y)$ for the original flow equation (3.2). □

Remark 2. The result in Theorem 1 also holds for the more general diffeomorphic surface evolution

$$\frac{\partial \varphi_t(y)}{\partial t} = a_{\Sigma_t}^{-1/2}(\varphi_t(y)) \cdot \int_{q\in\Sigma_t} K(\varphi_t(y), q) f(q, N(q), dN(q), d^2N(q), \dots, d^mN(q)) a_{\Sigma_t}^{-1/2}(q) d\sigma_t,$$

where f has continuous derivatives for each variable, Σ_0 is a smooth surface, m is an integer, and N is the surface normal. The proof follows along the same general lines as the proof we gave above. We rewrite the integral as an integral on the original surface and the function f as a function $G(\varphi, d\varphi, d^2\varphi, \dots, d^{m+1}\varphi)$. Then the equations of the derivatives of $d\varphi, d^2\varphi, \dots, d^{m+1}\varphi$ involve only the derivatives of the kernel and G.

We can show that the solution φ is in fact a diffeomorphism by standard arguments using Gronwall's lemma [46, Chap. 10].

THEOREM 2 (diffeomorphism). *If* φ_t *is the solution of the flow equation* (3.2) *on the interval* $[0, T]$*, then* φ_t *is a diffeomorphism for all* $t \in [0, T]$*.*

We have the following important consequence.

COROLLARY 1 (topology invariance and singularity free). *The solution of the flow equation* (3.2) *at each time t gives a smooth surface with the same topology as the initial surface as long as the solution exists.*

We are obviously interested in the long-time existence of the flow. Numerical evidence and analysis for simple surfaces so far suggest that the flow has a long-time solution. However, the interactions between remote parts of a surface make the long-time behavior difficult to analyze (see, for example, Figure 1). Analyzing the proof of Theorem 1, we see that being able to extend a solution beyond a given time $t < t_0$ depends only on the regularity of the surface at time t. More precisely, a standard lower-bound of how far a current solution of an ODE in a Banach space can be extended beyond t is directly related to the Lipschitz constant of the ODE. In our proof, the Lipschitz properties of the system rely on the surface only via upper-bounds on the second derivative of the normal in (3.6) and via lower-bounds on the local area a_Σ. So solutions of (3.9) can be extended in time as long as the surface does not develop singularities.

A consequence of this analysis is that, if we can exhibit an evolution starting at some Σ_0 on some interval $[0, T]$ on which the total curvature of the evolving surface remains bounded from above and the local area remains bounded from below, then this evolution is the only solution of (3.2) starting at Σ_0.[1] This property will be used in the next section when we provide simplified equations for surfaces of revolution.

3.2. Surfaces of revolution. In this subsection we analyze the diffeomorphic mean curvature flow for a surface of revolution, Σ. Because of the uniqueness of solutions, the solution of (3.2) must preserve rotational symmetry. Consequently, it suffices to evolve the profile curve of Σ. Therefore, here we derive an equation for the evolution of the profile curve for the diffeomorphic mean curvature flow equation (3.2). It is much more computationally efficient to solve the equation for the profile curve than it is to solve the full flow equation (3.2) for a triangulated surface.

We parameterize a surface of revolution by

$$\mathbf{x}(u, v) = (\alpha(u)\cos v, \alpha(u)\sin v, \beta(u)). \tag{3.10}$$

Here $u \in [-1, 1]$, $v \in [0, 2\pi]$, and the profile curve $\gamma(u) = (\alpha(u), \beta(u))$ satisfies suitable conditions. For a closed curve, we need $\alpha > 0$, while for an open curve, we require that $\alpha \geq 0$ with $\alpha = 0$ only at the end points, and also $\beta' = 0$ at the end points. Here and below α', β' denote the derivatives of the functions α, β. The orientation of the curve is taken to be counterclockwise.

We now derive the induced flow equation for the profile curve. First, we have

$$N = \frac{(-\beta'\cos v, -\beta'\sin v, \alpha')}{\sqrt{(\alpha')^2 + (\beta')^2}}, \tag{3.11}$$

$$d\sigma = \alpha\sqrt{(\alpha')^2 + (\beta')^2}dudv, \qquad 2H = \frac{\beta'}{\alpha\sqrt{(\alpha')^2 + (\beta')^2}} + \kappa, \tag{3.12}$$

where κ is the curvature of the profile curve and the normal vector N is outward pointing.

Using these expressions, we can characterize the evolution of the profile curve $\mathbf{x}(u, 0) = (\alpha(u), 0, \beta(u))$ as follows:

$$\begin{aligned}
&\frac{\partial\alpha(u)}{\partial t} = -a(u)^{-1/2}\\
&\cdot\int_{-1}^{1} e^{-\frac{1}{2\sigma^2}((\alpha(u)-\alpha(u_t))^2+(\beta(u)-\beta(u_t))^2)} g_1(\alpha(u)\alpha(u_t))H(u_t)\alpha(u_t)\beta'(u_t)a(u_t)^{-1/2}du_t,\\
&\frac{\partial\beta(u)}{\partial t} = a(u)^{-1/2}\\
&\cdot\int_{-1}^{1} e^{-\frac{1}{2\sigma^2}((\alpha(u)-\alpha(u_t))^2+(\beta(u)-\beta(u_t))^2)} g_0(\alpha(u)\alpha(u_t))H(u_t)\alpha(u_t)\alpha'(u_t)a(u_t)^{-1/2}du_t,\\
&a(u) = \int_{-1}^{1} e^{-\frac{1}{2\sigma^2}((\alpha(u)-\alpha(u_t))^2+(\beta(u)-\beta(u_t))^2)} g_0(\alpha(u)\alpha(u_t))\alpha(u_t)\sqrt{\alpha'(u_t)^2 + \beta'(u_t)^2}du_t,
\end{aligned} \tag{3.13}$$

[1] Total curvature plays a role in the analysis of the classical mean curvature flow [12, Thm. 3.4].

where

$$g_1(x) = \int_0^{2\pi} e^{x(\cos(v)-1)/\sigma^2}\cos(v)dv, \qquad g_0(x) = \int_0^{2\pi} e^{x(\cos(v)-1)/\sigma^2}dv.$$

For a discrete profile curve consisting of finite line segments, (3.13) is a system of ODEs which has a short-time solution and can be solved numerically using MATLAB. In section 5 we use this system of ODEs to study the long-time behavior of the solution for surfaces of revolution.

For the classical mean curvature flow of a surface of revolution the only possible singularities are on the axis of revolution [12]. We conjecture that for a closed profile curve if the curvature is bounded, then $\alpha > 0$ for all time. In fact, numerical results suggest that an even stronger result is true.

Conjecture 1 (long-time solution for surfaces of revolution). *There exists a unique solution for the flow equation* (3.13) *for all* $t \geq 0$.

3.3. Sphere evolution. When the surface is the sphere, we have an explicit solution for the diffeomorphic mean curvature flow equation.

Proposition 1 (sphere evolution). *If the initial surface* Σ_0 *is a sphere of radius* R_0*, then the solution of the diffeomorphic mean curvature flow* (3.2) *exists for all time, and at each time* t *the surface* Σ_t *is a sphere of radius* R_t*, where* R_t *satisfies the equation*

$$\frac{dR_t}{dt} = -\frac{(R_t^2+\sigma^2)e^{-2R_t^2/\sigma^2} + R_t^2 - \sigma^2}{R_t^3(1-e^{-2R_t^2/\sigma^2})}. \tag{3.14}$$

Proof. By symmetry and from the uniqueness of solutions, the evolving surface remains a sphere at all times. Equation (3.14) is a direct application of the general formulae with $\alpha(u_t) = R_t\cos(\pi u_t/2)$, $\beta(u_t) = R_t\sin(\pi u_t/2)$ at $u_0 = 1$. For example,

$$\begin{aligned} a_{\Sigma_t}(1) &= 2\pi\int_{-1}^{1} e^{-\frac{R_t^2}{\sigma^2}(1-\sin\frac{\pi u_t}{2})} R_t^2\frac{\pi}{2}\cos\frac{\pi u_t}{2}du_t \\ &= 2\pi R_t^2 e^{-\frac{R_t^2}{\sigma^2}}\int_{-1}^{1} e^{\frac{R_t^2 z}{\sigma^2}}dz \\ &= 2\pi\sigma^2(1-e^{-\frac{2R_t^2}{\sigma^2}}). \end{aligned}$$

Similar computations can be done for the other integral, leading to (3.14). One can check that the function $f(r)$,

$$r \mapsto -\frac{(r^2+\sigma^2)e^{-2r^2/\sigma^2} + r^2 - \sigma^2}{r^3(1-e^{-2r^2/\sigma^2})},$$

is well defined and differentiable over $\mathbb{R}$ (including 0), vanishes at 0, and is negative for $r > 0$. This implies that solutions starting at $R_0 > 0$ decrease without reaching 0, and can be extended to infinite time. □

A Taylor expansion of the equation at $R_t = 0$ yields $dR_t/dt \simeq -(2/3\sigma^2)R_t$ at $t = 0$, yielding an exponential decay of the radius.

By a similar argument, we can show that the following proposition holds.

Proposition 2 (cylinder evolution). *The evolution of a circular (infinite) cylinder exists and is unique for all* $t \geq 0$.

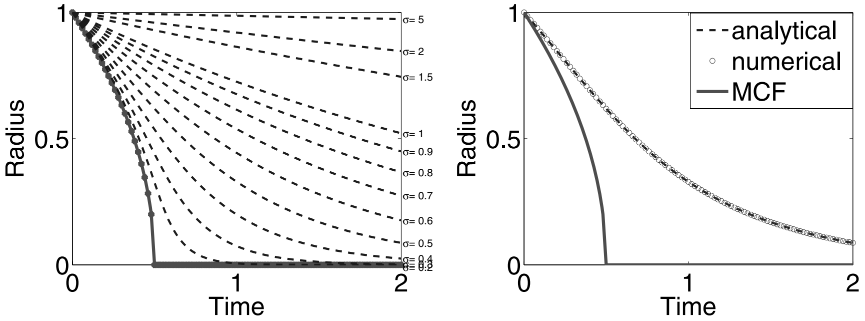


Fig. 2. *Diffeomorphic mean curvature flow of a sphere. Left: Flows with different kernel sizes, σ (dotted curves). The classical mean curvature flow is indicated by the solid curve with stars. Right: Validation of the numerical algorithm in section* 4 *(circles) by comparison with the analytical solution from Theorem* 1 *(dashed curves) with $\sigma = 0.5$. The classical mean curvature flow is indicated by the solid curve.*

We leave the proof to the reader. In this case, the first order expansion for the evolution of the radius is $dR_t/dt \simeq -(1/2\sigma^2)R_t$, yielding here also an exponential decay, at a rate slower than for the sphere.

Recall that for the classical mean curvature flow, a sphere vanishes into a point and the circular cylinder into the y axis at finite time [12]. However, the diffeomorphic mean curvature flow preserves surface topology for all times.

In the left panel of Figure 2 we compare the classical mean curvature flow for a sphere to the diffeomorphic mean curvature flow for a range of kernel sizes, σ. Notice that for the mean curvature flow the surface collapses to a point at time $t = 0.5$, whereas for all kernel sizes the solution of the diffeomorphic mean curvature flow exists for all times. Moreover, from the evolution equation (3.14), it is not hard to prove that for the unit sphere, as the kernel size converges to zero, the solution of the diffeomorphic mean curvature flow converges to that of the mean curvature flow over the interval $0 \leq t \leq 0.5$. This suggests the following conjecture.

Conjecture 2 (mean curvature flow limit). *Suppose, for an initial compact smooth surface, that the mean curvature flow exists for all times $t \in [0, T]$. Then for any time $T_1 < T$, on the time interval $[0, T_1]$ the diffeomorphic mean curvature flow converges uniformly to the mean curvature flow as the kernel size goes to zero.*

4. Numerical implementation.

In this section we give the details of the numerical implementation of diffeomorphic surface flows via the Runge–Kutta method. We focus on two aspects: the estimation of geometric quantities such as the normal and curvature on discrete surfaces, and the ODE solver for diffeomorphic surface flows.

4.1. Discrete differential geometry of surfaces.

Surfaces are discretized as triangulated meshes. Consequently, to numerically solve the diffeomorphic surface flow equations, we need to define discrete geometry quantities that approximate the normal and curvature functions on a smooth surface. Several such discretization methods have been described in the literature but none is universally used [19, 15, 31, 8]. We will use the discrete differential operators method of Meyer et al. [31], which is easy to implement and suitable for our examples.

Given a triangulated mesh with vertices v_i and faces f_j, the vertex one-ring $R_1(i)$ of a vertex v_i is the set of all adjacent vertices, and the face one-ring $F_1(i)$ of v_i is the set of all faces containing v_i. We determine the geometry at each vertex from the vertex one-ring of that vertex. If none of the triangles in $F_1(i)$ is obtuse, we define the area $A(v_i)$ at the vertex v_i to be the Voronoi area of that vertex:

$$A_{\text{Voronoi}} = \frac{1}{8} \sum_{j \in R_1(i)} (\cot \alpha_{ij} + \cot \beta_{ij}) \|v_i - v_j\|, \tag{4.1}$$

where α_{ij} and β_{ij} are the two angles opposite the edge $v_i v_j$ in the two triangles sharing that edge. However, if some of the triangles in $F_1(i)$ are obtuse, we define the area $A(v_i)$ to be the mixed area A_{mixed} described in [31].

Then the mean curvature normal vector $\overline{HN}$ at v_i is given by

$$\overline{HN}(v_i) = \frac{1}{4A(v_i)} \sum_{j \in R_1(i)} (\cot \alpha_{ij} + \cot \beta_{ij})(v_i - v_j). \tag{4.2}$$

We can also obtain the normal, mean curvature, and Gauss curvature formulae [31], which are not used in this paper.

4.2. ODE solution. We evolve a triangulated surface via its vertices; i.e., the surface vertices are used to discretize the flow equation (3.2). The resulting ODE system is solved numerically using a Runge–Kutta method.

The algorithm is as follows.

Algorithm 1. ODE solver for diffeomorphic surface flow.

```
1: Initialize the flow time $T$ and kernel size $\sigma$
2: Initialize surface with $v_i$ and $F_j$
3: Generate the one-ring neighborhood structure $R_1(i)$ and $F_1(i)$ for each vertex
4: while $t < T$ do
5:    Compute the geometry of the surface $\overline{HN}_i = \overline{HN}(v_i)$ and $A_i = A_{\text{mixed}}(v_i)$
6:    Compute the Gaussian kernel weights $K_{ij} = K(v_i, v_j)$
7:    Compute the local area weights $\text{loc}_i = (\sum_{j=\text{all}} K_{ij} A_j)^{-0.5}$
8:    Compute the flow speed term $u_i = -\text{loc}_i \sum_{j=\text{all}} K_{ij}\, \text{loc}_j\, \overline{HN}_j A_j$
9:    Obtain the new vertices from a Runge–Kutta solver $v_i = RK(v_i, u_i)$
10: end while
11: Output the surface
```

Remark 3. We did not attempt to determine an automatic stopping condition. We simply stopped after time T.

Remark 4. It is possible to use implicit ODE solvers.

Remark 5. In step 8, for each i it is possible to sum only over those indices j for which the Gaussian kernel K_{ij} exceeds a small threshold. Alternatively, for large kernel size one could use the fast Gauss transform [16].

Remark 6. For mean curvature flows of surfaces of revolution, we used a public-domain MATLAB toolbox for level set methods [33]. For mean curvature flows of triangulated meshes we used Algorithm 1 but with $u_i = -\overline{HN}_i$.

5. Results. We used MATLAB to implement the algorithm on a Pentium IV 3.2 GHz machine with 2 GB of RAM. In general for a synthetic surface with 7200 faces and 3600 vertices, one loop takes about 10 seconds. However, for cortical surface applications we used a C++ implementation that takes about 1 second per loop for a surface with 5000 vertices.

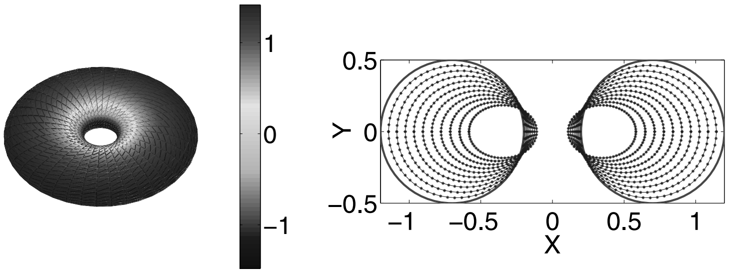


Fig. 3. *Flow of a fat torus. Left: The fat torus with mean curvature indicated by the grayscale. Right: Comparison between the profile curves obtained using the ODE solution of* (3.13) *for surfaces of revolution and those obtained using the numerical algorithm in section* 4.2*, respectively, indicated by dashed curves and plus symbols. The initial profile curve is shown with the two largest solid circles. Here $T = 0.5$, and $\sigma = 0.3$.*

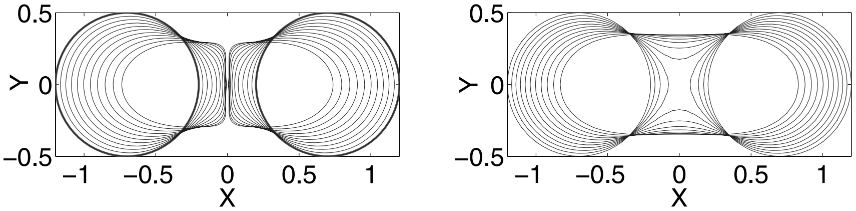


Fig. 4. *Comparison of flows for the fat torus. Left: Diffeomorphic mean curvature flow with $\sigma = 0.1$ and $T = 0.25$ obtained using the ODE algorithm for surfaces of revolution. Right: Classical mean curvature flow with $T = 0.2$.*

Sphere. In the right panel of Figure 2 we show the diffeomorphic mean curvature flow of a sphere for a kernel size of $\sigma = 0.3$ computed using Algorithm 1 (circles) and using the ODE (3.14) for the radius of the sphere (dashed curve). The agreement between the two methods is excellent. For comparison, we show the result for the classical mean curvature flow with a solid curve. For the algorithm, we generated an initial sphere with 642 vertices and 1280 faces using recursive subdivision of a cube. The solution of the ODE (3.14) was obtained using the MATLAB function ode45. As shown in Figure 1, three disconnected spheres can influence one another. It would be very interesting but difficult to analyze the long-time behavior of such interactions.

Circular torus. The circular torus was generated by rotating a circle about the y axis. The triangulated mesh had 3600 vertices and 7200 faces. We chose a "fat" torus with inner radius 0.2 and outer radius 1.2, as shown in Figure 3. In the right panel of Figure 3 we examined the solutions of the diffeomorphic mean curvature flow, obtained both by solving the surface of revolution flow equation (3.13) using the MATLAB function ode45 and by flowing the triangulated surface using Algorithm 1 indicated by dashed curves and plus symbols. The initial surface is represented by the two largest circles. The close agreement between the two sets of curves provides a mutual validation of both algorithms. For the surface flow we obtain the profile curves from the discrete surface by projecting the evolving base curves onto the plane.

Figure 4 provides a comparison between diffeomorphic flow with small σ (left)

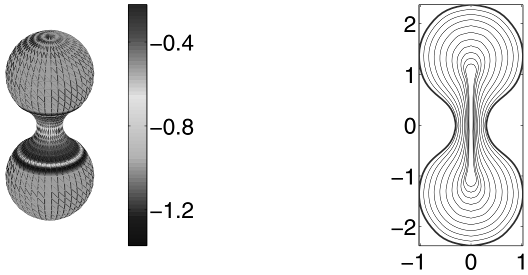


FIG. 5. *The flow of a dumbbell. Left: The dumbbell with mean curvature coloring. Right: The diffeomorphic mean curvature flow with $\sigma = 0.3$ and $T = 1$.*

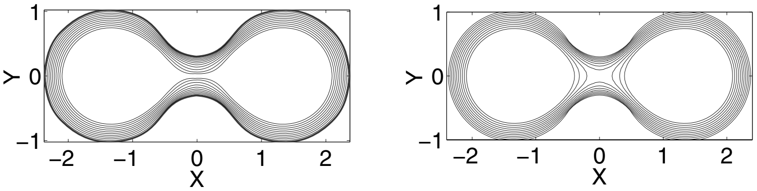


FIG. 6. *Comparison of flows for the dumbbell in Figure 5. Left: Diffeomorphic mean curvature flow with $\sigma = 0.1$ and $T = 0.25$. Right: Mean curvature flow with $T = 0.25$.*

and classical mean curvature flow (right). For the latter, the profile curve reaches the y axis, changes topology, and eventually becomes a sphere. On the other hand, for the diffeomorphic mean curvature flow, part of the curve very closely approaches the y axis but slows down as time increases to infinity and does not actually reach the y axis. In fact, the profile curve flows towards a semicircular shape. This experiment also illustrates the scale-dependent aspect for diffeomorphic flows. In Figures 3 (right) and 4 (left), diffeomorphic flows starting from the same surfaces have very different evolutions. The large value of σ in the first figure prevents the torus from collapsing on itself as it does in the second case, which is much closer to mean curvature evolution.

Dumbbell. Figures 5 and 6 show results for a dumbbell generated by a curve with neck shape $y = x^2 + c$ and $c = 0.3$. In the right-hand panel of Figure 6 we see that for the classical mean curvature flow, the thin neck breaks down and the dumbbell becomes two spheres. However, the diffeomorphic mean curvature flow does not break down even though, as we see in the right-hand panel of Figure 5, it flows towards two spheres connected by a very thin tube. One likely explanation for this shape is that the cylindrical part flows faster than the spherical part where the width of the neck is smaller than the kernel size.

Dumbbell with asymmetric ends. To illustrate typical problems encountered in real applications, we constructed a dumbbell shape with asymmetric ends. Figure 7 shows promising results. For classical mean curvature flow, the smaller end vanished quickly unlike in the diffeomorphic mean curvature flow.

Cortical surfaces. Figure 8 illustrates the application of diffeomorphic and classical mean curvature flows to a superior temporal gyrus cortical surface [35]. The voxel resolution of the image volume from which this surface was generated was 1 mm^3. For the diffeomorphic flows we used a kernel size of $\sigma = 0.3$. For surfaces with boundary,

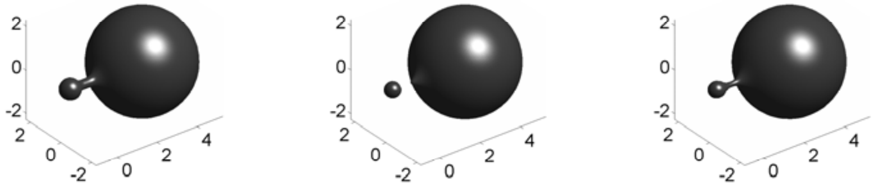


Fig. 7. *Comparison of flows for a dumbbell with asymmetric ends. Left: The dumbbell with asymmetric ends. Middle: Mean curvature flow for $T = 0.025$. Right: Diffeomorphic mean curvature flow with kernel size $\sigma = 0.1$ and $T = 0.025$.*

we simply flow the boundary vertex along with the closest interior vertex. The top row indicates the smoothing effect of both flows. Magnified views of a crest region of the surface indicate that mean curvature flow results in singular folds unlike diffeomorphic flows. Furthermore, the left-hand panel of Figure 9 shows that Hausdorff distances between the original surface and the final surfaces are within one voxel. The smoothing effect of the flows is also reflected in the mean curvature histograms in the right-hand panel of Figure 9.

6. Discussion. In this study, we proposed diffeomorphic surface flows as an alternative to classical mean curvature and Willmore flows. We obtained the flow equation for the elastic energy of a closed surface and proved the short-time existence and uniqueness of the flow. Then we examined the diffeomorphic mean curvature flow both by analyzing the case of a surface of revolution and by numerical experiments on arbitrary discrete surfaces. Our conjecture is that the solution continues to be well behaved for long times while preserving some of the characteristics of classical mean curvature flows (such as smoothing and decreasing area).

Furthermore, in computational anatomy applications, diffeomorphic surface flows can be used to evolve submanifolds of the brain such as planum temporale, superior temporal gyrus (STG), and cingulate cortical surfaces in neuroimaging studies of schizophrenia and auditory disorders. They can also be used for smoothing where the speed is controlled by changing the kernel size without changing the topology.

Diffeomorphic flows, while they clearly avoid topological changes, cannot, however, be considered as smoothing flows. Since they generate a diffeomorphic evolution, they cannot make a surface smoother (in terms of the number of derivatives) than it was initially. In particular, they cannot deal with surfaces corrupted by white noise. They can, however, have some smoothing effect, in the sense that they reduce the curvature of the surfaces on which they operate. There is an important scale factor in this regard, related to the scale of the kernel. Bumps larger than the kernel size will in general be removed in a way similar to classical flows, whereas small bumps are likely to survive after long time intervals. The choice of the kernel size therefore needs to be adapted to the roughness of the surface.

Several open problems remain. Among the numerical issues, there is a need to improve the discretization methods, especially in the case of the Willmore flow for which the computation of higher derivatives is a potential source of instability.

We have already mentioned the issue of long-term existence of the flow. As discussed in this paper, this requires controlling the smoothness of the surface during the evolution, which is made difficult by nonlocal interactions. The limit behavior of the evolution as the kernel size tends to zero is another open problem. It seems

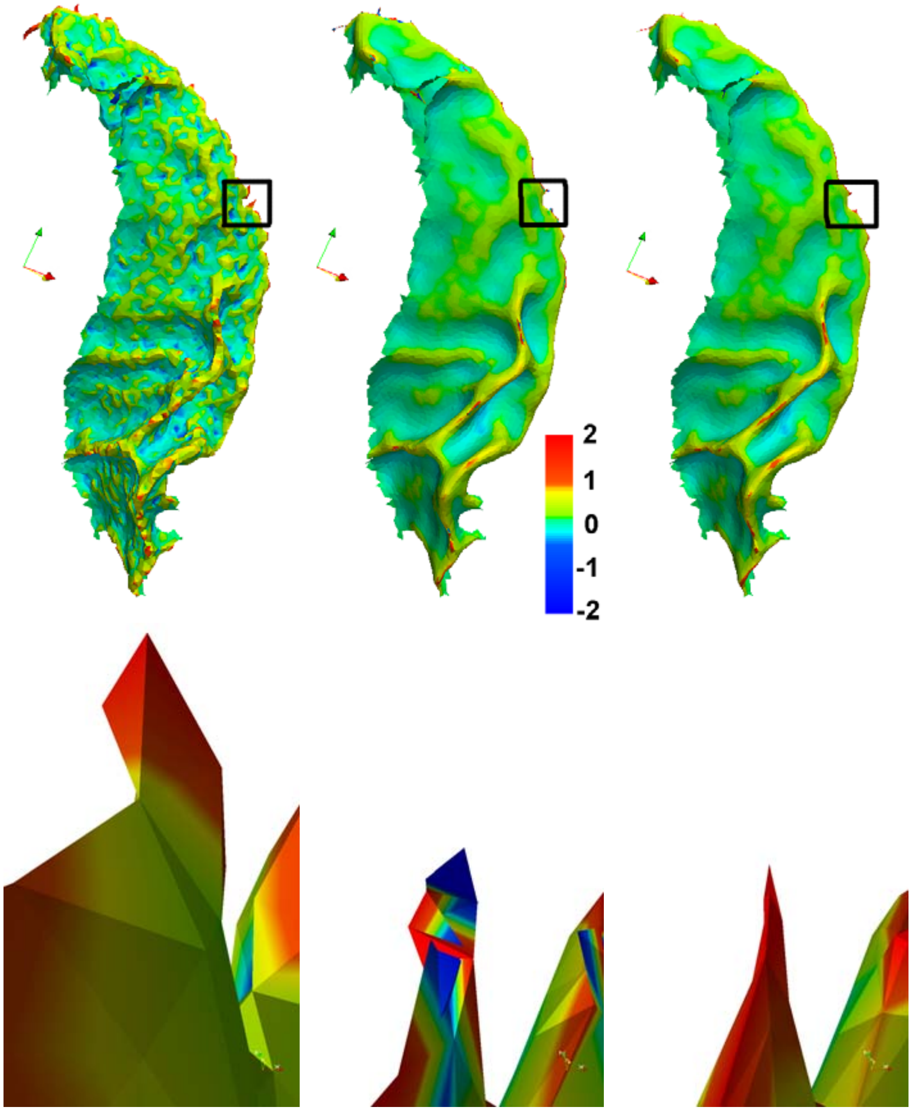


Fig. 8. *The top row shows STG cortical surfaces at $T = 0$ (left), $T = 1$ with mean curvature flow (middle), and $T = 1$ with diffeomorphic mean curvature flow using $\sigma = 0.3$ (right). Mean curvature is indicated by the vertical color bar. The bottom row shows the corresponding magnified view of the crest on the STG which is a subset of the region within the black borders indicated in the top row. The irregular color pattern in the magnified view of the mean curvature flow indicates that singularities occurred during the evolution with triangles crossing over.*

reasonable to expect that it should somewhat resemble the classical flows, but the nature of the convergence (and proof) needs to be investigated.

Another interesting issue, related to long-term evolution, is to characterize the limit shapes of the shrinking surfaces for the diffeomorphic mean curvature flow. Our experiments seem to indicate that such limit shapes exist and are not restricted to spheres.

It would be interesting also, for theoretical and practical purposes, to consider numerical schemes in which the kernel size is allowed to evolve with time, starting with

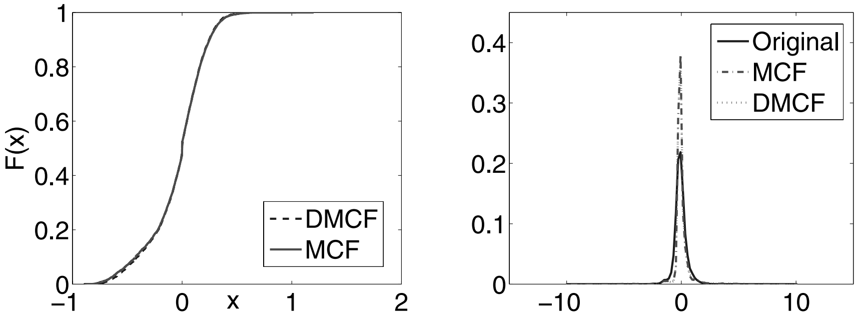


FIG. 9. *Left: Cumulative density function profiles of the oriented distances between the original surface and final surfaces. Right: Mean curvature histograms of the original and final surfaces at* $T = 1$.

a rigid evolution and progressively decreasing spatial smoothing to get closer to the mean curvature flow. The question here is to define sufficient conditions for such an annealed scheme that ensure a diffeomorphic evolution while providing a smoothing effect similar to that of the standard mean curvature flow.

While these topics provide interesting sources of future work, diffeomorphic surface flows already represent a promising family of surface evolutions. Initial experiments in this paper demonstrate several important features, regarding, in particular, the absence of topological change, that make them appropriate for a large range of practical situations.

Acknowledgments. We thank the reviewers for their extensive comments, and Dr. M. I. Miller, Dr. Y. Cao, and Dr. J. Glaunès for helpful discussions.

REFERENCES

[1] S. Akbarian, W. E. Bunney, Jr., S. G. Potkin, S. B. Wigal, J. O. Hagman, C. A. Sandman, and E. G. Jones, *Altered distribution of nicotinamide-adenine dinucleotide phosphate-diaphorase cells in frontal lobe of schizophrenics implies disturbances of cortical development*, Arch. Gen. Psychiatry, 50 (1993), pp. 169–77.

[2] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[3] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Springer-Verlag, New York, 2002.

[4] F. M. Benes and S. Berretta, *GABAergic interneurons: Implications for understanding schizophrenia and bipolar disorder*, Neuropsychopharmacology, 25 (2001), pp. 1–27.

[5] K. A. Brakke, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[6] Y. Cao, I. M. Miller, R. Winslow, and L. Younes, *Large deformation diffeomorphic metric mapping of vector fields*, IEEE Trans. Med. Imaging, 24 (2005), pp. 1216–1230.

[7] U. Clarenz, U. Diewald, G. Dziuk, M. Rumpf, and R. Rusu, *A finite element method for surface restoration with smooth boundary conditions*, Comput. Aided Geom. Design, 21 (2004), pp. 427–445.

[8] D. Cohen-Steiner and J. M. Morvan, *Restricted Delaunay triangulations and normal cycle*, in Proceedings of the 19th Annual Symposium on Computational Geometry (SCG '03), S. Fortune, ed., ACM Press, New York, 2003, pp. 312–321.

[9] K. Deckelnick, G. Zuik, and C. M. Elliott, *Computation of geometric partial differential equations and mean curvature flow*, Acta Numer., 14 (2005), pp. 139–232.

[10] M. C. Delfour and J. P. Zolésio, *Tangential calculus and shape derivatives*, in Shape Optimization and Optimal Design (Cambridge, 1999), Lecture Notes in Pure and Appl. Math. 216, Dekker, New York, 2001, pp. 37–60.

[11] P. Dupuis, U. Grenander, and M. I. Miller, *Variational problems on flows of diffeomorphisms for image matching*, Quart. Appl. Math., 56 (1998), pp. 587–600.

[12] K. Ecker, *Regularity Theory for Mean Curvature Flow*, Birkhäuser Boston, Boston, 2004.

[13] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.

[14] J. Glaunès, A. Trouvé, and L. Younes, *Modeling planar shape variation via Hamiltonian flows of curves*, in Statistics and Analysis of Shapes, H. Krim and A. Yezzi, Jr., eds., Birkhäuser Boston, Boston, 2006, pp. 335–363.

[15] J. Goldfeather and V. Interrante, *A novel cubic-order algorithm for approximating principal direction vectors*, ACM Trans. Graph., 23 (2004), pp. 45–63.

[16] L. Greengard and J. Strain, *The fast Gauss transform*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 79–94.

[17] U. Grenander and M. I. Miller, *Computational anatomy: An emerging discipline*, Quart. Appl. Math., 56 (1998), pp. 617–694.

[18] A. Gueziec and R. Hummel, *Exploiting triangulated surface extraction using tetrahedral decomposition*, IEEE Trans. Vis. Comput. Graph., 1 (1995), pp. 328–342.

[19] B. Hamann, *Curvature approximation for triangulated surfaces*, in Geometric Modelling, Comput. Suppl. 8, Springer-Verlag, Vienna, 1993, pp. 139–153.

[20] X. Han, C. Xu, U. Braga-Neto, and J. L. Prince, *Topology correction in brain cortex segmentation using a multiscale, graph-based algorithm*, IEEE Trans. Med. Imaging, 21 (2002), pp. 109–21.

[21] X. Han, C. Xu, and J. L. Prince, *A topology preserving deformable model using level sets*, in Proceedings of the IEEE Computer Science Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Vol. 2, 2001, pp. 765–770.

[22] W. Helfrich, *Elastic properties of lipid bilayers: Theory and possible experiments*, Z. Naturforsch., 28C (1973), pp. 693–703.

[23] K. Hildebrandt and K. Polthier, *Anisotropic filtering of non-linear surface features*, Comput. Graph. Forum, 23 (2004), pp. 391–400.

[24] S. C. Joshi and M. I. Miller, *Landmark matching via large deformation diffeomorphisms*, IEEE Trans. Image Process., 9 (2000), pp. 1357–1370.

[25] S. C. Joshi, J. Wang, M. I. Miller, D. C. VanEssen, and U. Grenander, *Differential geometry of the cortical surface*, in Vision Geometry IV, R. A. Melter, A. Y. Wu, F. L. Bookstein, and W. D. Green, eds., Proc. SPIE 2573, SPIE, Bellingham, WA, 1995, pp. 304–311.

[26] E. Kuwert and R. Schätzle, *The Willmore flow with small initial energy*, J. Differential Geom., 57 (2001), pp. 409–441.

[27] E. Kuwert and R. Schätzle, *Gradient flow for the Willmore functional*, Comm. Anal. Geom., 10 (2002), pp. 307–339.

[28] W. E. Lorensen and H. E. Cline, *Marching cubes: A high resolution* 3*D surface reconstruction algorithm*, in Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press, New York, 1987, pp. 163–169.

[29] U. F. Mayer and G. Simonett, *Self-intersections for the surface diffusion and the volume-preserving mean curvature flow*, Differential Integral Equations, 13 (2000), pp. 1189–1199.

[30] U. F. Mayer and G. Simonett, *A numerical scheme for axisymmetric solutions of curvature-driven free boundary problems, with applications to the Willmore flow*, Interfaces Free Bound., 4 (2002), pp. 89–109.

[31] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, *Discrete differential-geometry operators for triangulated* 2*-manifolds*, in Visualization and Mathematics III, H.-C. Hege and K. Polthier, eds., Springer-Verlag, Berlin, 2003, pp. 35–57.

[32] M. I. Miller, A. Trouvé, and L. Younes, *On the metrics and Euler-Lagrange equations of computational anatomy*, Annu. Rev. Biomed. Eng., 4 (2002), pp. 375–405.

[33] I. M. Mitchell and J. A. Templeton, *A toolbox of Hamilton–Jacobi solvers for analysis of nondeterministic continuous and hybrid systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 3414, Springer-Verlag, Berlin, 2005, pp. 480–494.

[34] J. C. C. Nitsche, *Boundary value problems for variational integrals involving surface curvatures*, Quart. Appl. Math., 51 (1993), pp. 363–387.

[35] J. T. Ratnanather, P. E. Barta, N. A. Honeycutt, N. G. Lee, H. M. Morris, A. C. Dziorny, M. K. Hurdal, G. D. Pearlson, and M. I. Miller, *Dynamic programming generation of boundaries of local coordinatized submanifolds in the neocortex: Application to the planum temporale*, NeuroImage, 20 (2003), pp. 359–377.

[36] J. T. Ratnanather, L. Wang, M. B. Nebel, M. Hosakere, X. Han, J. G. Csernansky, and M. I. Miller, *Validation of semiautomated methods for quantifying cingulate cortical metrics in schizophrenia*, Psych. Res.: Neuroimaging, 132 (2004), pp. 53–68.

[37] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[38] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.

[39] T. Tasdizen, R. Whitaker, P. Burchard, and S. Osher, *Geometric surface processing via normal maps*, ACM Trans. Graph., 22 (2003), pp. 1012–1033.

[40] A. Trouvé and L. Younes, *Local geometry of deformable templates*, SIAM J. Math. Anal., 37 (2005), pp. 17–59.

[41] M. Vaillant and J. Glaunès, *Surface matching via currents*, in Information Processing in Medical Imaging, Lecture Notes in Comput. Sci. 3565, Springer-Verlag, Berlin, 2005, pp. 381–392.

[42] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, Philadelphia, 1990.

[43] R. Whitaker, *Modeling deformable surfaces with level sets*, IEEE Comput. Graph. Appl., 24 (2004), pp. 6–9.

[44] T. J. Willmore, *Riemannian Geometry*, Oxford Sci. Publ., The Clarendon Press, Oxford University Press, New York, 1993.

[45] Z. Wood, H. Hoppe, M. Desbrun, and P. Schröder, *Removing excess topology from isosurfaces*, ACM Trans. Graph., 23 (2004), pp. 190–208.

[46] L. Younes, *Invariance, déformations et reconnaissance de formes*, Springer-Verlag, Berlin, 2004.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 825–843

THE NONLINEAR CRITICAL LAYER FOR KELVIN MODES ON A VORTEX WITH A CONTINUOUS VELOCITY PROFILE[*]

S. A. MASLOWE[†] AND N. NIGAM[†]

Abstract. We consider in this paper the propagation of neutral modes along a vortex with velocity profile $\bar{V}(r)$, r being the radial coordinate. In the linear inviscid stability theory for swirling flows, modes that are singular at some value of r denoted r_c, the critical point, are particularly significant. The singularity can be dealt with by adding viscous and/or nonlinear effects within a thin critical layer centered on the critical point. At high Reynolds numbers, the case of most interest in applications such as aeronautics and geophysical fluid dynamics, nonlinearity is the appropriate choice, although viscosity may still play a subtle role. We determine here the scaling and equations that govern the nonlinear critical layer. The method of characteristics is then employed to obtain an exact solution of the governing inviscid system composed of four coupled PDEs, of which two are nonlinear and two are linear. Finally, assuming zero phase change across the critical layer, solutions are obtained for the outer eigenvalue problem demonstrating the existence of modes not possible in a linear theory. This result may have important implications for the short wave cooperative instability mechanism that has received so much attention in the context of aircraft trailing vortices.

Key words. waves on vortices, Kelvin modes, nonlinear critical layers

AMS subject classification. 76D17

DOI. 10.1137/060658515

1. Introduction.

The propagation of helical perturbations to a columnar vortex seems to have been studied first by Lord Kelvin, whose results were published in 1880. In cylindrical coordinates (r, θ, z), the problem involves the investigation of infinitesimal perturbations (u_r, u_θ, u_z) superimposed on a flow with velocity profile $\{0, \bar{V}(r), 0\}$. Kelvin considered the case of a fluid in rigid rotation, i.e., $\bar{V} = \Omega_0 r$ contained within a cylinder of radius a. A single equation can be obtained for the pressure perturbation, and it is Bessel's equation of order m, where m is the azimuthal wavenumber. Imposition of the boundary conditions at $r = 0$ and $r = a$ leads to an eigenvalue problem for the frequency $\omega = \omega(k, m)$, where k is the axial wavenumber. The review article by Ash and Khorrami [1] is a convenient reference for the details.

In this paper, we are primarily interested in waves propagating on an unbounded vortex, and a model that has often been employed to study this phenomenon is the discontinuous Rankine vortex with velocity profile

$$\bar{V}(r) = \begin{cases} \Omega_0 r, & 0 \le r \le a, \\ \frac{\Omega_0 a^2}{r}, & r > a. \end{cases}$$

The solutions on either side of the discontinuity of vorticity at $r = a$ are matched using kinematic and pressure conditions, and this leads to an eigenvalue problem for the dispersion relation. Bessel functions are again involved, and the modal solutions obtained are termed Kelvin modes. The monograph by Saffman [2] details the analysis and presents dispersion curves for different azimuthal wavenumbers.

[*] Received by the editors April 29, 2006; accepted for publication (in revised form) June 1, 2007; published electronically January 16, 2008. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/siap/68-3/65851.html

[†] Department of Mathematics and Statistics, McGill University, Montreal, QC, H3A 2K6, Canada (maslowe@math.mcgill.ca, nigam@math.mcgill.ca). The second author was partially supported by the Fonds quebecois de la recherche sur la nature et les technologies.

Much of the recent research on the stability of vortices is motivated by the aircraft trailing vortex problem. A pair of counterrotating vortex filaments serves as a model that has been widely studied as representing the vortices shed from the wingtips of aircraft which, if they are jumbo jets, pose a major hazard to following aircraft. Given that the strength of the trailing vortices is related to the weight of the aircraft, it is clear that any smaller aircraft attempting to land behind the new Airbus A380 is at serious risk.

A number of theoretical investigations have treated the problem of a vortex subject to an external strain, this being a way of modelling the strain induced on one member of a pair of trailing vortices by the other. A long wave instability first explained by Crow [3] initially received the most attention, but more recently the short wave cooperative instability mechanism (often termed the elliptic instability) has attracted a great deal of interest. The latter mechanism involves interacting Kelvin waves. Specifically, Moore and Saffman [4] showed that for an arbitrary strained vortex two neutral modes are coupled by the strain field if a certain resonance condition is satisfied. They derived an approximate expression for the growth rate of the resonant modes by an asymptotic analysis valid when the strain field is small. The first quantitative investigation of this instability was by Tsai and Widnall [5], who employed the discontinuous Rankine vortex in their calculations. They found that the most unstable perturbations corresponded to a pair of Kelvin modes having zero frequency and azimuthal wavenumbers $m = \pm 1$.

Real vortices, however, have continuous profiles, and the use of Rankine vortices in theoretical studies was criticized in the review article by Spalart [6] (see section 2.2). Clearly, it is important to ask what effect the use of a profile whose vorticity is continuous might have on this instability mechanism. Sipp and Jacquin [7] have, in fact, recently done so, and they concluded that the "Widnall instabilities" will not occur because of the presence of a critical layer in the continuous case. Their argument, which is correct as far as it goes, is based on linear viscous stability calculations for the Lamb–Oseen vortex $\bar{V}(r) = (1 - e^{-r^2})/r$ which show that the neutral Kelvin modes required for the resonant interaction discussed in [4] and [5] would be damped in the continuous case.

In this paper, we reexamine the question and investigate the effect of retaining nonlinear terms in the critical layer rather than viscous terms, as is usually done in the theory of hydrodynamic stability. This is of interest in its own right as part of the theory of Kelvin modes, and its pertinence to the cooperative elliptic instability mechanism provides further motivation. The possibility that nonlinear critical layer modes could be neutral rather than damped was, in fact, suggested in [7, p. 265]. In section 3, we will determine a parameter that measures the relative importance of nonlinearity to viscosity. The larger the Reynolds number, the smaller the perturbation amplitude needs to be for nonlinearity to be the appropriate choice. A number sometimes cited as representative for trailing vortices behind jumbo jets is $Re = 10^7$. Clearly, this is large enough to motivate the formulation of a nonlinear approach. Even if the turbulence usually present is accounted for by introducing an effective Reynolds number, this value is still very large. Gerz and Ehret [8], in an investigation of the influence of wingtip vortices on atmospheric pollution caused by the jet exhaust, estimate the effective Re as $6 \cdot 10^5$ for the wake behind a Boeing 747.

Although the foregoing discussion focused on the trailing vortex problem, the results are pertinent to other applications in engineering and geophysical fluid dynamics. In turbomachinery, for example, the flow through a duct is sometimes modelled by representing the flow by a superposition of a solid body rotation and a potential axial vortex [9]. And in geophysical fluid dynamics, an important application is to hurricanes. In their numerical simulations of a hurricane, Chen, Brunet, and Yau [10] find that absorption of vortex Rossby waves at the critical level leads to an acceleration of the mean wind in the lower troposphere (the location of the critical layer is indicated in Figure 11 of their paper).

Before presenting the analysis, it is worth noting that the critical point singularity in a stratified shear flow has a very similar behavior to that occurring in a vortex. It is therefore possible to anticipate certain results based on those that have been demonstrated for stratified shear flows. For example, Miles [11] proved that when the local Richardson number is everywhere greater than 1/4, all singular modes must decay according to linear theory. However, when the critical layer is nonlinear and inviscid, singular neutral modes have been shown to exist (see section 3.1.1 of the review article by Maslowe [12]). For Kelvin modes on vortices, we will show that the same is true; i.e., inviscid nonlinear modes exist in regions of parameter space where they would be damped if viscosity were used to deal with the critical layer. As a result, we revive the possibility not only of cooperative instabilities but of other instability mechanisms that have been observed experimentally in which Kelvin modes interact to destroy vortices. In the experiments of Maxworthy, Hopfinger, and Redekopp [13], for example, unstable interactions between axisymmetric and helical waves were observed, the outcome of which depended on the amplitude of the axisymmetric mode.

The reason that neutral modes with nonlinear critical layers can exist when they would be damped in a linear, viscous theory is a direct result of the absence of any phase change across the singular critical point. This means essentially that terms with branch points are written simply as absolute values, whereas in the viscous theory, a term with a branch point at $r = r_c$ is written as $|r - r_c|\, e^{i\,\phi}$ for $r < r_c$, and the phase change ϕ is nonzero. In either case, the result must be derived by determining what outer solution can be matched to the critical layer solution.

Before proceeding with the analysis, we mention briefly that the simpler two-dimensional case of waves propagating only in the azimuthal direction, i.e., $k = 0$, has been investigated recently by Le Dizès [14] and by Balmforth, Smith, and Young [15]. Each paper points out the relevance of the results to plasma physics, as well as to vortices. The critical layer analysis in [14] is closer to our own, one reason being that the waves are forced in [15] and may be transient, but some brief comparisons will be made after presenting our own analysis for helical modes.

2. Formulation and outer expansion. We consider small-amplitude helical perturbations to a vortex with azimuthal velocity profile $\bar{V}(r)$ and a corresponding radial pressure distribution $\bar{p}(r)$. Away from the critical layer, the perturbations are sinusoidal with phase $\xi = kz + m\theta - \omega t$ and, because we are dealing with neutral modes, it will be convenient to use ξ as an independent variable. The momentum and continuity equations can then be written as

$$\left(\frac{m}{r}\,u_\theta - \omega\right)\frac{\partial u_r}{\partial \xi} + u_r\,\frac{\partial u_r}{\partial r} - \frac{u_\theta^2}{r} + k\,u_z\,\frac{\partial u_r}{\partial \xi} = -\frac{\partial p^*}{\partial r} + \frac{1}{Re}\,\frac{\partial^2 u_r}{\partial r^2}, \tag{2.1a}$$

$$\left(\frac{m}{r}\,u_\theta - \omega\right)\frac{\partial u_\theta}{\partial \xi} + u_r\,\frac{\partial u_\theta}{\partial r} + \frac{u_r\,u_\theta}{r} + k\,u_z\,\frac{\partial u_\theta}{\partial \xi} = -\frac{m}{r}\frac{\partial p^*}{\partial \xi} + \frac{1}{Re}\,\frac{\partial^2 u_\theta}{\partial r^2}, \tag{2.1b}$$

$$\left(\frac{m}{r}\,u_\theta - \omega\right)\frac{\partial u_z}{\partial \xi} + u_r\,\frac{\partial u_z}{\partial r} + k\,u_z\,\frac{\partial u_z}{\partial \xi} = -k\frac{\partial p^*}{\partial \xi} + \frac{1}{Re}\,\frac{\partial^2 u_z}{\partial r^2}, \tag{2.1c}$$

$$\text{and}\quad \frac{\partial (r\,u_r)}{\partial r} + m\,\frac{\partial u_\theta}{\partial \xi} + k\,r\,\frac{\partial u_z}{\partial \xi} = 0. \tag{2.1d}$$

Our analysis being primarily inviscid, we have retained in the above momentum equations only those viscous terms involving second derivatives with respect to r, because these terms will be the largest in the critical layer.

The scaling employed in the foregoing equations deserves some discussion because we wish to identify parameter regimes where analytical progress is possible. To begin, we denote by Ω_0 the angular velocity of the vortex at its center and use this to scale the frequencies and time. A characteristic length scale for the vortex denoted a is used to nondimensionalize r and the wavenumber k, while $\Omega_0\, a$ is employed to scale the velocities. Finally, the dimensionless pressure p^* is obtained by dividing the actual pressure by $\rho\,\Omega_0^2\, a^2$, where ρ is the constant density. After introducing this scaling into the momentum equations, the Reynolds number $Re = a^2\Omega_0/\nu$, where ν is the kinematic viscosity.

We next consider the linear, inviscid theory because it describes the perturbation to leading order in the outer region. A separation of variables can be achieved in the linearized equations by writing

$$u_r = \varepsilon\, u(r) \sin\xi, \tag{2.2a}$$

$$u_\theta = \bar{V}(r) + \varepsilon\, v(r) \cos\xi, \tag{2.2b}$$

$$u_z = \varepsilon\, w(r) \cos\xi, \tag{2.2c}$$

$$\text{and} \qquad p^* = \bar{p}(r) + \varepsilon\, p(r) \cos\xi, \tag{2.2d}$$

where $\varepsilon \ll 1$ is a dimensionless amplitude parameter. After linearizing and then substituting (2.2a)–(2.2d) into (2.1a)–(2.1d), we obtain the system

$$\gamma(r)\, u = 2\,\frac{\bar{V}}{r}\, v - p\,', \tag{2.3a}$$

$$\gamma(r)\, v = \frac{1}{r}\,(r\bar{V})'\, u - \frac{m}{r}\, p, \tag{2.3b}$$

$$\gamma(r)\, w = -k\, p, \tag{2.3c}$$

$$(r\, u)' = m\, v + k\, r\, w, \tag{2.3d}$$

where

$$\gamma(r) = m\,\frac{\bar{V}}{r} - \omega = m\,\bar{\Omega} - \omega. \tag{2.4}$$

Critical point singularities occur at any value of r for which $\gamma(r) = 0$.

Equations (2.3a)–(2.3d) can be combined into a single second order differential equation for $u(r)$, namely,

$$\gamma^2 D\{SD_* u\} - \left\{\gamma^2 + \frac{m\gamma}{r^2}\left(D[SD(r\bar{V})] - 3\,\frac{S}{r} D(r\bar{V})\right) - 2\,\bar{V}\, k^2\,\frac{S}{r}\, Q(r)\right\} u = 0, \tag{2.5}$$

where

$$D = \frac{d}{dr}, \quad D_* = \frac{d}{dr} + \frac{1}{r}, \quad S = \frac{r^2}{m^2 + k^2 r^2}, \quad \text{and} \quad Q(r) = \frac{D(r\bar{V})}{r}.$$

$Q(r)$ can be recognized as the vorticity of the mean flow nondimensionalized with respect to Ω_0, the angular velocity at the center of the vortex.

Equation (2.5) can be obtained from the equation derived by Howard and Gupta [16] for swirling flows by setting the axial velocity $W = 0$ in equation (18) of [16].

Noting the similarity of their equation (18) to the Taylor–Goldstein equation governing stratified shear flows, Howard and Gupta derived a Richardson number 1/4 stability theorem for swirling flows, employing an integral approach as in Howard's earlier paper on stratified flows [17]. This theorem, however, was limited to axisymmetric, i.e., $m = 0$, perturbations, thus underlining the importance of perturbations with $m \neq 0$, where only a bound on the growth rate could be obtained.

The mathematical similarities to the case of a stratified shear flow are nonetheless useful, and it will be seen that in our analysis the paper by Miles [18] is most pertinent. Miles used Frobenius expansions near the critical point to derive a number of important results, including the Richardson number 1/4 theorem. Following his approach and notation, we expand all terms in (2.5) around the critical point r_c to obtain a solution valid locally having the form

$$u(r) = A\,u_+(r) + B\,u_-(r), \tag{2.6}$$

where

$$u_\pm(r) = (r - r_c)^{\frac{1}{2}\,(1 \pm \nu)}\,w_\pm(r) \tag{2.7}$$

and the functions $w_\pm(r)$ are regular in the neighborhood of r_c. We have defined a local Richardson number analogous to the one arising in stratified shear flows by

$$J_c = \frac{2\,k^2\,\bar{V}_c\,Q_c}{r_c\,(\gamma_c')^2}, \tag{2.8}$$

and the parameter ν in (2.7) is related to J_c by $\nu = (1 - 4\,J_c)^{1/2}$.

Miles used arguments based on the variation of the Reynolds stress to prove a number of useful results that apply to singular neutral modes. For example, within the framework of linear theory, a neutral mode comprising part of a stability boundary must be proportional to one or the other of the Frobenius solutions. Even though the expression for the Reynolds stress is quite different in cylindrical coordinates, we show in the appendix that the same conclusion applies here; i.e., on a stability boundary with $J_c < 1/4$, the case we treat in this paper, either A or B must be zero in (2.6).

There are two exceptional cases, however, that should be mentioned. First, when J_c is small, the second term in the series for $w_-(r)$ in (2.7) is very large, becoming infinite as $J_c \to 0$. That case is of interest in the vortex problem because it arises, for example, when r_c is far from the center and the vorticity Q_c is then small. The Frobenius solution u_- in (2.6) can then be replaced by a linear combination of u_+ and u_- that is well behaved as $J_c \to 0$; the associated nonlinear critical layer theory has been developed by Caillol and Maslowe [19]. The second exceptional case occurs when J_c is greater than 1/4 so that the Frobenius exponents are complex. We do not treat that case here, but it does arise, for example, when there is an external forcing. Let us proceed now to the derivation of the critical layer equations for the case of $J_c \sim O(1)$ with ν real.

3. Critical layer scaling and governing equations.

There are several ways to determine the scaling for the nonlinear critical layer. Of these, the most direct is to try to get a balance between linear inertial terms and the nonlinear terms with the largest derivatives in r. Because we are dealing with a system, however, different equations yield different results. For the present Kelvin wave problem, either the v or the w momentum equation in (2.1) leads to the correct scaling, but that was not obvious a priori. That being the case, it is a great help to have an alternative method. The most reliable is to examine several terms in the outer expansion which proceeds in powers of ε, the amplitude parameter. The power of ε where the expansion first breaks down yields the critical layer thickness. We illustrate this below, but first the behavior of all variables near r_c in the linear problem must be determined.

Let us concentrate on the case $J_c < 1/4$ corresponding to (2.6)–(2.8). The most singular Frobenius solution is the one with the minus sign in front of ν; we denote this exponent δ. Because the pressure perturbation satisfies an equation nearly identical to the one satisfied by u (see p. 244 of [2]), it will have the same behavior near r_c. Although not a trivial exercise, consideration of the system (2.3a)–(2.3d) leads to the conclusion that

$$v \sim (r - r_c)^{\delta-1}, \quad w \sim (r - r_c)^{\delta-1}, \quad \text{and} \quad p \sim (r - r_c)^{\delta}, \tag{3.1}$$

where to be consistent with the first of equations (2.3)

$$\left(\frac{dp}{dr} - \frac{2\bar{V}}{r}\, v\right)_c \sim (r - r_c)^{1+\delta}. \tag{3.2}$$

Considering now higher order terms in the outer expansion, if we were to write (2.2) as an expansion in powers of ε, then the $O(\varepsilon^2)$ terms would involve second harmonics, and in particular (2.2a) would include a term $u_2(r)\sin 2\xi$. A single nonhomogeneous second order ODE can be derived for $u_2(r)$ by following exactly the same steps that were used to obtain (2.5), the equation for $u(r)$, from (2.3a)–(2.3d). (A detailed derivation of the Howard–Gupta equation is given in section 2.4 of [1].) For the purpose at hand, however, only the most singular of the nonhomogeneous terms need be considered. The foregoing procedure leads to the result that $u_2 \sim (r-r_c)^{2\delta-2}$, and we see that the first two terms in the expansion for u_r become the same order of magnitude when $(r - r_c) \sim \varepsilon^{\frac{1}{2-\delta}}$.

From the behavior deduced immediately above, we find that appropriate independent variables in the nonlinear critical layer are

$$\xi = k\, z + m\, \theta - \omega\, t \qquad \text{and} \qquad R = \frac{r - r_c}{\varepsilon^{\beta}}, \qquad \text{where} \qquad \beta = \frac{1}{2-\delta}. \tag{3.3}$$

When $k = 0$, $\delta = 0$ and $\beta = 1/2$, as in [14] and [15]. With the critical layer thickness now determined, we can deduce from (2.2b) that in a frame of reference rotating with the angular velocity at the critical point, the magnitude of the azimuthal velocity perturbation is equal to that of the mean flow. This follows by expanding $\bar{V}$ in a Taylor series about r_c, and from the behavior of $v(r)$ as given by (3.1), it can be seen that $\varepsilon\, v(r)$ is the same order of magnitude as the mean flow when $(r - r_c) \sim O(\varepsilon^{\beta})$. Noting that point, an appropriate scaling for $V(R, \xi)$, the azimuthal velocity in the critical layer, is given by

$$u_\theta - \bar{V}_c \sim \bar{V}_c{}'(r - r_c) + \varepsilon\, v(r) \cos\xi = \varepsilon^{\beta}[V(R, \xi) + \bar{\Omega}_c\, R], \tag{3.4}$$

where the R-term is included because it simplifies the governing equations. The remaining dependent variables in the case $J_c < 1/4$ (but not small) are scaled as

$$u_r = \varepsilon^{2\beta} U(R, \xi), \quad u_z = \varepsilon^{\beta} W(R, \xi), \quad \text{and} \quad p^* - \frac{1}{2}\, \bar{\Omega}_c{}^2 r^2 = \varepsilon^{2\beta} P(R, \xi). \tag{3.5}$$

Now substituting (3.3)–(3.5) into the governing equations (2.1), the nonlinear critical layer equations to lowest order are the following:

$$r_c \frac{\partial P}{\partial R} - 2\,\bar{V}_c V = 0, \tag{3.6a}$$

$$2\,\frac{\bar{V}_c}{r_c} U + \frac{m}{r_c}\frac{\partial P}{\partial \xi} = -\left[U\frac{\partial V}{\partial R} + \left(\frac{m}{r_c}V + kW\right)\frac{\partial V}{\partial \xi}\right] + \lambda\,\frac{\partial^2 V}{\partial R^2}, \tag{3.6b}$$

$$k\frac{\partial P}{\partial \xi} = -\left[U\frac{\partial W}{\partial R} + \left(\frac{m}{r_c}V + kW\right)\frac{\partial W}{\partial \xi}\right] + \lambda\,\frac{\partial^2 W}{\partial R^2}, \tag{3.6c}$$

(3.6d) and $r_c\frac{\partial U}{\partial R} + m\frac{\partial V}{\partial \xi} + kr_c\frac{\partial W}{\partial \xi} = 0.$

The parameter $\lambda = 1/Re\,\varepsilon^{3\beta}$; taking $\lambda \ll 1$, means that the nonlinear critical layer thickness ε^{β} is greater than that of the viscous critical layer, whose thickness is $Re^{-1/3}$. In most applications outside of the laboratory, that condition will be satisfied easily.

The critical layer problem is highly nonlinear, and the solution even at lowest order involves all the harmonics. In matching to the outer expansion, however, we can ignore higher harmonics because they decay more rapidly at large R than those terms involving the primary mode. After expanding the reduced pressure in (3.5) in a Taylor series about r_c and transforming to inner variables, we obtain the following asymptotic conditions:

$$U \sim u_0\,R^{\delta}\sin\xi, \quad V \sim \frac{\gamma_c' r_c}{m}\,R + v_0\,R^{\delta-1}\cos\xi, \quad W \sim w_0\,R^{\delta-1}\cos\xi,$$

$$\text{and} \qquad P \sim \frac{\gamma_c'\bar{V}_c}{m}\,R^2 + p_0\,R^{\delta}\cos\xi \qquad \text{as} \qquad R \to \infty. \tag{3.7}$$

Of the four constants u_0, v_0, w_0, and p_0, one is arbitrary. A convenient choice is $p_0 = 1$, and the system (2.3) can then be used to express the other three constants in terms of this constant.

3.1. The $\boldsymbol{\lambda = 0}$ limit. Given that the Reynolds number is very large in most applications, the $\lambda = 0$ limit is clearly of considerable interest even though we expect viscosity to still play a subtle role. For example, arbitrary functions that arise in integrating the governing PDEs can be determined uniquely only by introducing a small viscosity. Moreover, in related earlier studies thin viscous layers were required along streamlines separating open and closed regions in order for the vorticity and velocity components to be continuous. Most of the flow field is inviscid, though, and an exact solution of the system (3.6) will now be presented for that case.

We begin by writing the system (3.6) in the following matrix form:

$$\mathbf{A}(\mathbf{u})\,\frac{\partial \mathbf{u}}{\partial \xi} + \mathbf{B}(\mathbf{u})\,\frac{\partial \mathbf{u}}{\partial R} = \mathbf{c}, \tag{3.8}$$

where $\mathbf{A}$ and $\mathbf{B}$ are 4×4 matrices and the vectors $\mathbf{u}$ and $\mathbf{c}$ are given by

$$\mathbf{u}(R,\xi) = \begin{pmatrix} P(R,\xi) \\ U(R,\xi) \\ V(R,\xi) \\ W(R,\xi) \end{pmatrix} \qquad \text{and} \qquad \mathbf{c} = \begin{pmatrix} 2\,\Omega_c V(R,\xi) \\ -2\,\Omega_c U(R,\xi) \\ 0 \\ 0 \end{pmatrix}. \tag{3.9}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A}(\mathbf{u}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{m}{r_c} & 0 & \alpha & 0 \\ k & 0 & 0 & \alpha \\ 0 & 0 & m & kr_c \end{pmatrix}, \quad \mathbf{B}(\mathbf{u}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & U(R,\xi) & 0 \\ 0 & 0 & 0 & U(R,\xi) \\ 0 & r_c & 0 & 0 \end{pmatrix},$$

where $\alpha(R,\xi) = (m/r_c)V(R,\xi) + kW(R,\xi)$.

It is clear that $\mathbf{B}(\mathbf{u})$ becomes singular when $U = 0$. This tells us that solving the system by the method of lines (using a spectral collocation method in the ξ direction followed by integration in the R variable) will lead to a numerically stiff system since (3.6a)–(3.6d) turn into a system of differential-algebraic equations. Indeed, when $U = 0$, we are led to the constraints $V = 0$, $P =$ *constant*, $W =$ *constant*. This is also the reason why the small-λ limit of equations (3.6) is difficult (in the sense of stability) to simulate numerically. We should also note that the low-rank behavior of $\mathbf{A}(\mathbf{u})$ does not present computational difficulties. In order to deal with the possibly singular behavior of $\mathbf{B}$, we now introduce an alternate solution strategy.

3.2. Solution by the method of characteristics. If we denote by $\frac{dR}{d\xi} = \frac{1}{\mu}$ the slope of the characteristics, they are given by the roots of the characteristic polynomial $|\mathbf{A} - \mu\mathbf{B}| = 0$. For the system (3.6), this condition yields

$$det(\mathbf{A} - \mu\mathbf{B}) = \mu^2 r_c(\alpha - \mu U)^2 = 0,$$

where $\alpha = (m/r_c)V + kW$. Even though the system is not totally hyperbolic, i.e., there are only two characteristic directions, it develops that we can still solve (3.6) by integrating along the two characteristic directions given by

$$\frac{dR}{d\xi} = \frac{U}{(m/r_c)V + kW} \qquad \text{and} \qquad \frac{d\xi}{dR} = 0. \tag{3.10}$$

In accordance with the above result, we define a family of characteristics by

$$\left(\frac{\partial R}{\partial s}\right)_\tau = U \qquad \text{and} \qquad \left(\frac{\partial \xi}{\partial s}\right)_\tau = \frac{m}{r_c}\,V + kW, \tag{3.11}$$

where s measures distance along a characteristic and τ is a parameter identifying a particular characteristic. Eliminating $\frac{\partial P}{\partial \xi}$ from the azimuthal and axial momentum equations in (3.6) and using the characteristic equations (3.11), we obtain the following integral:

$$r_c\, V + 2\,\bar{V}_c\, R - \frac{m}{k}\, W = F(\tau). \tag{3.12}$$

A second integral can be obtained along the same family of characteristics by writing the directional derivative for the pressure, namely,

$$\left(\frac{\partial P}{\partial s}\right)_\tau = \left(\frac{\partial P}{\partial R}\right)_\xi \left(\frac{\partial R}{\partial s}\right)_\tau + \left(\frac{\partial P}{\partial \xi}\right)_R \left(\frac{\partial \xi}{\partial s}\right)_\tau. \tag{3.13}$$

Expressions for $\partial P/\partial R$ and $\partial P/\partial \xi$ are provided by (3.6a) and (3.6c), i.e., the radial and axial momentum equations. The characteristic equations (3.11) yield the other partial derivatives, and substitution of these four partial derivatives into (3.13) then allows us to integrate with respect to s to obtain the following expression for the pressure:

$$P = a_0\,\frac{m}{k}\,R\,W + \left(a_0 R - \frac{m}{kr_c^2}\,W\right) F(\tau) - a_0 \bar{V}_c R^2 - \frac{1}{2}\left[1 + \left(\frac{m}{kr_c}\right)^2\right] W^2 + G(\tau), \tag{3.14}$$

where $a_0 = 2\,\bar{V}_c/r_c^2$.

The arbitrary functions $F(\tau)$ and $G(\tau)$ are determined by matching to the outer solution and will be specified below. First, however, two more equations are required to complete the solution. One is provided by the radial momentum equation; i.e., (3.6a) gives us a simple expression for $\frac{\partial P}{\partial R}$, and we can integrate along the vertical lines $\xi = \mathit{const.}$, because they are characteristics. The fourth relationship that we require is obtained not by integrating along a characteristic but by defining the parameter τ in such a way that the continuity equation is satisfied; the characteristics $\tau = \mathit{const.}$ are then analogous to the streamlines in two-dimensional flows. Specifically, we require τ to satisfy the conditions

$$\frac{\partial \tau}{\partial \xi} = U \qquad \text{and} \qquad \frac{\partial \tau}{\partial R} = -(m/r_c)V - kW. \tag{3.15}$$

The numerical procedure that we use is to begin the integration at a large value of $|R|$ and then integrate toward the center of the critical layer. This allows us to determine an expression for τ that is valid as $R \to \infty$ that we can use to begin the integration. First, we integrate the second of equations (3.15); the resulting expression for τ contains an arbitrary function of ξ which can then be determined by differentiating with respect to ξ and comparing with the asymptotic behavior of U given in (3.7). It can be verified that the following expression for τ satisfies (3.15), as well as being consistent with the asymptotic conditions (3.7) and the continuity equation:

$$\tau = -\gamma_c' \frac{R^2}{2} - u_0 \, |R|^\delta \cos \xi. \tag{3.16}$$

Noting that $\tau \sim R^2$ for large $|R|$, it can be seen that $F \sim \sqrt{\tau}$ in (3.12) and that $G \sim \tau$ in (3.14). By substituting into (3.12) and (3.14) and using the asymptotic conditions (3.7), we find more precisely that

$$F(\tau) = (r\,\bar{V})_c' \sqrt{-2\,\tau/\gamma_c'} \qquad \text{and} \qquad G = \frac{2\,\bar{V}_c\,(r\,\bar{V})_c'}{\gamma_c'\,r_c^2}\,\tau. \tag{3.17}$$

3.3. Numerical procedure. As mentioned in the previous section, we begin integration of the characteristic equations for P, V, W, and the radial variable R at large initial values of R. We first locate $\tau(\xi)$ for this large value of R using (3.16). The values of $\tau(\xi)$ will subsequently be decreased by a *fixed* amount, d_τ, which in turn changes V, W, and R through the second of equations (3.15). We implement a backward Euler update:

$$d_\tau = (\tau - \tau_{old}) = -\left(\frac{m}{r_c}V + kW\right)(R_{old} - R). \tag{3.18}$$

Here the subscripts $_{old}$ refer to the values of the variable at the previous τ-step. We also write an Euler update to compute the new values of P in terms of its current values, and use this to update V, W, and R. We are thus led to a nonlinear system of four equations comprising (3.12), (3.14), (3.18), and

$$r_c(P - P_{old}) - 2\bar{V}_c V(R - R_{old}) = 0.$$

The nonlinear system needs to be solved for the new values of (P, V, W, R) at each τ-step. The manifold on which these solutions evolve seems particularly sensitive to initial guesses in V and W; hence, we retain only the updated P and R values from each iteration. We can then explicitly calculate from these the updated values of V and W. We note here that though it may seem appealing to solve for V and W in terms of P and R using (3.12) and (3.14), the latter is a quadratic in W. This would force us to make a choice for the sign of the root (and is, indeed, the source of the numerical sensitivity to initial guesses in V and W). Instead, we use all four equations to derive explicit expressions for V and W in terms of the other variables.

Certain constants must be fixed in order to initiate the integration from above and below the critical layer, so the procedure is discussed briefly here. If we assume that in crossing the critical layer there is no phase change, then absolute values are used to deal with the branch point at r_c in (2.5)–(2.7), the Frobenius solution for u. Two limiting cases were treated in [19]; in the first, it was assumed that the vorticity $Q_c \ll 1$, whereas the second case assumed waves that are long in the axial direction, i.e., $k \ll 1$. In both limits, it was found as in previous nonlinear critical layer studies that there was no phase change across the critical layer. It is therefore likely that the same result is generally true for vortical flows with $J_c < 1/4$. Consistent with the assumption of zero phase change, we take u_0 in (3.7) to be the same on both sides of the critical layer, in which case U will be an even function. Observing that U is differentiated with respect to R in the continuity equation, i.e., the last of equations (3.6), it is clear that V and W will be odd functions of R so that u_0 and w_0 in (3.7) must have opposite signs above and below the critical layer. Similar considerations lead us to conclude that P is an even function.

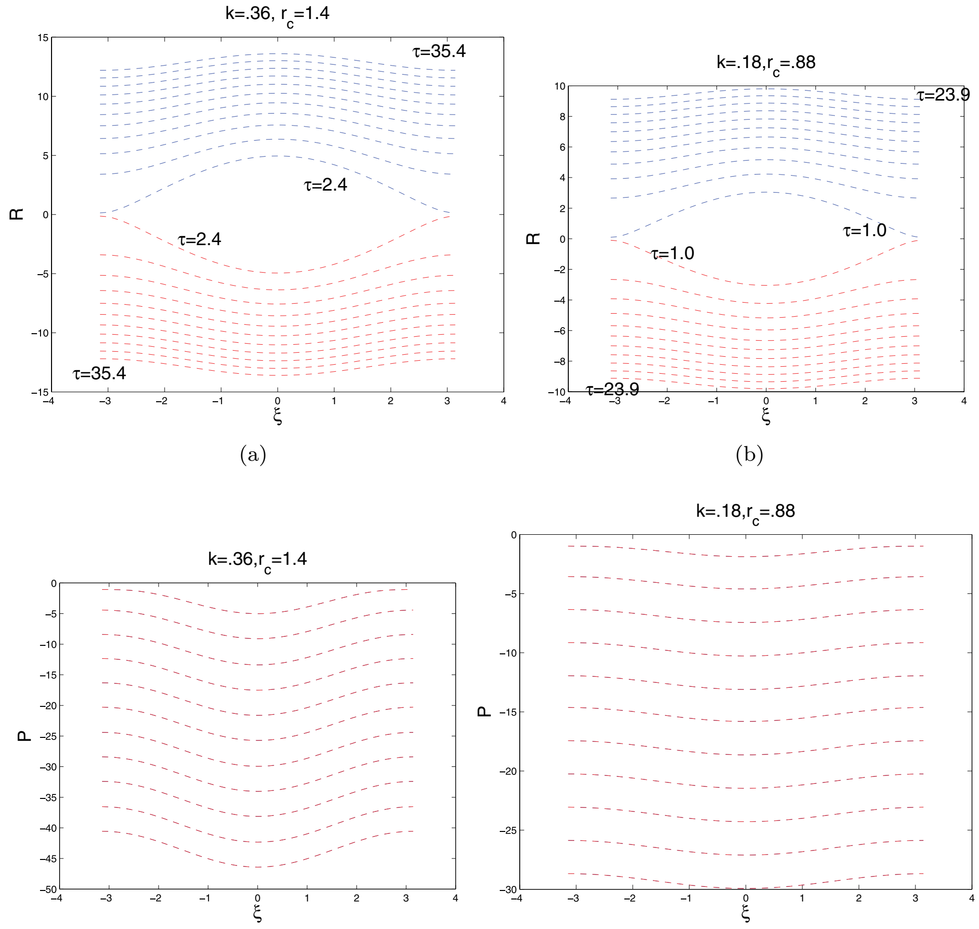
The foregoing considerations were used to compute the solution illustrated in Figures 3.1–3.4 for a Lamb–Oseen vortex profile. Initial values were obtained from (3.7), and the computation was initiated at large values of τ on both sides of the critical layer; i.e., τ is even in R. This beginning value of τ depends on the choice of k, r_c, and γ_c'.

In Figure 3.1, we present the characteristics R and pressures P obtained by the procedure described above. We repeated the procedure with several values of d_τ, and present the converged solutions. We show the results of two experiments; in Figures 3.1(a) and 3.1(c), $k = 0.36$, $m = 1$, $r_c = 1.4$, $\gamma_c' = -0.424907$. In Figures 3.1(b) and 3.1(d), $k = 0.18$, $m = 1$, $r_c = 0.88$, $\gamma_c' = -0.53425$. The procedure used is to decrease τ in increments of 0.05 until a characteristic is reached that goes beyond the corners at $R = 0$ and $\xi = \pm\pi$. This occurs on the characteristic $\tau = 2.4$ in the first experiment and on $\tau = 1.0$ in the second.

For clarity, only a few of the characteristics are shown. The similarity of Figures 3.1(a) and 3.1(c) compared with the streamline patterns in the stratified shear flow computations in Figure 2 of Maslowe [20] is striking, even including a cusp at the corners. Although the characteristics are not streamlines in our problem, they do serve the same purpose mathematically by providing a solution of the continuity equation; thus to this extent comparisons are valid.

Figure 3.2 shows that the nonlinear critical layer equations (3.8)–(3.9) completely eliminate the singular behavior exhibited by the linear problem. We show the fields for the azimuthal and axial velocity components V and W, respectively, because these are the most singular according to (3.8). We again chose $k = 0.36$, $r_c = 1.4$, $\gamma_c' = -0.424960$. For this particular example, $\delta = 0.230$, so that $v \sim |r - r_c|^{-0.770}$ as $|r - r_c| \to 0$ in the outer problem and w has the same behavior.

Our solution is not yet complete because we must still deal with the region of closed characteristics. In addition, there are discontinuities in V and W at the corners where the separatrices meet, although the radial velocity U is continuous. Distor-



(a) (b)

(c) (d)

FIG. 3.1. *Characteristic curves of R and P as functions of ξ, for varying τ. Left figures: $r_c = 1.4$, $k = 0.36$. Right figures: $r_c = 0.88$, $k = 0.18$. Blue curves: integration from large R-values above the critical layer toward the critical layer. Red curves: integration from below. Particularly note the cusp-like behavior near $R = 0$ at $\xi = \pm\pi$.*

tions in the mean flow, as well as thin viscous layers, would be needed to completely eliminate these discontinuities. This is not surprising in view of previous research on nonlinear critical layers in stratified shear flows. Both Haberman [21], in the small Richardson number case, and Troitskaya [22], who studied the forced wave problem with $J_c > 1/4$, found that a temperature jump took place across the critical layer and the vorticity was also discontinuous. The need for these mean flow corrections in our Kelvin wave problem is clear in Figure 3.2(b), where the axial velocity component W is illustrated and it can be seen that W is discontinuous at the corners $(R, \xi) = (0, \pm\pi)$.

While such a vortex sheet is permitted in an inviscid problem, mean flow distortions are required to make this a valid solution in the sense of being the limit of a viscous flow as $Re \to \infty$. A comprehensive study of the $\lambda \sim O(1)$ problem would be required with consideration given to the limit as $\lambda \to 0$. Although we have not

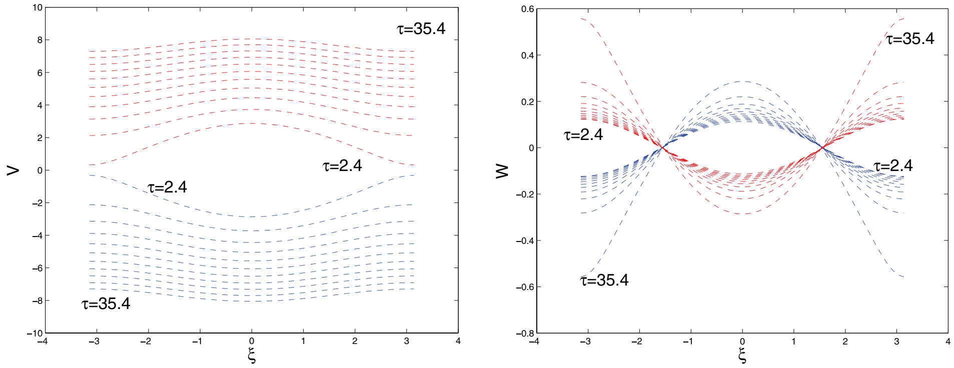


Fig. 3.2. *Shown are curves of V and W as functions of ξ for different τ-steps. Left: the V-component of velocity. Right: the W-component of velocity. For both, $k = 0.36$, $r_c = 1.4$.*

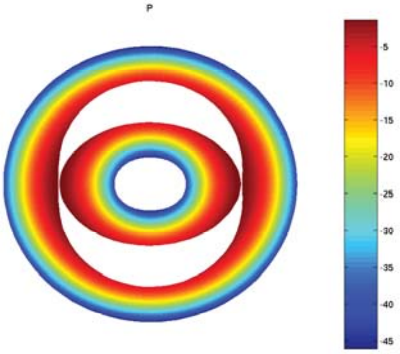
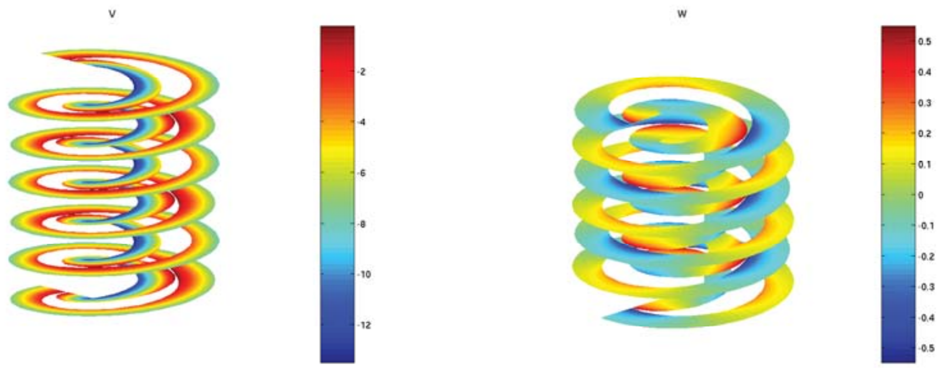


Fig. 3.3. *The characteristic solution for the pressure P, in (r, ξ) coordinates. Here $\xi \in [-\pi, \pi]$. Recall $R = \frac{r-r_c}{\epsilon^\beta}$; in this plot, $R \in [-8.29, 8.29]$. We have scaled the radial variable to enable easier reading of the graph. Note that P is continuous as we move across the critical layer.*

yet completed such an analysis, we can still anticipate to some extent the changes in mean flow based on the results in [21] and [22] and the small vorticity analysis of Caillol and Maslowe [19]. Before discussing these distortions further, however, let us proceed to the analysis of the closed characteristics region because related issues arise there.

3.4. The closed characteristics region. For a steady inviscid two-dimensional flow, a general solution of the Euler equations is $\nabla^2\psi = f(\psi)$, where ψ is the streamfunction and the function f is arbitrary. By considering a small viscous perturbation to the basic flow, Batchelor [23] proved that $\nabla^2\psi$, the vorticity, must be a constant within a region of closed streamlines. The value of the constant can be determined only by matching to the solution outside the separatrix. This is the counterpart of a difficulty here, namely, that the functions $F(\tau)$ and $G(\tau)$ are no longer determined by the outer region. We follow a procedure comparable to what is done in the case of a plane parallel flow, but the three-dimensionality of the present problem naturally adds complications.

To begin, we can exploit the analogy between rotating and stratified flows. For a stratified flow, Grimshaw [24] extended Batchelor's result by showing that the temper-



(a) $V(r, \theta, z)$

(b) $W(r, \theta, z)$

FIG. 3.4. *V and W as functions of r, θ, and z for a fixed t. Recall that P, V, W are periodic in $\xi = m\theta + kz - \omega t$. The spirals correspond to the surfaces $(10(r_c + R(\xi)), \theta, z)$ for varying τ (the radial variable is scaled to improve readability). The colors on the surfaces indicate the values of V and W.*

ature is constant within a region of closed streamlines. This is intuitively reasonable because both the vorticity and the thermal energy equation are diffusion equations, for the vorticity and heat, respectively. If we compare the first of equations (3.6) with the vertical momentum equation for a stratified flow, it can be seen that the azimuthal velocity V is the equivalent of the temperature. In both cases, the pressure gradient balances some force, the buoyancy force or the linearized centrifugal force. We will therefore assume that the axial component of vorticity $\frac{\partial V}{\partial R} + \bar{\Omega}_c$ is constant in the region of closed flow. This was proved in [19] for the small vorticity case by projecting onto a plane $z = const.$, and it must be nearly true in general. Integrating now with respect to R, we obtain the following expression for $V(R, \xi)$ inside the separatrices:

$$V(R, \xi) = (\bar{\Omega}_c + G_0)\, R + g(\xi), \tag{3.19}$$

where G_0 is the axial component of vorticity in the original frame of reference. We fix G_0 by matching the velocity V at $\xi = 0$ and determine $g(\xi)$ by matching V along the separatrices. It develops that we must take $g(\xi) = 0$ in order to preserve the symmetry in our solution.

The determination of $W(\xi, R)$ is less clear cut. As an approximation, we will again assume that the vorticity, this time the azimuthal component, is constant within the separatrices. In the critical layer, to lowest order, the azimuthal vorticity is given by $-\frac{\partial W}{\partial R}$, and, as a result, we obtain

$$W(R, \xi) = H_0\, R + h(\xi), \tag{3.20}$$

where H_0 and $h(\xi)$ are determined in the same way as their counterparts governing V and again symmetry requires $h(\xi) = 0$.

Returning now to the question of changes in the mean flow as the critical layer is crossed, to account for these we would include $O(\varepsilon^{\beta})$ mean flow components $\bar{V}_1(r)$ and $\bar{W}_1(r)$, say, in equations (2.2). These would then be expanded in Taylor series about the critical point r_c, and additional terms would appear in the matching conditions for V and W in (3.7). We included such terms in order to investigate their effect on a trial and error basis, and, as expected from earlier studies, including these mean flow jumps does remove some of the discontinuities. Viscosity, however, is still required to smooth out derivatives in velocity that appear in the equations defining the vorticity. The results of these experiments are not presented here because the procedure is not rigorous. It is nonetheless worth mentioning that a sign change in $\frac{\partial W}{\partial R}$ seems necessary to eliminate the discontinuity in W at the corners $(R, \xi) = (0, \pm\pi)$. It turned out, surprisingly, that introducing a jump in the mean flow vorticity gradient $Q(r)$ was the most effective way to accomplish this; i.e., as a consequence of nonlinearity, a change in $\bar{V}$ can significantly modify the behavior of W.

4. The eigenvalue problem. We now outline the procedure for solving (2.5) numerically for neutral modes with no phase change across the critical point. Results will be presented for the Lamb–Oseen vortex profile $\bar{V}(r) = (1 - e^{-r^2})/r$. The range of integration is from $r = 0$ to $r \to \infty$, and, because (2.5) has a regular singular point at $r = 0$ and an irregular singular point at infinity, series solutions are required at both ends. A Runge–Kutta method was used to carry out the integration.

Near the origin, the solution can be represented by a Frobenius expansion having the form

$$u = u_0\, r^{|m|-1}\, [1 + \zeta_1 r^2 + O(r^4)],$$

so that the radial perturbation velocity vanishes at the center of the vortex except in the case $m = 1$, the so-called bending mode, and then is continuous. Far from the center of the vortex, the velocity profile can be approximated as a potential vortex so that $\bar{V} \sim r^{-1}$. For a potential vortex, the pressure perturbation satisfies a modified Bessel equation (see pp. 341–342 of [1]) so that the pressure can be approximated using the asymptotic expansion for K_m, the solution that decays exponentially. To determine an expression for u valid for large r, the asymptotic result for p can then be substituted into the first two equations of the system (2.3) to obtain

$$u \sim u_\infty \frac{e^{-kr}}{\sqrt{kr}} \left(1 + \frac{\kappa_1}{kr} + \frac{\kappa_2}{k^2 r^2} + O[(kr)^{-3}]\right).$$

Near the critical layer, we employ a linear combination of the two Frobenius solutions, as in (2.6) and (2.7). Using the above conditions to initiate the integration, we integrate toward the critical layer from either side. All variables are real, and if we let $2\,\eta = |r - r_c|$ and choose $B = 1$ as the arbitrary constant, then as the critical layer is approached from the vortex center, we write

$$u(\eta^-) = A\, u_+(\eta^-) + u_-(\eta^-) \quad \text{or else} \quad u(\eta^+) = A\, u_+(\eta^+) + u_-(\eta^+)$$

if the critical layer is approached from outside. Requiring u'/u, as well as the constant A to be the same on either side of the critical layer, gives us enough conditions to determine the constants u_0 and u_∞, as well as the dispersion relation $\omega(k)$ for a given value of m.

In Figure 4.1, a dispersion curve is illustrated for the bending mode $m = 1$. The local Richardson number at the critical point is also shown. The bending mode is the most important, and it is clear from Figures 5 and 11 of Leweke and Williamson [25] that this is the mode that arises naturally in their experiments. Because the $m = 1$ mode essentially disappears in the $k = 0$ limit treated in [14] and [15], the importance of generalizing the theory to helical modes is evident. The solutions that we obtained

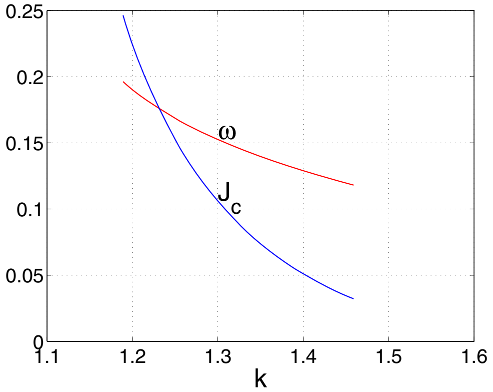


Fig. 4.1. *The dispersion relation $\omega(k)$ and Richardson number at the critical point defined in (2.8) for the $m = 1$ bending mode.*

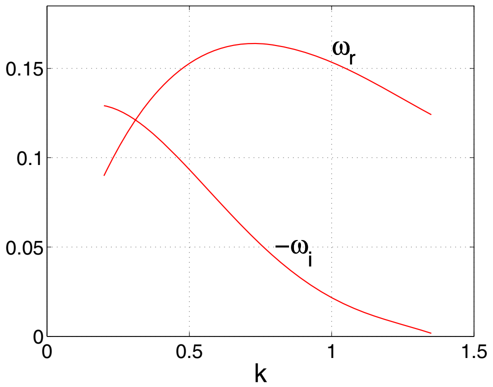


Fig. 4.2. *Variation of the frequency and decay rate ω_i as a function of wavenumber according to linear theory for a Lamb–Oseen vortex $\bar{V}(r) = \left(1 - \exp(-r^2)\right)/r$.*

for $m \geq 2$ had very large values of k, never being smaller than 7.80. Given that short waves are damped by viscosity this is likely the reason that they are not observed in the experiments.

To compare the results when there is a phase change with those in Figure 4.1, we have written a program that avoids the singularity in (2.5) by indenting the contour of integration. Because γ_c' is negative, the integration path passes above the singularity in the complex r plane, corresponding to the viscous limit as $Re \to \infty$ or to the initial value problem as $t \to \infty$. The damping rate is quite large for long waves, but it is small for $k \geq 1.20$. This may mean that it takes very little nonlinearity to generate a neutral mode if the wavelength is not long. The reason that the frequencies in Figures 4.1 and 4.2 are not far apart for $k \geq 1.4$ is that $r_c \geq 2.8$ and the singularity is weak that far from the center of the vortex. It is practically a potential vortex there, so how the singularity is crossed has little effect on the frequency for a given wavelength.

5. Conclusions. The differential equation governing the eigenvalue problem for helical waves propagating on a vortex has a critical point singularity if for some value of r the frequency $\omega = m\bar{\Omega}(r)$, where $\bar{\Omega}(r)$ is the angular velocity of the vortex. Our paper treats the class of waves for which this condition is satisfied and for which ε, a dimensionless amplitude parameter, is small enough so that linear theory is a good approximation outside a thin critical layer. At high Reynolds numbers, however, we have shown that nonlinear effects cannot be neglected in this layer even if ε is very small. Whether nonlinearity or viscosity is dominant depends on the parameter $\lambda = 1/(Re\,\varepsilon^{3\beta})$, where $1/2 > \beta > 2/3$. It can be recognized that $\lambda^{1/3}$ is the ratio of the viscous to the nonlinear critical layer thickness, so the conditions for validity of the classical viscous theory are not only $\varepsilon \ll 1$ but $\varepsilon^{\beta} \ll Re^{-1}$ as well. This limits the role of the linear viscous theory to laboratory experiments, where the Reynolds number is much smaller than in such applications as aircraft trailing vortices.

In our analysis of the nonlinear critical layer, the similarities were noted between the helical modes on a vortex and the propagation of nonlinear waves in a stratified shear flow. This enabled us to benefit from the knowledge gained in studies of the latter. In particular, we know that the scaling and details of the matching depend very much on the value of J_c, the local Richardson number at the critical point (i.e., the equivalent local Richardson number defined in (2.8) above). There are three different regimes corresponding to J_c greater than $1/4$, J_c less than $1/4$, and finally $J_c \sim O(\varepsilon^{1/2})$. Even for the stratified shear flow critical layer, however, questions remain to be answered about the mean flow distortion because only the cases of $J_c \sim O(\varepsilon^{1/2})$ [21] and $J_c > 1/4$ [22] have been treated.

The system of four coupled PDEs that govern the nonlinear critical layer was derived in section 3, and these equations are the same in all three regimes. The matching conditions, however, are different in the three cases. We have focused primarily on the Richardson number less than $1/4$ case, but for some vortex profiles the case $J_c > 1/4$ may also be of interest, particularly if there is some external forcing. An analytical solution of the inviscid governing equations was found by the method of characteristics in section 3.1, and this solution shows that the problem is highly nonlinear. One measure of the nonlinearity is that within the critical layer, all the higher harmonics are the same order of magnitude as the fundamental perturbation mode. Another is that there are discontinuities in the mean flow, particularly in the axial velocity induced by the wave. While vortex sheets are often employed as models in inviscid fluid dynamics, in real flows a thin shear layer is present in which viscosity is required to smooth out the discontinuity.

An analysis of the case where Q_c, the vorticity, is small at the critical point [19] confirmed that both the azimuthal and axial vorticity components are different on either side of the critical layer. The complexity of that analysis made it clear that a numerical solution of equations (3.6) including viscosity, i.e., with $\lambda \sim O(1)$, would be desirable in order to avoid having to deal with higher order terms in both the inner and outer expansions. We have initiated such a study and have been successful in obtaining solutions of (3.6) for moderate values of λ. Dealing with small values, however, is a computational challenge that will be overcome only after we have devised a way to solve (3.6) and (3.7) as a boundary value problem. That means finding a way to impose the asymptotic behavior below the critical layer without knowing the phase change in advance, as was done by Haberman [26] for the unstratified parallel shear flow, where the singularity, being logarithmic, is not as strong.

To conclude, we address the question of observability of the nonlinear waves described in this paper, whether in the atmosphere, the laboratory, or in numerical simulations. The mathematical similarities to the corresponding stratified shear flow problem make this the obvious place to look for some idea of what might be expected. Stratified shear flows with nonlinear critical layers have a structure resembling radar observations of what meteorologists call Kelvin–Helmholtz billows (see section 3 of [20]). Despite the fact that these billows contain localized turbulent layers, the large scale coherence of the wave is maintained. However, in the laboratory it has not yet been possible to achieve large enough Reynolds numbers to compare with the theory. As a consequence, its greatest utility has been in numerical simulations, where structural details first revealed by the nonlinear critical layer theory appeared several years later in computational work (see section 5 of [12] and section 4.6 of [27]).

Despite the mathematical analogies, the physical context is sufficiently different that it is not clear to what extent experience with stratified shear flows can be extrapolated to the trailing vortex problem. In the latter case, even though the flow is often laminar in the vortex core, its environment and upstream history are such that it is likely to be turbulent elsewhere in an aircraft wake. Of course, nonlinear waves play a role in some turbulence theories, but it would be difficult to identify a nonlinear wave in observations because of the danger in making detailed measurements. Numerical simulations are a possibility, but the Reynolds numbers are too low in computations reported to date. It would appear, therefore, that experiments offer the most promise. The paper by Delisi and Robins [28] reports an investigation to determine the effect of stratification on the trajectory of a pair of vortices. There is a table in their paper giving the Reynolds numbers for experiments reported in several papers. These are all large enough so that with only a slight forcing, a wave could be generated satisfying the requirements of our theory, and such experiments are currently under consideration. Suppose we consider a small perturbation with $\varepsilon = 0.03$ at the value $Re = 2.5 \cdot 10^4$ of the experiments in [28]. The calculation is not sensitive to the value of β; using $J_c = 0.09$ to compute β, we find that $\lambda = 0.01$, which is clearly in the nonlinear critical layer regime.

Appendix: Variation of the Reynolds stress. It is well known in hydrodynamic stability theory that valuable information about neutral modes can be obtained by evaluating the Reynolds stress for an unstable perturbation and then taking the limit as the growth rate tends to zero. From the energy equation for an unstable perturbation, as discussed in the review article by Stuart [29], it can be seen that the energy exchange between the mean flow and the perturbation is given by the integral

$$E = -\int_0^{r_1} \bar{\Omega}'(r)\, \overline{U_r U_\theta}\, r^2\, dr,$$

where the Reynolds stress $\tau = -\overline{U_r U_\theta}$. In this appendix we use capital letters for the velocity fluctuations, as in most texts on turbulence. The overline signifies an average over one wavelength in ξ, and we write the radial perturbation velocity as

$$U_r = \hat{U}_r\,(r)\, e^{i\,\xi} + \hat{U}_r^*\,(r)\, e^{-\,i\,\xi},$$

where $*$ indicates the complex conjugate.

The complex amplitude $\hat{U}_r$ is a combination of the two Frobenius solutions

$$\hat{U}_r = AX^+(r) + BX^-(r),$$

where X^+ and X^- are identical to the series $u_\pm$ in (2.6) and (2.7) above. The reason for changing our notation is to agree with that used by Miles [18] in order to permit comparison with his very similar development for stratified shear flows. One small difference is that, owing to the cylindrical geometry, the lower limit of our domain is $r = 0$.

In order to compute τ, the Reynolds stress, we need an expression for $\hat{U}_\theta\,(r)$. The system (2.3) yields a simple relationship between $\hat{U}_\theta\,(r)$ and $\hat{U}_r\,(r)$, namely,

$$\hat{U}_\theta = S\left[k^2\frac{Q}{\gamma} + \frac{m}{r}\,D_*\right]\hat{U}_r.$$

Using this expression now for $\hat{U}_\theta$, the Reynolds stress can be expressed in terms of the Frobenius solutions as follows:

$$\begin{aligned}\tau &= 2S(r)\frac{m}{r}\,\mathrm{Im}\Big[\hat{U}_r'\hat{U}_r^*\Big]\\ &= 2S(r)\frac{m}{r}\,\mathrm{Im}\Big[|A|^2X^{*+}X^{'+} + |B|^2X^{*-}X^{'-} + A^*BX^{*+}X^{'-} + AB^*X^{*-}X^{'+}\Big].\end{aligned}$$

The differential equation satisfied by the Reynolds stress is

$$\begin{aligned}\frac{d\tau}{dr} &= 2mD\left(\frac{S(r)}{r}\right)\mathrm{Im}(\hat{U}_r'\hat{U}_r^*) - 2\frac{m}{r}S(r)\,\mathrm{Im}\Big[\hat{U}_r''\hat{U}_r^*\Big] = -4m\frac{S(r)}{r^2}\,\mathrm{Im}\Big[\hat{U}_r'\hat{U}_r^*\Big]\\ &= -2\,\frac{\tau}{r}, \quad \text{which is readily integrated to obtain} \quad \tau(r) = \frac{\tau_0}{r^2}.\end{aligned}$$

The r^{-2} behavior of τ contrasts with the case of a parallel shear flow, where $\tau =$ *const.*; however, in both cases we must consider the possibility of the constant being discontinuous across r_c, as in the case of the Blasius boundary layer, for example.

For the vortices that are of primary interest in this paper, $\tau_0 = 0$ for $r < r_c$, because the Reynolds stress must be finite at $r = 0$. And, with the exception of the limiting case $J_c \to 0$, the Frobenius solutions (2.7) show that $\tau = 0$ at the critical point. Therefore, we must also have $\tau_0 = 0$ for $r > r_c$. If J_c is less than 1/4, the Reynolds stress near the critical point is given by

$$-\overline{U_rU_\theta} = 2\nu m r_c\frac{S(r_c)}{r^2}\,\mathrm{Im}[AB^*], \qquad r > r_c,$$

$$\text{and} \qquad -\overline{U_rU_\theta} = 2\nu m r_c\frac{S(r_c)}{r^2}\,\mathrm{Im}[AB^*\,e^{-i\pi(1+\nu)}], \qquad r < r_c.$$

It follows that either $A = 0$ or $B = 0$, as was shown by Miles for stratified shear flows.

For $J_c > 1/4$, on the other hand, the corresponding expressions are

$$-\overline{U_rU_\theta} = -\mu m r_c\frac{S(r_c)}{r^2}(|B|^2 - |A|^2), \qquad r > r_c,$$

$$\text{and} \qquad -\overline{U_rU_\theta} = -\mu m r_c\frac{S(r_c)}{r^2}(|A|^2e^{\pi\mu} - |B|^2e^{-\pi\mu}), \qquad r < r_c,$$

where $\mu = (4\,J_c - 1)^{1/2}$. The stress is zero when $|B| = |A|\,e^{\pi\mu}$. Finally, if there is no phase change, as is the case when the critical layer is nonlinear, the constants A and B are real, so a neutral mode can be a linear combination of both Frobenius solutions.

Acknowledgments. The authors are indebted to Stéphane LeDizès for suggesting this investigation and to Philippe Caillol and Dominique Perrault-Joncas for assistance in carrying out parts of the analysis and computation. Thomas Leweke has been most helpful in sharing with us his expertise on aircraft trailing vortices.

REFERENCES

[1] R. L. Ash and M. R. Khorrami, *Vortex stability*, in Fluid Vortices, S. I. Green, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 317–372.

[2] P. G. Saffman, *Vortex Dynamics*, Cambridge University Press, New York, 1992, Chapter 12.

[3] S. C. Crow, *Stability theory for a pair of trailing vortices*, AIAA J., 8 (1970), pp. 2172–2179.

[4] D. W. Moore and P. G. Saffman, *The instability of a straight vortex filament in a strain field*, Proc. Roy. Soc. London Ser. A, 346 (1975), pp. 413–425.

[5] C.-Y. Tsai and S. E. Widnall, *The stability of short waves on a straight vortex filament in a weak externally imposed strain field*, J. Fluid Mech., 73 (1976), pp. 721–733.

[6] P. R. Spalart, *Airplane trailing vortices*, in Annual Review of Fluid Mechanics, Vol. 30, Annu. Rev. Fluid Mech. 30, Annual Reviews, Palo Alto, CA, 1998, pp. 107–138.

[7] D. Sipp and A. L. Jacquin, *Widnall instabilities in vortex pairs*, Phys. Fluids, 15 (2003), pp. 1861–1874.

[8] T. Gerz and T. Ehret, *Wingtip vortices and exhaust jets during the jet regime of aircraft wakes*, Aerospace Sci. Techn., 7 (1997), pp. 463–474.

[9] V. V. Golubev and H. M. Atassi, *Sound propagation in an annular duct with mean potential swirling flow*, J. Sound Vibration, 198 (1996), pp. 601–616.

[10] Y. Chen, G. Brunet, and M. K. Yau, *Spiral bands in a simulated hurricane. Part* II: *Wave activity diagnostics*, J. Atmospheric Sci., 60 (2003), pp. 1239–1256.

[11] J. W. Miles, *On the stability of heterogeneous shear flows. Part* 2, J. Fluid Mech., 16 (1963), pp. 209–227.

[12] S. A. Maslowe, *Critical layers in shear flows*, in Annual Review of Fluid Mechanics, Vol. 18, Annu. Rev. Fluid Mech. 18, Annual Reviews, Palo Alto, CA, 1986, pp. 405–432.

[13] T. Maxworthy, E. J. Hopfinger, and L. G. Redekopp, *Wave motions on vortex cores*, J. Fluid Mech., 151 (1985), pp. 141–165.

[14] S. Le Dizès, *Non-axisymmetric vortices in two-dimensional flows*, J. Fluid Mech., 406 (2000), pp. 175–198.

[15] N. J. Balmforth, S. G. Smith, and W. R. Young, *Disturbing vortices*, J. Fluid Mech., 426 (2001), pp. 95–133.

[16] L. N. Howard and A. S. Gupta, *On the hydrodynamic and hydromagnetic stability of swirling flows*, J. Fluid Mech., 14 (1962), pp. 463–476.

[17] L. N. Howard, *Note on a paper of John W. Miles*, J. Fluid Mech., 10 (1961), pp. 509–512.

[18] J. W. Miles, *On the stability of heterogeneous shear flows*, J. Fluid Mech., 10 (1961), pp. 496–508.

[19] P. Caillol and S. A. Maslowe, *The small vorticity nonlinear critical layer for Kelvin modes on a vortex*, Stud. Appl. Math., 118 (2007), pp. 221–254.

[20] S. A. Maslowe, *The generation of clear air turbulence by nonlinear waves*, Stud. Appl. Math., 51 (1972), pp. 1–16.

[21] R. Haberman, *Wave-induced distortions of a slightly stratified shear flow: A nonlinear critical-layer effect*, J. Fluid Mech., 58 (1973), pp. 727–735.

[22] Yu. I. Troitskaya, *The viscous-diffusion nonlinear critical layer in a stratified shear flow*, J. Fluid Mech., 233 (1991), pp. 25–48.

[23] G. K. Batchelor, *On steady laminar flow with closed streamlines at large Reynolds number*, J. Fluid Mech., 1 (1956), pp. 177–190.

[24] R. H. J. Grimshaw, *On steady recirculating flows*, J. Fluid Mech., 39 (1969), pp. 695–703.

[25] T. Leweke and C. H. K. Williamson, *Cooperative elliptic instability of a vortex pair*, J. Fluid Mech., 360 (1998), pp. 85–119.

[26] R. Haberman, *Critical layers in parallel flows*, Stud. Appl. Math., 51 (1972), pp. 139–161.

[27] C. Staquet, *Two-dimensional secondary instabilities in a strongly stratified shear layer*, J. Fluid Mech., 296 (1995), pp. 73–126.

[28] D. P. Delisi and R. E. Robins, *Short-scale instabilities in trailing wake vortices in a stratified fluid*, AIAA J., 38 (2000), pp. 1916–1923.

[29] J. T. Stuart, *Hydrodynamic stability*, in Laminar Boundary Layers, L. Rosenhead, ed., Clarendon Press, Oxford, UK, 1963, pp. 492–579.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 844–868

PARTIALLY REFLECTED DIFFUSION*

A. SINGER[†], Z. SCHUSS[‡], A. OSIPOV[§], and D. HOLCMAN[¶]

Abstract. The radiation (reactive or Robin) boundary condition for the diffusion equation is widely used in chemical and biological applications to express reactive boundaries. The underlying trajectories of the diffusing particles are believed to be partially absorbed and partially reflected at the reactive boundary; however, the relation between the reaction constant in the Robin boundary condition and the reflection probability is not well defined. In this paper we define the partially reflected process as a limit of the Markovian jump process generated by the Euler scheme for the underlying Itô dynamics with partial boundary reflection. Trajectories that cross the boundary are terminated with probability $P\sqrt{\Delta t}$ and otherwise are reflected in a normal or oblique direction. We use boundary layer analysis of the corresponding master equation to resolve the nonuniform convergence of the probability density function of the numerical scheme to the solution of the Fokker–Planck equation in a half-space, with the Robin constant κ. The boundary layer equation is of the Wiener–Hopf type. We show that the Robin boundary condition is recovered if and only if trajectories are reflected in the conormal direction $\boldsymbol{\sigma n}$, where $\boldsymbol{\sigma}$ is the (possibly anisotropic) constant diffusion matrix and $\boldsymbol{n}$ is the unit normal to the boundary. Otherwise, the density satisfies an oblique derivative boundary condition. The constant κ is related to P by $\kappa = rP\sqrt{\sigma_n}$, where $r = 1/\sqrt{\pi}$ and $\sigma_n = \boldsymbol{n}^T\boldsymbol{\sigma n}$. The reflection law and the relation are new for diffusion in higher dimensions.

Key words. stochastic differential equations, reactive boundary condition, Markovian jump process, Wiener–Hopf boundary layer equation

AMS subject classifications. 60H35, 60J50, 81S40

DOI. 10.1137/060663258

1. Introduction.

The Fokker–Planck equation (FPE) with radiation (also called reactive or Robin) boundary conditions is widely used to describe diffusion in a biological cell with chemical reactions on its surface [1], [2], [3], [4], [5], [6], [7], [8], [9]. The Robin boundary conditions are used in [2], [4], [5], [6] as a homogenization of mixed Dirichlet–Neumann boundary conditions given on scattered small absorbing windows in an otherwise reflecting boundary. The mixed boundary conditions may represent, e.g., ligand binding or pumping out ions at sites on the boundary of a biological cell and no flux through the remaining boundary. The reactive rate constant in the Robin boundary conditions is chosen in the homogenization process so that the decay rate of the survival probability is the same as that in the mixed Dirichlet–Neumann boundary value problem.

The definition of the Itô stochastic dynamics

$$\dot{x} = a(x,t) + \sqrt{2\sigma(x,t)}\,\dot{w} \tag{1.1}$$

*Received by the editors June 19, 2006; accepted for publication (in revised form) September 5, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siap/68-3/66325.html

[†] Department of Mathematics, Program in Applied Mathematics, Yale University, 10 Hillhouse Ave., PO Box 208283, New Haven, CT 06520-8283 (amit.singer@yale.edu).

[‡] Department of Applied Mathematics, Tel-Aviv University, Ramat-Aviv 69978 Tel-Aviv, Israel (schuss@post.tau.ac.il). The research of this author was partially supported by a research grant from TAU.

[§] Institute of Mathematics, The Hebrew University, Jerusalem 91904, Israel (andreiosipov2@gmail.com).

[¶] Department of Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel (david.holcman@weizmann.ac.il), and Départment de Mathématiques et de Biologie, Ecole Normale Supérieure, 46 Rue d'Ulm, 75005 Paris, France. The research of this author was supported by the program "Chaire d'Excellence."

on the positive axis with total or partial reflection at the origin was given first by Feller [10] for the one-dimensional case with $a(x,t)$ and $\sigma(x,t)$ independent of t, as a limit of Itô processes, which are terminated when they reach the boundary or moved instantaneously to a point $x=\rho_j>0$ with probability p_j. When $p_j\to 1$ and $\rho_j\to 0$ with

$$\lim_{j\to\infty}\frac{1-p_j}{\rho_j}=c, \tag{1.2}$$

where c is a constant, the partially reflected process converges to a limit. The transition probability density function (pdf) of the limit process, $p(y,t\,|\,x,s)\,dy=\Pr\{x(t)\in(y,y+dy)\mid x(s)=x\}$, is the solution of the FPE

$$\frac{\partial p(y,t\,|\,x,s)}{\partial t}=-\frac{\partial[a(y,t)p(y,t\,|\,x,s)]}{\partial y}+\frac{\partial^2[\sigma(y,t)p(y,t\,|\,x,s)]}{\partial y^2} \tag{1.3}$$

or, equivalently,

$$\frac{\partial p(y,t\,|\,x,s)}{\partial t}=-\frac{\partial J(y,t\,|\,x,s)}{\partial y}\quad\text{for all}\quad y,x>0,$$

where

$$J(y,t\,|\,x,s)=a(y,t)p(y,t\,|\,x,s)-\frac{\partial[\sigma(y,t)p(y,t\,|\,x,s)]}{\partial y}, \tag{1.4}$$

is the flux. The initial condition is

$$p(y,t\,|\,x,s)\to\delta(y-x)\quad\text{as}\quad t\downarrow s, \tag{1.5}$$

and the radiation boundary condition is

$$-J(0,t\,|\,x,s)=\kappa p(0,t\,|\,x,s), \tag{1.6}$$

where κ is a constant related to the constant c and to the values of the coefficients at the boundary. The no flux and Dirichlet boundary conditions are recovered if $c=0$ or $c=\infty$, respectively. Feller's method does not translate into a Brownian dynamics simulation of the limit process, because his approximations are continuous-time Itô processes. Skorokhod [11] defines the reflection process inside the boundary. Several numerical schemes have been proposed for simulating this process (see, e.g., [11], [12], [13], [14]). The main issue there is to approximate the local time spent on the boundary.

The definition of a diffusion process with absorbing or reflecting boundaries as limits of Markovian jump processes, which is the basis for all simulations, gives in the limit diffusion processes with well-defined boundary behavior. However, the definition of a diffusion process with partially reflecting boundaries as a limit of Markovian jump processes gives different diffusions for different jump processes. This is expressed in different relations between the termination probability of the jump process and the boundary conditions for the FPEs (see, e.g., [8]). The process $x(t)$ defined by (1.1) with partially absorbing boundaries can be defined as the limit of the solutions of the Markovian jump processes generated by the Euler scheme

$$x_{\Delta t}(t+\Delta t)=x_{\Delta t}(t)+a(x_{\Delta t}(t),t)\Delta t+\sqrt{2\sigma(x_{\Delta t}(t),t)}\,\Delta w(t,\Delta t)\quad\text{for}\quad t\geq s, \tag{1.7}$$

$$x_{\Delta t}(s)=x \tag{1.8}$$

in the interval $x > 0$, for $0 \leq t - s \leq T$, with $\Delta t = T/N$, $t - s = iT/N$ ($i = 0, 1, \ldots, N$), where for each t the random variables $\Delta w(t, \Delta t)$ are normally distributed and independent with zero mean and variance Δt. The partially absorbing boundary condition for (1.7) has to be chosen so that the pdf $p_{\Delta t}(x, t)$ of $x_{\Delta t}(t)$ converges to the solution of (1.3)–(1.6). At a partially reflecting boundary for (1.7), the trajectories are reflected with probability (w.p.) R and otherwise terminated (absorbed), once they cross the origin. We show below that keeping R constant (e.g., $R = 1/2$) as $\Delta t \to 0$ leads to the convergence of the pdf $p_{\Delta t}(x, t)$ to the solution of the FPE with an absorbing rather than the Robin boundary condition. Thus the reflection probability R must increase to 1 as $\Delta t \to 0$ in order to yield the Robin condition (1.6). Moreover, the reactive constant κ is related to the limit

$$\lim_{\Delta t \to 0} \frac{1 - R}{\sqrt{\Delta t}} = P. \tag{1.9}$$

The reflecting boundary condition is recovered for $P = 0$, while the absorbing boundary condition is obtained for $P = \infty$. Motivated by these considerations, we design the following simple boundary behavior for the simulated trajectories that cross the boundary, identified by $x_{\Delta t}(t) + a(x_{\Delta t}(t), t)\Delta t + \sqrt{2\sigma(x_{\Delta t}(t), t)}\,\Delta w < 0$:

(1.10)
$$x_{\Delta t}(t + \Delta t) = \begin{cases} -(x_{\Delta t}(t) + a(x_{\Delta t}(t), t)\Delta t + \sqrt{2\sigma(x_{\Delta t}(t), t)}\,\Delta w) & \text{w.p. } 1 - P\sqrt{\Delta t}, \\ \text{terminate trajectory otherwise.} \end{cases}$$

The exiting trajectory is normally reflected w.p.

$$R = 1 - P\sqrt{\Delta t} \tag{1.11}$$

and is otherwise terminated (absorbed). The scaling of the termination probability with $\sqrt{\Delta t}$ reflects the fact that the discrete unidirectional diffusion current at any point, including the boundary, is $O(1/\sqrt{\Delta t})$ (see [15], [16]). This means that the number of discrete trajectories hitting or crossing the boundary in any finite time interval increases as $1/\sqrt{\Delta t}$. Therefore, to keep the efflux of trajectories finite as $\Delta t \to 0$, the termination probability of a crossing trajectory, $1 - R$, has to be $O(\sqrt{\Delta t})$. The pdf $p_{\Delta t}(x, t)$, however, does not converge to the solution $p(x, t)$ of (1.3)–(1.6) on the boundary, as discussed in section 2. This is due to the formation of a boundary layer, as is typical for diffusion approximations of Markovian jump processes that jump over the boundary [17], [18], [19]. The boundary layer equations are typically Wiener–Hopf integral equations. The Wiener–Hopf boundary layer equation for the particular case of a partially reflected Brownian motion on the positive axis (i.e., $a(x, t) = 0$ and $\sigma(x, t) = \sigma$ in (1.7)) was recently solved in [8], and the relationship $\kappa = P\sqrt{\sigma}/\sqrt{\pi}$ was found.

The convergence of the pdf of an Euler scheme has been studied in [20], [21] for the higher-dimensional problem with oblique reflection. Bounds on the integral norm of the approximation error are given for the solution of the backward Kolmogorov equation. These, however, do not resolve the boundary layer of the pdf of the numerical solution. The solution of the forward equation for the Euler scheme converges nonuniformly to the solution of the FPE due to the appearance of a boundary layer in the first order spatial derivative. This distorts the boundary flux and gives incorrect boundary conditions. A boundary layer expansion is needed to capture the boundary phenomena.

The derivation of the radiation condition has a long history. Collins and Kimball [22] (see also [23]) derived the radiation boundary condition (1.6) for the limit $p(x,t) = \lim_{\Delta t\to 0} p_{\Delta t}(x,t)$ from an underlying discrete random walk model on a semi-infinite one-dimensional lattice with partial absorption at the endpoint. Their model assumes constant diffusion coefficient and vanishing drift, for which they find the reactive constant in terms of the absorption probability and the diffusion coefficient. Previous simulation schemes that recover the Robin boundary condition [1], [24], [25], [26], [27] make use of the explicit solution to the half-space FPE with linear drift term and constant diffusion coefficient with a Robin condition. In [28] and the references therein, the specular reflection method near a reflecting boundary has been shown to be superior to other methods such as rejection, multiple rejection, and interruption.

An apparent paradox arises when using (1.7) and other schemes: while the pdf $p_{\Delta t}(y,t\,|\,x,s)$ of the solution of (1.7), (1.8), (1.10), (1.11) converges to the solution of the FPE (1.3) and the initial condition (1.5), each approximant $p_{\Delta t}(y,t\,|\,x,s)$ does not satisfy the boundary condition (1.6), not even approximately; that is, the error does not decay as $\Delta t \to 0$. For a general diffusion coefficient and drift term, the boundary condition is not satisfied even for the case of a reflecting boundary condition. This problem plagues other schemes as well. The apparent paradox is due to the nonuniform convergence of $p_{\Delta t}(y,t\,|\,x,s)$ to the solution $p(y,t\,|\,x,s)$ of the FPE, caused by a boundary layer in $p_{\Delta t}(y,t\,|\,x,s)$, as is typical of boundary behavior of diffusion approximations to Markovian jump processes. The limit $p(y,t\,|\,x,s)$, however, satisfies the boundary condition (1.6) for some κ. Our analysis can be extended to other schemes in a straightforward way. It is well known that the Euler scheme produces an $O(\sqrt{\Delta t})$ error in estimating the mean first passage time to reach an absorbing boundary. There are several recipes to reduce the discretization error to $O(\Delta t)$ [29], [30], [31], [32], [33]. Another manifestation of the boundary layer is that the approximation error of the pdf near absorbing or reflecting boundaries is $O(\sqrt{\Delta t})$, and some methods, including [1], [34], reduce this error to $O(\Delta t)$. Thus, we expect the formation of a boundary layer of size $O(\sqrt{\Delta t})$ for the Euler scheme (1.7) with the boundary behavior (1.10).

This paper is concerned with the convergence of the partially reflecting Markovian jump process generated by (1.7), (1.10) in one and higher dimensions. We show that this scheme, with the additional requirement that the pdf converges to the solution of the FPE with a given Robin boundary condition, defines a unique diffusion process with partial reflection at the boundary. This definition is then generalized to higher dimensions. In contrast to Collins and Kimball's [22] discrete scheme, this definition is not restricted to lattice points, and the drift and diffusion coefficients may vary. The advantage of the current suggested design (1.10) is its simplicity, which is both easily and efficiently implemented and amenable to analysis. There is no need to make any assumptions on the structure of the diffusion coefficient or the drift. From the theoretical point of view, it serves as a physical interpretation for the behavior of diffusive trajectories near a reactive boundary.

Our main result in the one-dimensional case is the relation between the reactive "constant" $\kappa(t)$ and the absorption parameter P for the dynamics (1.1) on the positive axis with drift and with a variable diffusion coefficient,

$$\kappa(t) = rP\sqrt{\sigma(0,t)}, \quad r = \frac{1}{\sqrt{\pi}}. \tag{1.12}$$

The relation (1.12) is new for diffusion with variable coefficients. The value $r = 1/\sqrt{\pi}$ is different from values obtained for other schemes, e.g., from the value $r = 1/\sqrt{2}$, predicted by the discrete random walk theory of radiation boundaries [22]. Values of r for other schemes are given in [8]. We show the effect of using (1.12) in numerical simulations.

The scheme (1.10) is generalized to diffusion with drift and anisotropic constant diffusion matrix $\boldsymbol{\sigma}(t)$ in the half-space, $x_1 > 0$, with partial oblique reflection. We show that the Robin boundary condition is recovered if and only if trajectories are reflected in the direction of the unit vector

$$\boldsymbol{v} = \frac{\boldsymbol{\sigma n}}{\|\boldsymbol{\sigma n}\|}, \tag{1.13}$$

where $\boldsymbol{n}$ is the unit normal to the boundary. The radiation parameter $\kappa(\boldsymbol{x}, t)$ in the d-dimensional Robin boundary condition and the absorption parameter $P(\boldsymbol{x})$ are related by

$$\kappa(\boldsymbol{x}, t) = rP(\boldsymbol{x})\sqrt{\sigma_n(t)}, \quad x_1 = 0, \tag{1.14}$$

with r given in (1.12) and $\sigma_n(t) = \boldsymbol{n}^T\boldsymbol{\sigma}(t)\boldsymbol{n}$. The relation (1.14) is new for higher-dimensional diffusion in a half-space with drift and anisotropic diffusion matrix.

In the most common case of constant isotropic diffusion our result extends to domains with curved boundaries. This is due to the fact that a smooth local mapping of the domain to a half-space with an orthogonal system of coordinates preserves the constant isotropic diffusion matrix, though the drift changes according to Itô's formula. In this case the vector $\boldsymbol{v}$ coincides with the normal $\boldsymbol{n}$.

2. Boundary layer analysis in one dimension.

The aim of the boundary layer analysis below is to examine the convergence of the pdf $p_{\Delta t}(y, t \,|\, x, s)$ of the solution $x_{\Delta t}(t)$ of (1.7), (1.8) to the solution $p(y, t \,|\, x, s)$ of (1.3)–(1.6), and to find the relation between the parameter P of (1.10) and the reactive constant κ in (1.6). Using abbreviated notation, the pdf $p_{\Delta t}(y, t \,|\, x, s) = p_{\Delta t}(y, t)$ satisfies the forward Kolmogorov equation [15], [16], [17], [18], [19], [35]

$$\begin{aligned} p_{\Delta t}(y, t + \Delta t) = \int_0^\infty \frac{p_{\Delta t}(x, t)}{\sqrt{4\pi\sigma(x, t)\Delta t}} &\left\{ \exp\left[-\frac{(y - x - a(x, t)\Delta t)^2}{4\sigma(x, t)\Delta t} \right] \right. \\ &\left. + (1 - P\sqrt{\Delta t}) \exp\left[-\frac{(y + x + a(x, t)\Delta t)^2}{4\sigma(x, t)\Delta t} \right] \right\} dx. \end{aligned} \tag{2.1}$$

For $P = 0$ the pdf $p_{\Delta t}(y, t)$ satisfies the boundary condition

$$\frac{\partial p_{\Delta t}(0, t)}{\partial y} = 0, \tag{2.2}$$

which is obtained by differentiation of (2.1) with respect to y at $y = 0$. If $P \neq 0$, we obtain

$$\frac{\partial p_{\Delta t}(0, t + \Delta t)}{\partial y} = \frac{p_{\Delta t}(0, t)P}{\sqrt{4\pi\sigma(0, t)}} + O(\sqrt{\Delta t}), \tag{2.3}$$

which holds also in the limit $\Delta t \to 0$. However, the order of the limits $\Delta t \to 0$ and $y \downarrow 0$ matters; indeed,

$$\lim_{\Delta t \to 0} \lim_{y \downarrow 0} \frac{\partial p_{\Delta t}(y, t)}{\partial y} \neq \lim_{y \downarrow 0} \lim_{\Delta t \to 0} \frac{\partial p_{\Delta t}(y, t)}{\partial y}. \tag{2.4}$$

The limit of (2.3) is not the boundary condition that the limit function $p(y,t) = \lim_{\Delta t\to 0} p_{\Delta t}(y,t)$ (for $y > 0$) satisfies. To find the boundary condition of $p(y,t)$, in either case $P = 0$ or $P \neq 0$, we show below that $p(y,t)$ satisfies the FPE (1.3) and the initial condition (1.5) for all $y > 0$. Since for $P = 0$ the simulation preserves probability (the population of trajectories),

$$0 = \frac{d}{dt}\int_0^\infty p(x,t)\,dx = -\frac{\partial[\sigma(0,t)p(0,t)]}{\partial y} + a(0,t)p(0,t) = J(0,t). \tag{2.5}$$

Equation (2.5) is the no flux boundary condition. The discrepancy between (2.5) and (2.2) is due to the nonuniform convergence of $p_{\Delta t}(y,t)$ to its limit $p(y,t)$ in the interval. There is a boundary layer of width $O(\sqrt{\Delta t})$, in which the boundary condition (2.2) for $p_{\Delta t}(y,t)$ changes into the boundary condition (2.5) that $p(y,t)$ satisfies. To analyze the discrepancy between (2.2) and (2.5), we introduce the local variable $y = \eta\sqrt{\Delta t}$ and the boundary layer solution

$$p_{BL}(\eta,t) = p_{\Delta t}(\eta\sqrt{\Delta t},t). \tag{2.6}$$

Changing variables $x = \xi\sqrt{\Delta t}$ in the integral (2.1) gives

$$\begin{aligned} p_{BL}(\eta,t+\Delta t) = \int_0^\infty \frac{p_{BL}(\xi,t)}{\sqrt{4\pi\sigma(\xi\sqrt{\Delta t},t)}} &\left\{ \exp\left[-\frac{\left(\eta-\xi-a(\xi\sqrt{\Delta t},t)\sqrt{\Delta t}\right)^2}{4\sigma(\xi\sqrt{\Delta t},t)}\right]\right. \\ &\left. + (1-P\sqrt{\Delta t})\exp\left[-\frac{\left(\eta+\xi+a(\xi\sqrt{\Delta t},t)\sqrt{\Delta t}\right)^2}{4\sigma(\xi\sqrt{\Delta t},t)}\right]\right\} d\xi. \end{aligned} \tag{2.7}$$

The boundary layer solution has an asymptotic expansion in powers of $\sqrt{\Delta t}$:

$$p_{BL}(\eta,t) \sim p_{BL}^{(0)}(\eta,t) + \sqrt{\Delta t}\,p_{BL}^{(1)}(\eta,t) + \Delta t\,p_{BL}^{(2)}(\eta,t) + \cdots. \tag{2.8}$$

Expanding all functions in (2.7) in powers of $\sqrt{\Delta t}$ and equating similar orders, we obtain integral equations that the asymptotic terms of (2.8) must satisfy. The leading order $O(1)$ term gives the Wiener–Hopf-type equation on the half-line

$$p_{BL}^{(0)}(\eta,t) = \int_0^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\left\{\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right] + \exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\} d\xi \tag{2.9}$$

for $\eta > 0$. The kernel

$$K(\eta,\xi) = \exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right] + \exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right] \tag{2.10}$$

is an even function of η and ξ; i.e., $K(\eta,\xi) = K(-\eta,\xi) = K(\eta,-\xi) = K(-\eta,-\xi)$. Therefore, we extend $p_{BL}^{(0)}(\xi,t)$ to the entire line as an even function ($p_{BL}^{(0)}(\xi,t) = p_{BL}^{(0)}(-\xi,t)$) and rewrite (2.9) as

$$p_{BL}^{(0)}(\eta,t) = \int_{-\infty}^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right] d\xi \tag{2.11}$$

for $-\infty < \eta < \infty$. The only solution of the integral equation (2.11) is the constant function, that is, $p^{(0)}_{BL}(\eta,t) = f(t)$, independent of η. This follows immediately from the Fourier transform of (2.11), whose right-hand side is a convolution.

Away from the boundary layer the solution admits an outer solution expansion

$$p_{OUT}(y,t) \sim p^{(0)}_{OUT}(y,t) + \sqrt{\Delta t} p^{(1)}_{OUT}(y,t) + \cdots , \tag{2.12}$$

where $p^{(0)}_{OUT}$ satisfies the Fokker–Planck equation (1.3) and the initial condition (1.5). Indeed, the integrals in (2.1) are of Laplace type with the small parameter Δt. For interior points $y \gg \sqrt{\Delta t}$, the second integral, which represents only boundary interactions, is negligible relative to the first. We change variables in (2.1) by setting

$$\eta = \frac{y - x - a(x,t)\Delta t}{\sqrt{2\sigma(x,t)\Delta t}},$$

and extend integration over the entire line in the first integral and expand all functions in powers of $\sqrt{\Delta t}$. The resulting integrals are moments of the normal distribution. We obtain

$$\frac{p_{\Delta t}(y,t+\Delta t) - p_{\Delta t}(y,t)}{\Delta t} = -\frac{\partial[a(y,t)p_{\Delta t}(y,t)]}{\partial y} + \frac{\partial^2[\sigma(y,t)p_{\Delta t}(y,t)]}{\partial y^2} + O(\sqrt{\Delta t}).$$

The leading term in the expansion of $p_{\Delta t}(y,t)$ is $p^{(0)}_{OUT}(y,t)$, which therefore satisfies the Fokker–Planck equation (1.3). The initial condition (1.5) is recovered from the Gaussian integral as $\Delta t \to 0$. The boundary condition that $p^{(0)}_{OUT}(y,t)$ satisfies can be determined only after the boundary layer is resolved by matching. The leading order matching condition of the boundary layer and the outer solutions is

$$\lim_{\eta\to\infty} p^{(0)}_{BL}(\eta,t) = p^{(0)}_{OUT}(0,t).$$

Therefore

$$p^{(0)}_{BL}(\eta,t) = p^{(0)}_{OUT}(0,t). \tag{2.13}$$

The matching condition at order $\sqrt{\Delta t}$ gives

$$\eta \frac{\partial p^{(0)}_{OUT}(0,t)}{\partial y} + p^{(1)}_{OUT}(0,t) \sim p^{(1)}_{BL}(\eta,t) \quad \text{for} \quad \eta \to \infty,$$

which means that $p^{(1)}_{BL}(\eta,t)$ is asymptotically a linear function of η; therefore the limit of its derivative is a constant. Thus the matching condition reduces to

$$\lim_{\eta\to\infty} \frac{\partial p^{(1)}_{BL}(\eta,t)}{\partial \eta} = \frac{\partial p^{(0)}_{OUT}(0,t)}{\partial y}. \tag{2.14}$$

The first order boundary layer term satisfies the integral equation

(2.15)

$$p^{(1)}_{BL}(\eta,t) = \int_0^\infty \frac{p^{(1)}_{BL}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}} \left\{ \exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right] + \exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right] \right\} d\xi$$

$$
\begin{aligned}
&- P\int_0^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]d\xi\\
&- \frac{\sigma_y(0,t)}{2\sigma(0,t)}\int_0^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\,\xi\left\{\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]+\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi\\
&+ \frac{\sigma_y(0,t)}{4\sigma(0,t)^2}\int_0^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\,\xi\left\{(\eta-\xi)^2\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]+(\eta+\xi)^2\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi\\
&+ \frac{2a(0,t)}{4\sigma(0,t)}\int_0^\infty \frac{p_{BL}^{(0)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\left\{(\eta-\xi)\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]-(\eta+\xi)\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi.
\end{aligned}
$$

Evaluating explicitly the last four integrals in (2.15) and using (2.13) gives

$$
\begin{aligned}
p_{BL}^{(1)}(\eta,t) = {}&\int_0^\infty \frac{p_{BL}^{(1)}(\xi,t)}{\sqrt{4\pi\sigma(0,t)}}\left\{\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]+\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi\\
&- \frac{P}{2}\,p_{OUT}^{(0)}(0,t)\operatorname{erfc}\left(\frac{\eta}{2\sqrt{\sigma(0,t)}}\right)\\
&+ \frac{\sigma_y(0,t)-a(0,t)}{\sqrt{\pi\sigma(0,t)}}p_{OUT}^{(0)}(0,t)\exp\left[-\frac{\eta^2}{4\sigma(0,t)}\right].
\end{aligned}
\tag{2.16}
$$

Differentiating (2.16) with respect to η and integrating by parts, we obtain

$$
\begin{aligned}
\frac{\partial p_{BL}^{(1)}(\eta,t)}{\partial\eta} = {}&\frac{1}{\sqrt{4\pi\sigma(0,t)}}\int_0^\infty \frac{\partial p_{BL}^{(1)}(\xi,t)}{\partial\eta}\left\{\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]-\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi\\
&+ \frac{P}{2\sqrt{\pi\sigma(0,t)}}\,p_{OUT}^{(0)}(0,t)\exp\left[\frac{-\eta^2}{4\sigma(0,t)}\right]-\frac{\sigma_y(0,t)-a(0,t)}{2\sqrt{\pi}\,\sigma(0,t)^{3/2}}p_{OUT}^{(0)}(0,t)\,\eta\exp\left[\frac{-\eta^2}{4\sigma(0,t)}\right].
\end{aligned}
\tag{2.17}
$$

Setting

$$
g(\eta,t)=\frac{\partial p_{BL}^{(1)}(\eta,t)}{\partial\eta}-\frac{P}{2\sqrt{\pi\sigma(0,t)}}\,p_{OUT}^{(0)}(0,t)\exp\left[-\frac{\eta^2}{4\sigma(0,t)}\right],
\tag{2.18}
$$

we rewrite (2.17) as

$$
g(\eta,t)=\phi(\eta,t)+\frac{1}{\sqrt{4\pi\sigma(0,t)}}\int_0^\infty g(\xi,t)\left\{\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]-\exp\left[-\frac{(\eta+\xi)^2}{4\sigma(0,t)}\right]\right\}d\xi,
\tag{2.19}
$$

where

$$
\begin{aligned}
\phi(\eta,t) = {}&\frac{P}{\sqrt{8\pi\sigma(0,t)}}\,p_{OUT}^{(0)}(0,t)\exp\left[\frac{-\eta^2}{8\sigma(0,t)}\right]\operatorname{erf}\left(\frac{\eta}{\sqrt{8\sigma(0,t)}}\right)\\
&- \frac{\sigma_y(0,t)-a(0,t)}{2\sqrt{\pi}\,\sigma(0,t)^{3/2}}p_{OUT}^{(0)}(0,t)\,\eta\exp\left[\frac{-\eta^2}{4\sigma(0,t)}\right].
\end{aligned}
\tag{2.20}
$$

Since $\phi(\eta,t)$ is an odd function of η, we can define $g(\eta,t)$ for negative values as an odd function by setting $g(\eta,t)=-g(-\eta,t)$ for $\eta<0$. Then (2.19) can be rewritten as

$$
g(\eta,t)=\phi(\eta,t)+\frac{1}{\sqrt{4\pi\sigma(0,t)}}\int_{-\infty}^\infty g(\xi,t)\exp\left[-\frac{(\eta-\xi)^2}{4\sigma(0,t)}\right]d\xi,
\tag{2.21}
$$

which in Fourier space is

$$\hat{g}(k,t) = \frac{\hat{\phi}(k,t)}{1-\exp[-\sigma(0,t)k^2]}. \tag{2.22}$$

Using the Wiener–Hopf method, we decompose

$$\hat{g}(k,t) = \hat{g}_+(k,t) + \hat{g}_-(k,t), \tag{2.23}$$

where $g_+(\eta) = g(\eta)\chi_{[0,\infty)}(\eta)$, $g_-(\eta) = g(\eta)\chi_{(-\infty,0]}(\eta)$. The Fourier transform $\hat{g}(k,t)$ exists in the sense of distributions, and $\hat{g}_\pm(k,t)$ are analytic in the upper and lower halves of the complex plane, respectively. Taylor's expansion of $\hat{\phi}(k,t)$ in (2.20) gives

$$\hat{\phi}(k,t) = 2ip^{(0)}_{OUT}(0,t)\left\{\frac{P\sqrt{\sigma(0,t)}}{\sqrt{\pi}} - [\sigma_y(0,t) - a(0,t)]\right\}k + O(k^3) \quad \text{as} \quad k \to 0. \tag{2.24}$$

The nonzero poles of (2.22) split evenly between $\hat{g}_+(k,t)$ and $\hat{g}_-(k,t)$, and using $\hat{g}_+(k,t) = -\hat{g}_-(-k,t)$, the pole at the origin gives

$$\hat{g}_+(k,t) = ip^{(0)}_{OUT}(0,t)\left\{\frac{P}{\sqrt{\pi\sigma(0,t)}} - \frac{\sigma_y(0,t) - a(0,t)}{\sigma(0,t)}\right\}\frac{1}{k} + O(k) \quad \text{as} \quad k \to 0. \tag{2.25}$$

Inverting the Fourier transform $\hat{g}_+(k,t)$, by closing the contour of integration around the lower half-plane, we obtain

$$\lim_{\eta\to\infty} \frac{\partial p^{(1)}_{BL}(\eta,t)}{\partial\eta} = p^{(0)}_{OUT}(0,t)\left\{\frac{P}{\sqrt{\pi\sigma(0,t)}} - \frac{\sigma_y(0,t) - a(0,t)}{\sigma(0,t)}\right\}. \tag{2.26}$$

The matching condition (2.14) implies

$$\frac{\partial p^{(0)}_{OUT}(0,t)}{\partial y} = p^{(0)}_{OUT}(0,t)\left\{\frac{P}{\sqrt{\pi\sigma(0,t)}} - \frac{\sigma_y(0,t) - a(0,t)}{\sigma(0,t)}\right\}. \tag{2.27}$$

Multiplying by $\sigma(0,t)$ and rearranging, we obtain the radiation boundary condition

$$-J(0,t) = \frac{\partial}{\partial y}\left[\sigma(0,t)p^{(0)}_{OUT}(0,t)\right] - a(0,t)p^{(0)}_{OUT}(0,t) = \frac{P\sqrt{\sigma(0,t)}}{\sqrt{\pi}}p^{(0)}_{OUT}(0,t). \tag{2.28}$$

Since $p(y,t) = p^{(0)}_{OUT}(y,t)$, the reactive "constant" in (1.6) is

$$\kappa(t) = \frac{P\sqrt{\sigma(0,t)}}{\sqrt{\pi}}. \tag{2.29}$$

3. Numerical simulations in one dimension.

The explicit analytical solution of the FPE (1.3) with the initial condition (1.5) and the radiation boundary condition (1.6) for the case of vanishing drift ($a = 0$) and constant diffusion coefficient ($\sigma(x,t) = \sigma$) was first given by Bryan in 1891 [36] (see [37, sect. 14.2, p. 358]):

$$\begin{aligned} p(x,t\,|\,x_0) = {}& \frac{1}{\sqrt{4\pi\sigma t}}\left[\exp\left\{-\frac{(x-x_0)^2}{4\sigma t}\right\} + \exp\left\{-\frac{(x+x_0)^2}{4\sigma t}\right\}\right] \\ & - \frac{\kappa}{\sigma}\exp\left\{\frac{\kappa(x+x_0+\kappa t)}{\sigma}\right\}\operatorname{erfc}\left[\frac{x+x_0+2\kappa t}{\sqrt{4\sigma t}}\right]. \end{aligned} \tag{3.1}$$

The first term in (3.1) is the fundamental solution of (1.3) and (1.5) with a reflecting boundary condition, whereas the second term may be transformed into

$$-\frac{\kappa}{\sqrt{\pi\sigma^3 t}}\int_0^\infty \exp\left\{-\frac{\kappa\xi}{\sigma}\right\}\exp\left\{-\frac{(x+x_0+\xi)^2}{4\sigma t}\right\}d\xi,$$

which represents the density due to a line of exponentially decreasing sinks extending from $-x_0$ to $-\infty$. The method of Laplace transforming (1.3) with respect to t was later employed [1], [38] to obtain explicit analytical solution for the FPE (1.3)–(1.5) with a constant diffusion coefficient and a (not necessarily vanishing) constant drift term $a(x,t)=a$:

$$\begin{aligned} &p(x,t\,|\,x_0)\\ &\quad=\frac{1}{\sqrt{4\pi\sigma t}}\left[\exp\left\{-\frac{(x-x_0-at)^2}{4\sigma t}\right\}+\exp\left\{-\frac{ax_0}{\sigma}-\frac{(x+x_0-at)^2}{4\sigma t}\right\}\right]\\ &\qquad-\frac{2\kappa+a}{2\sigma}\exp\left\{\frac{ax+\kappa[x+x_0+(\kappa+a)t]}{\sigma}\right\}\operatorname{erfc}\left[\frac{x+x_0+(2\kappa+a)t}{\sqrt{4\sigma t}}\right]. \end{aligned} \tag{3.2}$$

Setting $\kappa=0$ in (3.2) reduces to Smoluchowski's [39] explicit analytical solution for a reflecting boundary with a constant drift term, while setting $a=0$ reduces to Bryan's solution (3.1).

We conducted several numerical experiments in which $n=10^7$ trajectories were simulated according to the Euler scheme (1.7) with the boundary behavior (1.10). The diffusion coefficient was constant $\sigma=1$, and the reactive constant was $\kappa=1$, giving $P=\sqrt{\pi}$ in (2.29). The trajectories were initially located at $x_0=1$, and their statistics were collected at time $t=1$ and compared to the predicted $p(x,t=1\,|\,x_0=1)$. The convergence of the scheme was tested by using four different time steps, $\Delta t=10^{-1},10^{-2},10^{-3},10^{-4}$.

The first experiment corresponds to a vanishing drift $a=0$. Figure 1 shows the convergence of the numerical scheme to the analytic solution (3.1). The rate of convergence of the numerical scheme to the analytic solution is $\sqrt{\Delta t}$. This is demonstrated, for example, by the survival probability

$$p_{sur}(x_0,t)=\int_0^\infty p(x,t\,|\,x_0)\,dx$$

of finding the trajectory inside the domain at time t, that is, the probability that the trajectory was not absorbed prior to t. Integrating (3.1) gives $p_{sur}(1,1)=0.77095\ldots$ for $\sigma=\kappa=1$. The survival probability is estimated numerically by the ratio of the number of survived (unabsorbed) trajectories n_{sur} and the total number of simulated trajectories $n=10^7$. Table 1 shows that the convergence rate of the estimated survival probability to its analytic value is $\sqrt{\Delta t}$, as predicted by our boundary layer analysis. The statistical estimation (variance) error due to the finite number of simulated trajectories is $\sqrt{p_{sur}(1-p_{sur})/n}=0.00013\ldots$, which is an order of magnitude smaller than the smallest (bias) error obtained for $\Delta t=10^{-4}$ (see Table 1).

In the second experiment, the drift term $a=-1$ shifts the density leftward and causes more trajectories to react with the boundary. Figure 2 shows the convergence of the numerical scheme to the analytic solution (3.2).

The final experiment corresponds to a reflecting boundary, $P=\kappa=0$, and a constant nonvanishing drift toward the boundary $a=-1$. We simulated $n=10^8$ trajectories to obtain a finer resolution at the boundary. Figure 3 shows a comparison

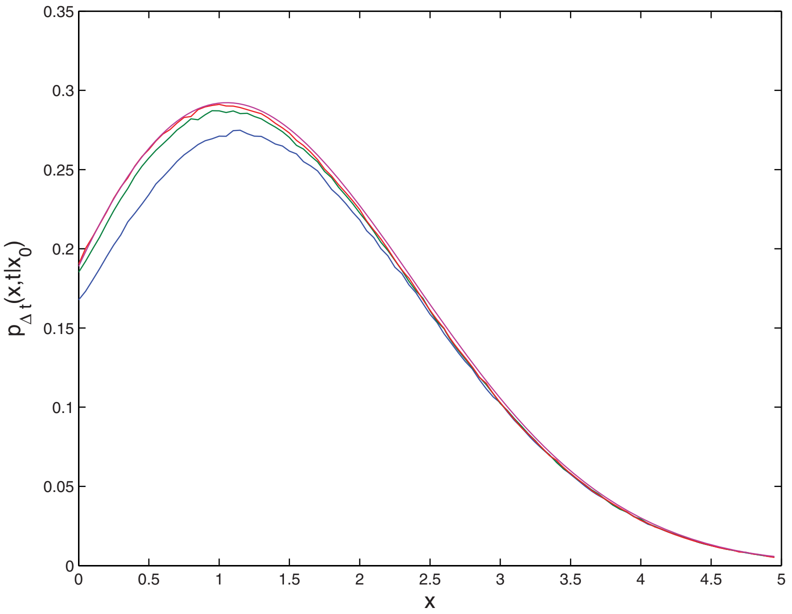


Fig. 1. *No drift: the analytical solution* (3.1) *(magenta) and the three numerical densities* $\Delta t = 10^{-1}$ *(blue),* $\Delta t = 10^{-2}$ *(green),* $\Delta t = 10^{-3}$ *(red) approaching it from below. The numerical density of* $\Delta t = 10^{-4}$ *is not shown because it is difficult to distinguish it from the analytic density. (Parameters:* $\sigma = \kappa = x_0 = t = 1$, $a = 0$, $P = \sqrt{\pi}$, $n = 10^7$.*)*

Table 1

Survival probability: the difference between the analytic value of the survival probability $p_{sur} = 0.77095\ldots$ *and its numerical estimation* n_{sur}/n *decreases by roughly* $\sqrt{10}$ *whenever* Δt *is decreased by an order of magnitude. (Parameters:* $\sigma = \kappa = x_0 = t = 1$, $a = 0$, $n = 10^7$.*)*

Δt	n_{sur}	$p_{sur} - n_{sur}/n$
10^{-1}	7253450	0.0456
10^{-2}	7577156	0.0132
10^{-3}	7670969	0.0039
10^{-4}	7698523	0.0011

between the analytical solution (3.2) and the numerical densities for $\Delta t = 10^{-1}, 10^{-2}$. The no flux condition $J = 0$ of a reflecting boundary together with (1.4) gives a negative boundary derivative, $p_y(0,t) = -p(0,t) < 0$. In particular, the analytic solution (3.2) satisfies $p_y(0,1) = -p(0,1) = -(2+\sqrt{\pi})/(2\sqrt{\pi}) \approx -1.06$. The numerical densities, however, are flat at the boundary. Their first derivatives vanish at the boundary, as predicted in (2.2) and shown in Figure 3. The first derivative changes from 0 to $O(1)$ on an interval of length $O(\sqrt{\Delta t})$, manifesting a boundary layer behavior, though there is no such behavior in the density itself.

4. Diffusion in $\mathbb{R}^d$ with partial oblique reflection at the boundary. We consider the d-dimensional stochastic dynamics

$$\dot{\boldsymbol{x}} = \boldsymbol{a}(\boldsymbol{x},t) + \sqrt{2}\boldsymbol{B}(t)\,\dot{\boldsymbol{w}} \tag{4.1}$$

in the half-space

$$\Omega = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_1 > 0\},$$

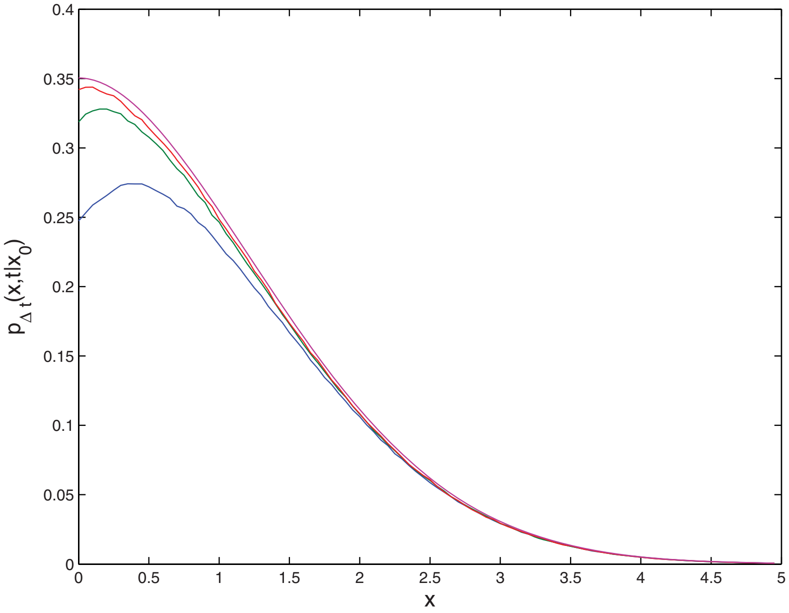


FIG. 2. *Drift, $a = -1$: the analytical solution* (3.2) *(magenta) and the numerical densities $\Delta t = 10^{-1}$ (blue), $\Delta t = 10^{-2}$ (green), $\Delta t = 10^{-3}$ (red) that approach it from below. (Parameters: $\sigma = \kappa = x_0 = t = 1$, $P = \sqrt{\pi}$, $n = 10^7$.)*

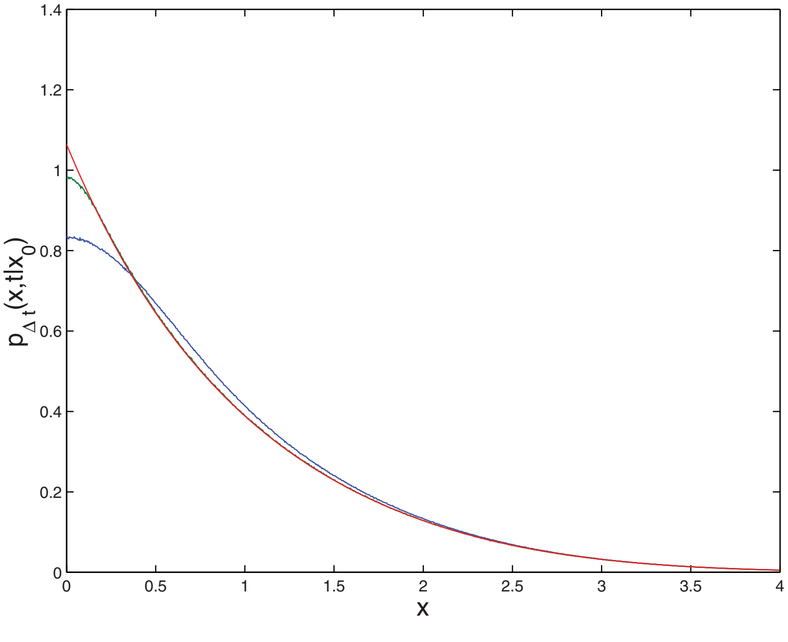


FIG. 3. *Drift, $a = -1$, reflecting boundary $P = \kappa = 0$: the analytic solution* (3.2) *(red) and the numerical densities $\Delta t = 10^{-1}$ (blue) and $\Delta t = 10^{-2}$ (green) with $n = 10^8$ simulated trajectories to obtain a finer boundary resolution. (Parameters: $\sigma = \kappa = x_0 = t = 1$.)*

where $\boldsymbol{w}$ is a vector of d independent Brownian motions and we assume that the diffusion tensor $\boldsymbol{\sigma}(t) = \boldsymbol{B}(t)\boldsymbol{B}^T(t)$ is uniformly positive definite for all $t \geq s$. The case of space-dependent diffusion involves many technically complicated calculations and will be considered in a separate paper. We use henceforth the abbreviation $\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}$. The radiation condition (1.6) becomes

$$-\boldsymbol{J}(\boldsymbol{y},t\,|\,\boldsymbol{x},s)\cdot\boldsymbol{n} = \kappa(\boldsymbol{y},t)p(\boldsymbol{y},t\,|\,\boldsymbol{x},s), \quad \text{for} \quad \boldsymbol{y}\in\partial\Omega,\ \boldsymbol{x}\in\Omega, \tag{4.2}$$

where the components of the flux vector $\boldsymbol{J}(\boldsymbol{y},t\,|\,\boldsymbol{x},s)$ are defined by

$$J^k(\boldsymbol{y},t\,|\,\boldsymbol{x},s) = -[a^k(\boldsymbol{y},t)p(\boldsymbol{y},t\,|\,\boldsymbol{x},s)] + \sum_{j=1}^{d}\frac{\partial}{\partial y_j}\left[\sigma^{j,k}p(\boldsymbol{y},t\,|\,x,s)\right], \tag{4.3}$$

where $\sigma^{j,k}$ are the elements of the diffusion matrix $\boldsymbol{\sigma}$. The Fokker–Plank equation for the pdf of $\boldsymbol{x}(t)$ can be written as

$$\frac{\partial p(\boldsymbol{y},t\,|\,\boldsymbol{x},s)}{\partial t} = -\nabla_{\boldsymbol{y}}\cdot\boldsymbol{J}(\boldsymbol{y},t\,|\,\boldsymbol{x},s) \quad \text{for all} \quad \boldsymbol{y},\boldsymbol{x}\in\Omega. \tag{4.4}$$

If $\boldsymbol{x}\in\Omega$, but

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{a}(\boldsymbol{x},t)\Delta t + \sqrt{2}\boldsymbol{B}(t)\,\Delta\boldsymbol{w}(t,\Delta t) \notin \Omega,$$

the Euler scheme for (4.1) with oblique reflection in $\partial\Omega$ reflects the point $\boldsymbol{x}'$ obliquely in the constant direction of $\boldsymbol{v}$ to a point $\boldsymbol{x}''\in\Omega$, as described below. First, we denote by $\boldsymbol{x}'_B$ the normal projection of a point $\boldsymbol{x}'$ on $\partial\Omega$, that is, $\boldsymbol{x}'_B = \boldsymbol{x}' - (\boldsymbol{x}'\cdot\boldsymbol{n})\boldsymbol{n}$. Then we write the Euler scheme for (4.1) with partially reflecting boundary as

$$\boldsymbol{x}(t+\Delta t) = \begin{cases} \boldsymbol{x}' \quad \text{for} \quad \boldsymbol{x}'\in\Omega, \\ \boldsymbol{x}'' \quad \text{w.p.} \quad 1 - P\left(\boldsymbol{x}'_B\right)\sqrt{\Delta t} \quad \text{if} \quad \boldsymbol{x}'\notin\Omega, \\ \text{terminate trajectory w.p. } P\left(\boldsymbol{x}'_B\right)\sqrt{\Delta t} \quad \text{if} \quad \boldsymbol{x}'\notin\Omega. \end{cases} \tag{4.5}$$

The value of the termination probability $P\left(\boldsymbol{x}'_B\right)\sqrt{\Delta t}$, which varies continuously in the boundary, is evaluated at the normal projection of the point $\boldsymbol{x}'$ on the boundary. The oblique reflection in the direction of the unit vector $\boldsymbol{v}$ ($v_1 \neq 0$) is defined by

$$\boldsymbol{x}'' = \boldsymbol{x}' - \frac{2x'_1}{v_1}\boldsymbol{v}. \tag{4.6}$$

Note that $x''_1 = -x'_1$ guarantees that the reflected point of a crossing trajectory is inside the domain Ω. The fact that the normal components of $\boldsymbol{x}''$ and $\boldsymbol{x}'$ are of equal lengths makes the high-dimensional boundary layer analysis similar to that in one dimension. Normal reflection corresponds to $\boldsymbol{v} = \boldsymbol{n} = (1,0,\dots,0)$.

We note that for a point $\boldsymbol{y}\in\Omega$, we can write $\Pr\{\boldsymbol{x}'' = \boldsymbol{y}\} = \Pr\{\boldsymbol{x}' = \boldsymbol{y}'\}$, where

$$\boldsymbol{y} = \boldsymbol{y}' - \frac{2\boldsymbol{y}'\cdot\boldsymbol{n}}{v_1}\boldsymbol{v} \tag{4.7}$$

is the oblique reflection of $\boldsymbol{y}'$ (see Figure 4). Given $\boldsymbol{y}$, (4.7) defines $\boldsymbol{y}'$ as

$$\boldsymbol{y}' = \boldsymbol{y} - 2\frac{y_1}{v_1}\boldsymbol{v}. \tag{4.8}$$

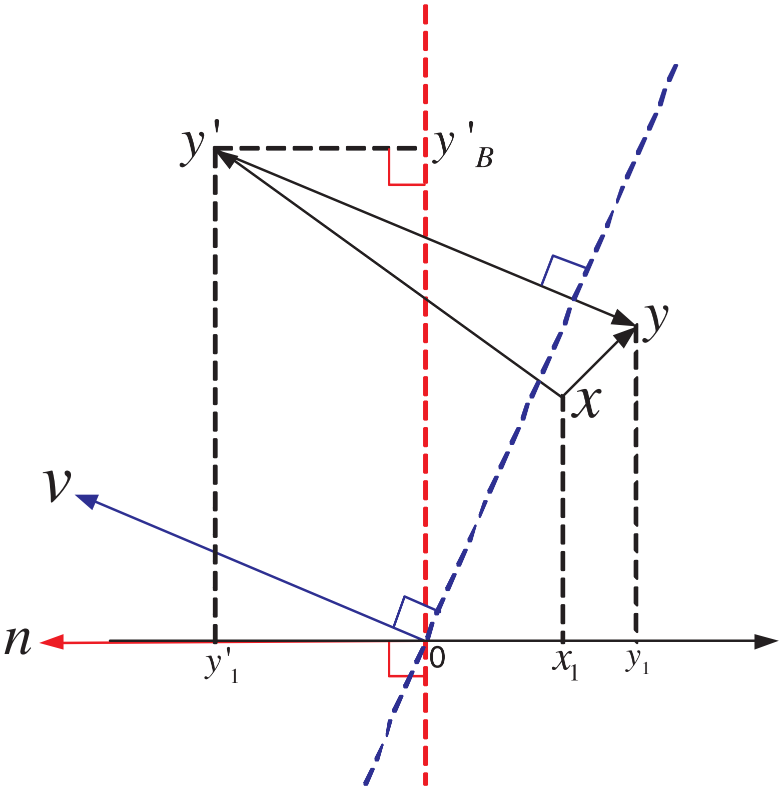


FIG. 4. *A simulated trajectory can get from* $\boldsymbol{x}$ *to* $\boldsymbol{y}$ *in a single time step* Δt *in two different ways:* (i) *directly from* $\boldsymbol{x}$ *to* $\boldsymbol{y}$*, without crossing the boundary, and* (ii) *by crossing the boundary from* $\boldsymbol{x}$ *to* $\boldsymbol{y}'$ *and reflection in the oblique direction* $\boldsymbol{v}$ *with probability* $1-P(\boldsymbol{y}'_B)\sqrt{\Delta t}$ *to* $\boldsymbol{y}$*. The reflection law* (4.5)–(4.7) *satisfies* $y'_1=-y_1$*.*

As in the one-dimensional case, the forward Kolmogorov equation is

$$p_{\Delta t}(\boldsymbol{y},t+\Delta t)=\int_{x_1>0}\frac{p_{\Delta t}(\boldsymbol{x},t)}{(4\pi\Delta t)^{d/2}\sqrt{\det\boldsymbol{\sigma}}}\left\{\exp\left[-\frac{\mathcal{B}(\boldsymbol{x}+\boldsymbol{a}(\boldsymbol{x},t)\Delta t,\boldsymbol{y})}{4\Delta t}\right]\right.$$
$$\left.+\left(1-P(\boldsymbol{y}'_B)\sqrt{\Delta t}\right)\exp\left[-\frac{\mathcal{B}(\boldsymbol{x}+\boldsymbol{a}(\boldsymbol{x},t)\Delta t,\boldsymbol{y}')}{4\Delta t}\right]\right\}d\boldsymbol{x},\tag{4.9}$$

where

$$\mathcal{B}(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y}).\tag{4.10}$$

We construct a boundary layer of width $O(\sqrt{\Delta t})$ in the normal direction to the boundary. The layer extends infinitely in the $d-1$ directions tangent to the boundary

$$p_{BL}(\eta_1,y_2,\dots,y_d,t)=p_{\Delta t}(\eta_1\sqrt{\Delta t},y_2,\dots,y_d,t).\tag{4.11}$$

In other words, $p_{BL}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)=p_{\Delta t}(\eta_1\sqrt{\Delta t}\,\boldsymbol{n}+\boldsymbol{y}_B,t)$, where $\boldsymbol{y}_B=(0,y_2,y_3,\dots,y_d)$. As in the one-dimensional case, we assume the asymptotic expansion

$$p_{BL}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)\sim p_{BL}^{(0)}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)+\sqrt{\Delta t}\,p_{BL}^{(1)}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)+\cdots\tag{4.12}$$

and substitute

$$\boldsymbol{x} = \boldsymbol{y}_B + \sqrt{\Delta t}\,\boldsymbol{\xi} \tag{4.13}$$

in the integral (4.9). We obtain

$$\begin{aligned} p_{BL}(\eta_1 \boldsymbol{n} + \boldsymbol{y}_B, t + \Delta t) = \int_{\xi_1 > 0} & \frac{p_{BL}(\xi_1 \boldsymbol{n} + \boldsymbol{y}_B + \sqrt{\Delta t}\,\boldsymbol{\xi}_B, t)}{(4\pi)^{d/2}\sqrt{\det \boldsymbol{\sigma}}} \\ \times & \left\{ \exp\left[-\frac{\mathcal{B}(\boldsymbol{\xi} + \boldsymbol{a}(\boldsymbol{y}_B, t)\sqrt{\Delta t}, \eta_1 \boldsymbol{n})}{4} \right] + (1 - P(\boldsymbol{y}'_B)\sqrt{\Delta t}) \right. \\ \times & \left. \exp\left[-\frac{1}{4}\mathcal{B}\left(\boldsymbol{\xi} + \boldsymbol{a}(\boldsymbol{y}_B, t)\sqrt{\Delta t},\ \eta_1 \boldsymbol{n} - \frac{2\eta_1}{v_1}\boldsymbol{v} \right) \right] \right\} d\boldsymbol{\xi} + O(\Delta t). \end{aligned} \tag{4.14}$$

We calculate separately the integral of the first and second terms in the braces. Substituting

$$\boldsymbol{z} = \boldsymbol{\sigma}^{-1/2}(\boldsymbol{\xi} - \eta_1 \boldsymbol{n}) \tag{4.15}$$

in the first integral of (4.14) transforms the domain of integration into

$$\boldsymbol{z} \cdot \tilde{\boldsymbol{n}} > -\frac{\eta_1}{\sqrt{\sigma_n}}, \tag{4.16}$$

where $\tilde{\boldsymbol{n}} = \frac{\boldsymbol{\sigma}^{1/2}\boldsymbol{n}}{\|\boldsymbol{\sigma}^{1/2}\boldsymbol{n}\|}$ is a unit vector and $\sigma_n = \boldsymbol{n}^T \boldsymbol{\sigma} \boldsymbol{n} = \|\boldsymbol{\sigma}^{1/2}\boldsymbol{n}\|^2$. Similarly, we transform the second integral by substituting $\boldsymbol{z}' = \boldsymbol{\sigma}^{-1/2}\left(\boldsymbol{\xi} - \eta_1 \boldsymbol{n} + \frac{2\eta_1}{v_1}\boldsymbol{v}\right)$. Using the expansion (4.12), we obtain at the leading order the integral equation

$$\begin{aligned} & p^{(0)}_{BL}(\eta_1 \boldsymbol{n} + \boldsymbol{y}_B, t) \\ & = \frac{1}{(4\pi)^{d/2}} \int_{\boldsymbol{z}\cdot\tilde{\boldsymbol{n}} > -\frac{\eta_1}{\sqrt{\sigma_n}}} p^{(0)}_{BL}\left((\eta_1 + \sqrt{\sigma_n}\,\boldsymbol{z}\cdot\tilde{\boldsymbol{n}})\boldsymbol{n} + \boldsymbol{y}_B, t\right) \exp\left[-\frac{\|\boldsymbol{z}\|^2}{4} \right] d\boldsymbol{z} \\ & \quad + \frac{1}{(4\pi)^{d/2}} \int_{\boldsymbol{z}'\cdot\tilde{\boldsymbol{n}} > \frac{\eta_1}{\sqrt{\sigma_n}}} p^{(0)}_{BL}\left((-\eta_1 + \sqrt{\sigma_n}\,\boldsymbol{z}'\cdot\tilde{\boldsymbol{n}})\boldsymbol{n} + \boldsymbol{y}_B, t\right) \exp\left[-\frac{\|\boldsymbol{z}'\|^2}{4} \right] d\boldsymbol{z}'. \end{aligned}$$

Integrating in the $d-1$ directions orthogonal to $\tilde{\boldsymbol{n}}$ yields

$$\begin{aligned} p^{(0)}_{BL}(\eta_1 \boldsymbol{n} + \boldsymbol{y}_B, t) & = \frac{1}{\sqrt{4\pi}} \int_{-\frac{\eta_1}{\sqrt{\sigma_n}}}^{\infty} p^{(0)}_{BL}\left((\eta_1 + \sqrt{\sigma_n}\,u)\boldsymbol{n} + \boldsymbol{y}_B, t\right) \exp\left[-\frac{u^2}{4} \right] du \\ & + \frac{1}{\sqrt{4\pi}} \int_{\frac{\eta_1}{\sqrt{\sigma_n}}}^{\infty} p^{(0)}_{BL}\left((-\eta_1 + \sqrt{\sigma_n}\,u)\boldsymbol{n} + \boldsymbol{y}_B, t\right) \exp\left[-\frac{u^2}{4} \right] du \\ & = \frac{1}{\sqrt{4\pi\sigma_n}} \int_0^{\infty} p^{(0)}_{BL}\left(u\boldsymbol{n} + \boldsymbol{y}_B, t\right) \left\{ \exp\left[-\frac{(u - \eta_1)^2}{4\sigma_n} \right] + \exp\left[-\frac{(u + \eta_1)^2}{4\sigma_n} \right] \right\} du. \end{aligned}$$

This is the same leading order integral equation as that of the one-dimensional case (2.9); thus the solution is independent of η_1, and matching to the outer solution gives

$$p^{(0)}_{BL}(\eta_1 \boldsymbol{n} + \boldsymbol{y}_B, t) = p^{(0)}_{OUT}(\boldsymbol{y}_B, t). \tag{4.17}$$

To evaluate the $O(\sqrt{\Delta t})$ terms, we expand in the first integral in (4.14)

$$\begin{aligned}\mathcal{B}(\boldsymbol{\xi}+\boldsymbol{a}(\boldsymbol{y}_B,t)\sqrt{\Delta t},\eta_1\boldsymbol{n}) &= (\boldsymbol{\xi}-\eta_1\boldsymbol{n})\cdot\boldsymbol{\sigma}^{-1}(\boldsymbol{\xi}-\eta_1\boldsymbol{n})\\ &\quad+\sqrt{\Delta t}\,2\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{\sigma}^{-1}(\boldsymbol{\xi}-\eta_1\boldsymbol{n}),\end{aligned}\tag{4.18}$$

and in the second integral

$$\begin{aligned}\mathcal{B}\left(\boldsymbol{\xi}+\boldsymbol{a}(\boldsymbol{y}_B,t)\sqrt{\Delta t},\ \eta_1\boldsymbol{n}-\frac{2\eta_1}{v_1}\boldsymbol{v}\right) &= \left(\boldsymbol{\xi}-\eta_1\boldsymbol{n}+\frac{2\eta_1}{v_1}\boldsymbol{v}\right)\cdot\boldsymbol{\sigma}^{-1}\left(\boldsymbol{\xi}-\eta_1\boldsymbol{n}\frac{2\eta_1}{v_1}\boldsymbol{v}\right)\\ &\quad+\sqrt{\Delta t}\,2\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{\sigma}^{-1}\left(\boldsymbol{\xi}-\eta_1\boldsymbol{n}\frac{2\eta_1}{v_1}\boldsymbol{v}\right).\end{aligned}\tag{4.19}$$

The $O(\sqrt{\Delta t})$ contribution of the drift term for the first exponential term is

$$\begin{aligned}&-\frac{1}{4}\int_{\xi_1>0}\frac{p_{OUT}^{(0)}(\boldsymbol{y}_B,t)}{(4\pi)^{d/2}\sqrt{\det\boldsymbol{\sigma}}}\exp\left\{-\frac{\mathcal{B}(\boldsymbol{\xi},\eta_1\boldsymbol{n})}{4}\right\}\left[2\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{\sigma}^{-1}(\boldsymbol{\xi}-\eta_1\boldsymbol{n})\right]d\boldsymbol{\xi}\\ &=-\frac{1}{4}\frac{p_{OUT}^{(0)}(\boldsymbol{y}_B,t)}{\sqrt{4\pi}}2\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{\sigma}^{-1/2}\tilde{\boldsymbol{n}}\int_{-\eta_1/\sqrt{\sigma_n}}^{\infty}ue^{-u^2/4}\,du\\ &=-\frac{1}{2}\frac{p_{OUT}^{(0)}(\boldsymbol{y}_B,t)}{\sqrt{\pi\sigma_n}}\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\exp\left\{\frac{-\eta_1^2}{4\sigma_n}\right\}.\end{aligned}\tag{4.20}$$

The second exponential has the same contribution, so the overall contribution of the drift to the $O(\sqrt{\Delta t})$ term is

$$-\frac{p_{OUT}^{(0)}(\boldsymbol{y}_B,t)}{\sqrt{\pi\sigma_n}}\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\exp\left\{\frac{-\eta_1^2}{4\sigma_n}\right\}.\tag{4.21}$$

Now, we expand

$$\begin{aligned}&p_{BL}^{(0)}\left((\eta_1+\sqrt{\sigma_n}\,\boldsymbol{z}\cdot\tilde{\boldsymbol{n}})\boldsymbol{n}+\boldsymbol{y}_B+\sqrt{\Delta t}\,(\boldsymbol{\sigma}^{1/2}\boldsymbol{z})_B,t\right)=p_{BL}^{(0)}\left((\eta_1+\sqrt{\sigma_n}\,\boldsymbol{z}\cdot\tilde{\boldsymbol{n}})\boldsymbol{n}+\boldsymbol{y}_B,t\right)\\ &+\sqrt{\Delta t}\,\nabla p_{BL}^{(0)}\left((\eta_1+\sqrt{\sigma_n}\,\boldsymbol{z}\cdot\tilde{\boldsymbol{n}})\boldsymbol{n}+\boldsymbol{y}_B,t\right)\cdot(\boldsymbol{\sigma}^{1/2}\boldsymbol{z})_B+O(\Delta t).\end{aligned}\tag{4.22}$$

Together with (4.17), the expansion (4.22) reduces to

$$\begin{aligned}&p_{BL}^{(0)}\left((\eta_1+\sqrt{\sigma_n}\,\boldsymbol{z}\cdot\tilde{\boldsymbol{n}})\boldsymbol{n}+\boldsymbol{y}_B+\sqrt{\Delta t}\,(\boldsymbol{\sigma}^{1/2}\boldsymbol{z})_B,t\right)\\ &=p_{OUT}^{(0)}\left(\boldsymbol{y}_B,t\right)+\sqrt{\Delta t}\,\nabla p_{OUT}^{(0)}\left(\boldsymbol{y}_B,t\right)\cdot(\boldsymbol{\sigma}^{1/2}\boldsymbol{z})_B+O(\Delta t).\end{aligned}$$

Integrating as above, we obtain the $O(\sqrt{\Delta t})$ integral equation as

$$\begin{aligned}&p_{BL}^{(1)}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)\\ &=\frac{1}{\sqrt{4\pi\sigma_n}}\int_0^{\infty}p_{BL}^{(1)}\left(u\boldsymbol{n}+\boldsymbol{y}_B,t\right)\left\{\exp\left[-\frac{(u-\eta_1)^2}{4\sigma_n}\right]+\exp\left[-\frac{(u+\eta_1)^2}{4\sigma_n}\right]\right\}du\\ &\quad-\frac{P\left(\boldsymbol{y}_B'\right)p_{OUT}^{(0)}\left(\boldsymbol{y}_B,t\right)}{\sqrt{4\pi\sigma_n}}\int_0^{\infty}\exp\left[-\frac{(u+\eta_1)^2}{4\sigma_n}\right]du\\ &\quad+\frac{1}{\sqrt{4\pi}}\int_{\frac{\eta_1}{\sqrt{\sigma_n}}}^{\infty}\nabla p_{OUT}^{(0)}(\boldsymbol{y}_B,t)\cdot\left(2\boldsymbol{\sigma}^{1/2}u\tilde{\boldsymbol{n}}-\frac{2\eta_1}{v_1}\boldsymbol{v}\right)_B\exp\left[-\frac{u^2}{4}\right]du\\ &\quad-\frac{p_{OUT}^{(0)}(\boldsymbol{y}_B,t)}{\sqrt{\pi\sigma_n}}\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\exp\left\{\frac{-\eta_1^2}{4\sigma_n}\right\}.\end{aligned}$$

Differentiating with respect to η_1 and integrating by parts (as was done in the one-dimensional case), we arrive at the integral equation

$$\begin{aligned}
&\frac{\partial p^{(1)}_{BL}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)}{\partial n}\\
&= \frac{1}{\sqrt{4\pi\sigma_n}}\int_0^\infty \frac{\partial p^{(1)}_{BL}\left(u\boldsymbol{n}+\boldsymbol{y}_B,t\right)}{\partial n}\left\{\exp\left[-\frac{(u-\eta_1)^2}{4\sigma_n}\right]-\exp\left[-\frac{(u+\eta_1)^2}{4\sigma_n}\right]\right\}du\\
&\quad-\frac{P\left(\boldsymbol{y}'_B\right)p^{(0)}_{OUT}\left(\boldsymbol{y}_B,t\right)}{\sqrt{4\pi\sigma_n}}\exp\left[\frac{-\eta_1^2}{4\sigma_n}\right]\\
&\quad+\nabla p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\cdot\left\{-\frac{1}{\sqrt{\pi\sigma_n}}\left[\frac{\boldsymbol{\sigma n}}{\sigma_n}-\frac{\boldsymbol{v}}{v_1}\right]\eta_1\exp\left[\frac{-\eta_1^2}{4\sigma_n}\right]-\boldsymbol{v}\frac{\operatorname{erfc}\left(\frac{\eta_1}{2\sqrt{\sigma_n}}\right)}{v_1}\right\}_B\\
&\quad+\frac{p^{(0)}_{OUT}(\boldsymbol{y}_B,t)}{\sqrt{\pi\sigma_n}}\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\frac{\eta_1}{2\sigma_n}\exp\left[\frac{-\eta_1^2}{4\sigma_n}\right].
\end{aligned}$$

The Wiener–Hopf method requires the extension of the erfc function discontinuously as an odd function, that is, to define $\widetilde{\operatorname{erfc}}(x)=\operatorname{sgn}(x)\operatorname{erfc}(|x|)$. Following the calculations of the one-dimensional case, it remains to determine the small k behavior of the Fourier transform of $\widetilde{\operatorname{erfc}}(x)$. Using

$$(4.23)\qquad \int_{-\infty}^{\infty}\widetilde{\operatorname{erfc}}\left(\frac{\eta}{2\sqrt{\sigma_n}}\right)\exp\{ik\eta\}\,d\eta\sim 2ik\int_0^\infty\operatorname{erfc}\left(\frac{\eta}{2\sqrt{\sigma_n}}\right)\eta\,d\eta=2ik\sigma_n,$$

we obtain, as in (2.24),

$$\begin{aligned}
\hat{\phi}(k)\sim 2ik\Bigg\{&\frac{P\left(\boldsymbol{y}'_B\right)p^{(0)}_{OUT}\left(\boldsymbol{y}_B,t\right)\sqrt{\sigma_n}}{\sqrt{\pi}}-2\sigma_n\nabla p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\cdot\left[\frac{\boldsymbol{\sigma n}}{\sigma_n}-\frac{\boldsymbol{v}}{2v_1}\right]_B\\
&+p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\Bigg\}\quad\text{as}\quad k\to 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\lim_{\eta_1\to\infty}\frac{\partial p^{(1)}_{BL}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)}{\partial n}\\
&=\left\{\frac{P\left(\boldsymbol{y}'_B\right)p^{(0)}_{OUT}\left(\boldsymbol{y}_B,t\right)}{\sqrt{\pi\sigma_n}}-2\nabla p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\cdot\left[\frac{\boldsymbol{\sigma n}}{\sigma_n}-\frac{\boldsymbol{v}}{2v_1}\right]_B+p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\frac{\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}}{\sigma_n}\right\}.
\end{aligned}$$

Combining with the matching condition

$$(4.24)\qquad \lim_{\eta\to\infty}\frac{\partial p^{(1)}_{BL}(\eta_1\boldsymbol{n}+\boldsymbol{y}_B,t)}{\partial n}=\frac{\partial p^{(0)}_{OUT}(\boldsymbol{y}_B,t)}{\partial n},$$

we obtain

$$\begin{aligned}
&\frac{\partial p^{(0)}_{OUT}(\boldsymbol{y}_B,t)}{\partial n}\\
&=\left\{\frac{P\left(\boldsymbol{y}_B\right)p^{(0)}_{OUT}\left(\boldsymbol{y}_B,t\right)}{\sqrt{\pi\sigma_n}}-2\nabla p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\cdot\left[\frac{\boldsymbol{\sigma n}}{\sigma_n}-\frac{\boldsymbol{v}}{2v_1}\right]_B+p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\frac{\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}}{\sigma_n}\right\}.
\end{aligned}$$

The requirement that the pdf of the limiting diffusion process satisfies the Robin boundary condition leads to the only possible choice,

$$\boldsymbol{v} = \frac{\boldsymbol{\sigma n}}{\|\boldsymbol{\sigma n}\|}. \tag{4.25}$$

Otherwise, we obtain an oblique derivative boundary condition. Since $\boldsymbol{y}'_B \to \boldsymbol{y}_B$ as $\Delta t \to 0$, we obtain the Robin boundary condition

$$\begin{aligned}-\boldsymbol{J}_{OUT}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n} &= \nabla p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\cdot\boldsymbol{\sigma n} - p^{(0)}_{OUT}(\boldsymbol{y}_B,t)\boldsymbol{a}(\boldsymbol{y}_B,t)\cdot\boldsymbol{n}\\ &= \frac{P\left(\boldsymbol{y}_B\right)p^{(0)}_{OUT}\left(\boldsymbol{y}_B,t\right)\sqrt{\sigma_n}}{\sqrt{\pi}}.\end{aligned}$$

The reflection direction $\boldsymbol{v}$ of crossing trajectories is the conormal direction $\boldsymbol{\sigma n}$. Normal reflection (i.e., replacing $\boldsymbol{v}$ by $\boldsymbol{n}$) gives rise to the boundary normal flux if and only if $\boldsymbol{n}$ is an eigenvector of the diffusion tensor $\boldsymbol{\sigma}$. The limit of the outer solution as $\Delta t \to 0$ is the solution of the Fokker–Planck equation (4.4) with the radiation boundary condition

$$-\boldsymbol{J}(\boldsymbol{y},t)\cdot\boldsymbol{n} = \kappa(\boldsymbol{y})p\left(\boldsymbol{y},t\right) \quad \text{for} \quad \boldsymbol{y}\in\partial\Omega, \tag{4.26}$$

where the reactive "constant" is

$$\kappa(\boldsymbol{y}) = \frac{P(\boldsymbol{y})\sqrt{\sigma_n}}{\sqrt{\pi}}. \tag{4.27}$$

Note that normal reflection will not recover the normal flux of the radiation condition if $\boldsymbol{n}$ is not an eigenvector of $\boldsymbol{\sigma}$.

5. Numerical simulations in two dimensions. To illustrate the conormal reflection law (4.25) in the Euler scheme (4.5)–(4.7) in the half-plane $x \geq 0$, we ran several numerical experiments. The simulations show the convergence of the pdf of the numerical solution to that of the FPE with the radiation boundary condition (4.26)–(4.27). Unlike in the one-dimensional case, no explicit solution of the anisotropic Robin problem for the FPE in the half-plane is available, so we compare the statistics of the simulated trajectories with a numerical solution of the FPE. The latter is constructed by the stable Crank–Nicolson scheme on lattice points, where in each time step the sparse linear system is solved by the conjugate gradient method.

In all numerical experiments the initial point is $(x_0, y_0) = (0.3, 0)$, and the statistics are collected at time $T = 0.5$. We choose the reactive constant $\kappa = 1$ and the diffusion matrix $\boldsymbol{B}$ in (4.1),

$$\boldsymbol{B} = \begin{pmatrix} 0.3 & 0.4 \\ 0 & 1 \end{pmatrix},$$

which gives the anisotropic diffusion tensor

$$\boldsymbol{\sigma} = \boldsymbol{B}\boldsymbol{B}^T = \begin{pmatrix} 0.25 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

We simulate $n = 10^7$ trajectories with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ in each experiment.

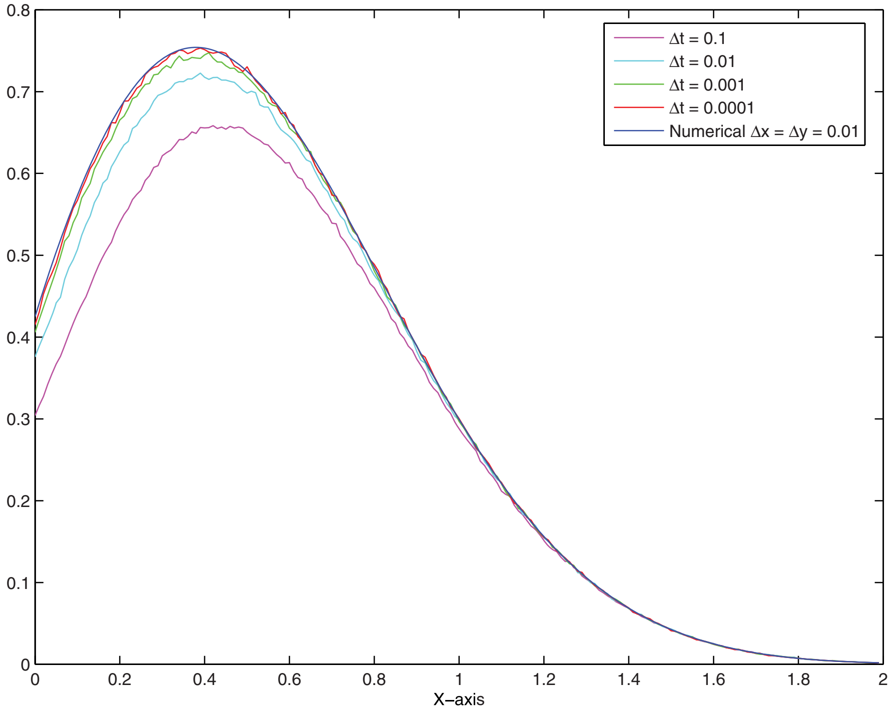


Fig. 5. *The marginal density of $x(T)$ with no drift and correct oblique reflection (the first experiment). The numerical solution of the FPE (blue) with grid size $\Delta x = 0.01$ and estimates from the simulation of $n = 10^7$ trajectories with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.*

In the first experiment the drift vanishes ($\boldsymbol{a} = \mathbf{0}$). The normal $\boldsymbol{n} = (1, 0)$ and the conormal $\boldsymbol{\sigma n} = (0.25, 0.4)$ point in different directions. The simulated trajectories are reflected in the conormal direction according to the prescription (4.25). The simulated and the numerical solutions of the FPE give the marginal densities shown in Figures 5 and 6. Figure 5 shows the marginal density of $x(T)$,

$$p(x, T \,|\, x_0, y_0) = \int_{-\infty}^{\infty} p(x, y, T \,|\, x_0, y_0)\, dy,$$

while Figure 6 shows the marginal density of $y(T)$,

$$p(y, T \,|\, x_0, y_0) = \int_{0}^{\infty} p(x, y, T \,|\, x_0, y_0)\, dx.$$

Table 2 gives the computed survival probability and indicates the convergence rate.

We illustrate the importance of using the correct reflection law in the second experiment, in which the simulated trajectories are reflected in the normal direction $\boldsymbol{n} = (1, 0)$. Clearly, the marginal density of $x(T)$ coincides with that of the first experiment, because both oblique and normal reflections have the same x-coordinate (see (4.6)). However, the plot of the marginal density of $y(T)$ differs significantly from that in the previous experiment. It is apparent from the comparison to the numerical solution of the FPE that the simulation does not recover the Robin boundary condition in the limit $\Delta t \to 0$ (see Figure 7). Note that the peak of the density is at $y > 0$,

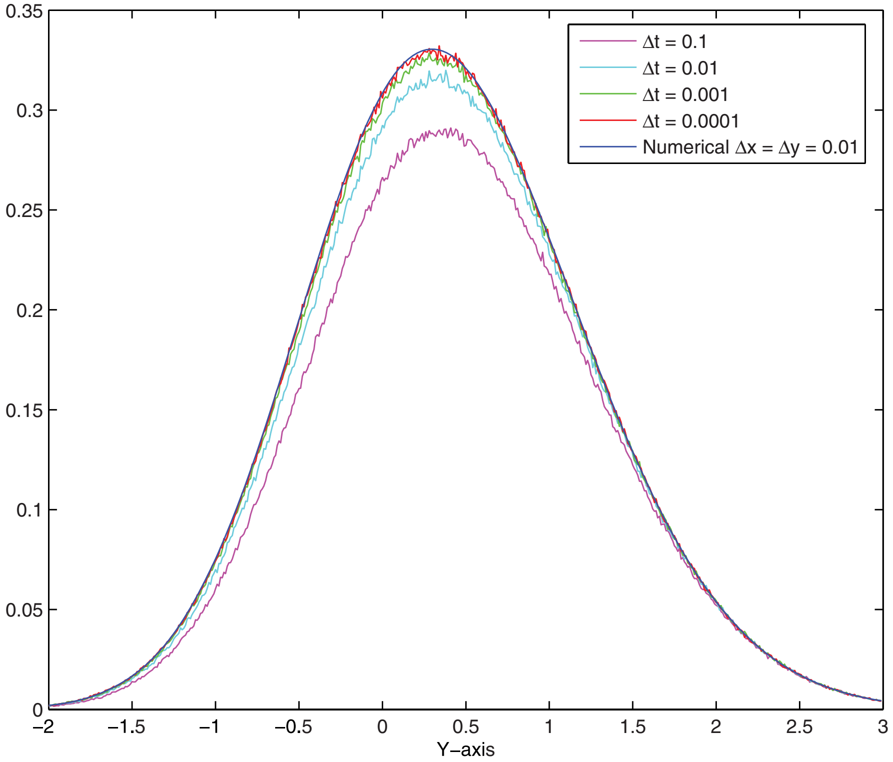


Fig. 6. *The marginal density of $y(T)$ with no drift and correct oblique reflection (the first experiment). The numerical solution of the FPE (blue) with grid size $\Delta x = 0.01$ and estimates from the simulation of $n = 10^7$ trajectories with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.*

Table 2

Survival probability for $\boldsymbol{a} = \mathbf{0}$. The third column lists the error between the numerical value of the survival probability $p_{sur} = 0.6799545$ from the solution of the FPE and its estimate n_{sur}/n from the simulation. The error decreases by about $\sqrt{10}$ whenever Δt is decreased by an order of magnitude, indicating the convergence rate $\sqrt{\Delta t}$ of the simulation.

Δt	n_{sur}	$p_{sur} - n_{sur}/n$
10^{-1}	5986662	0.0814708
10^{-2}	6449991	0.0351379
10^{-3}	6707318	0.0094052
10^{-4}	6775672	0.0025698

though the reflection is normal. This is due to the anisotropy of the diffusion tensor, which causes the probability flux density vector to have a positive y component.

In the third experiment the drift is the constant vector $\boldsymbol{a} = (-1, 0)$, and the diffusion tensor is as in the first experiment. The density is shifted toward the boundary (see Figures 8 and 9). The results are summarized in Table 3.

6. Summary and discussion. We have defined a diffusion process with partially reflecting boundary as a limit of Markovian jump processes generated by the Euler scheme for the dynamics in a half-space with partial absorption of exiting trajectories and partial oblique reflection in the boundary. We derived an expression for the radiation constant in the Robin boundary condition for the one-dimensional

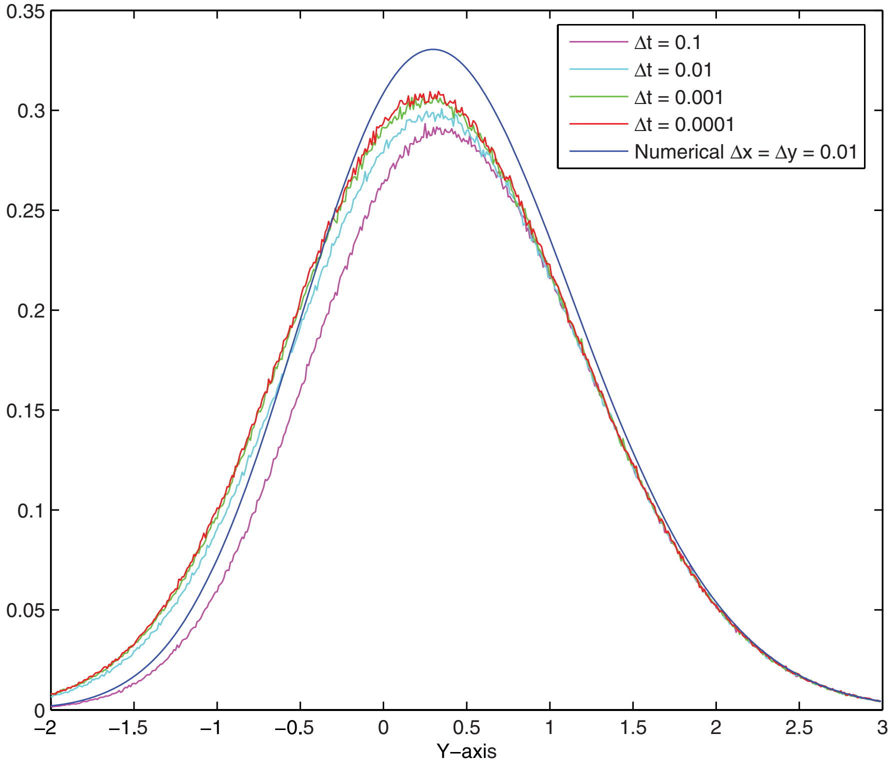


FIG. 7. *The marginal density of $y(T)$ with no drift and with normal reflection (the second experiment). The numerical solution of the FPE (blue) with grid size $\Delta x = 0.01$ and estimates from the simulation of $n = 10^7$ trajectories with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.*

Fokker–Planck equation for the case of diffusion with variable drift and diffusion coefficients, as a function of the absorption probability. We found that the Euler scheme for a diffusion in a half-space with variable drift and constant anisotropic diffusion has to be reflected in a particular oblique direction in order to recover the Robin boundary condition. Also for this case we found the radiation "constant" as a function of the local absorption probability on the boundary. We found a boundary layer of width $O(\sqrt{\Delta t})$ in the pdf of the Euler scheme and solved the boundary layer equation, which is of Wiener–Hopf type.

The boundary layer of $p_{\Delta t}(y,t)$ makes the calculation of the boundary flux nontrivial. The net boundary flux of the simulation profile $p_{\Delta t}(y,t)$ is

$$(6.1) \qquad -J_{\Delta t}(0,t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P\sqrt{\Delta t}}{\sqrt{4\pi\sigma\Delta t}} \int_{-\infty}^{0} dy \int_{0}^{\infty} p_{\Delta t}(x,t) \exp\left\{-\frac{(x-y)^2}{4\pi\sigma\Delta t}\right\} dx,$$

which is the probability of the trajectories that propagate out of the domain per unit time, discounted by the probability of trajectories returned into the domain by the partially reflecting Euler scheme. Changing the order of integration and then changing the variable of integration into $z = x/2\sqrt{\sigma\Delta t}$ gives

$$(6.2) \quad -J_{\Delta t}(0,t) = P\sqrt{\sigma} \int_{0}^{\infty} \operatorname{erfc}(z) p_{\Delta t}(2z\sqrt{\sigma\Delta t},t)\, dz = \frac{P\sqrt{\sigma}}{\sqrt{\pi}} p_{BL}^{(0)}(0,t) + O(\sqrt{\Delta t}).$$

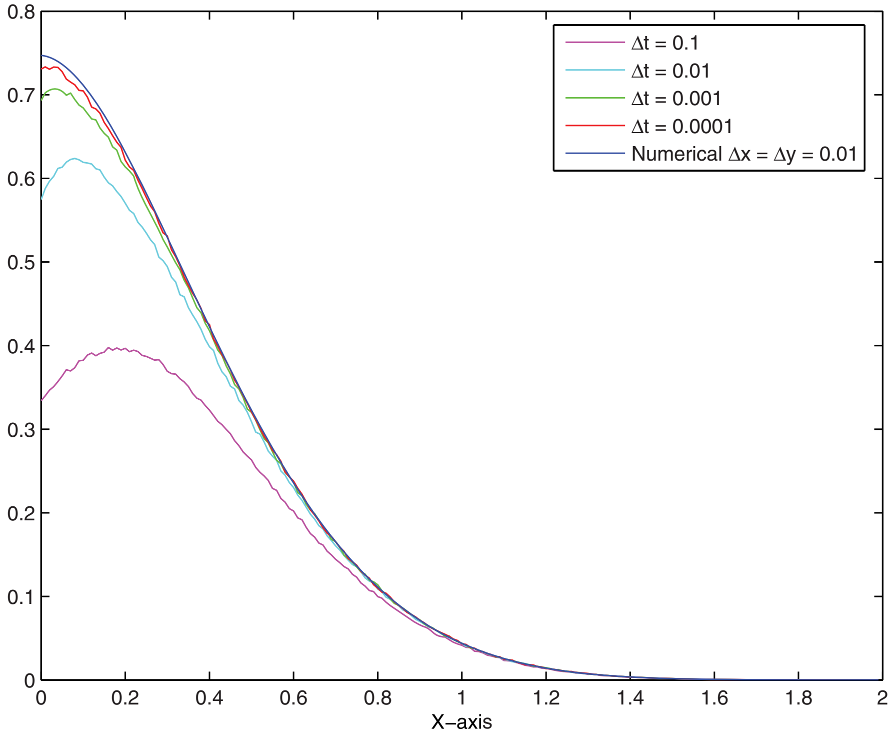


FIG. 8. *The marginal density of $x(T)$ with drift $\boldsymbol{a} = (-1, 0)$ and correct oblique reflection (the third experiment). The numerical solution of the FPE (blue) with grid size $\Delta x = 0.01$ and estimates from the simulation of $n = 10^7$ trajectories with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.*

This straightforward calculation of the flux gives the correct radiation constant, provided that

$$p_{BL}^{(0)}(0,t) = p_{OUT}^{(0)}(0,t). \tag{6.3}$$

The latter, however, depends on the mode of reflecting a trajectory from x' outside to x'' inside the domain. We have shown that for $x'' = -x'$ the provision holds; however, for other schemes, e.g., $x'' = -\alpha x'$ $(\alpha \neq 1)$, the provision (6.3) fails in general, though (6.2) still holds. On the other hand, the differential form of the flux, (1.4), has to be obtained from (6.1) in the limit $\Delta t \to 0$, which is not the case for $p_{\Delta t}(y,t)$, though it is for $p_{OUT}(y,t)$. This shows up in spades in the multidimensional case, because although (6.3) holds for any direction of reflection, the differential form of the flux is obtained in the limit only if the correct direction of oblique reflection is chosen.

The generalization of the multidimensional case to domains with curved boundaries and to a variable diffusion tensor $\boldsymbol{\sigma}(\boldsymbol{x},t)$ is not straightforward and will be done separately. Note that if the diffusion tensor is constant, but anisotropic, a local orthogonal mapping of the boundary to a plane converts the diffusion tensor from constant to variable, as can be seen from Itô's formula. However, as mentioned in section 1, in the most common case of constant isotropic diffusion, our result extends to domains with curved boundaries because the mapping leaves the Laplacian unchanged, though the drift changes according to Itô's formula. In this case the vector $\boldsymbol{v}$ coincides with the normal $\boldsymbol{n}$.

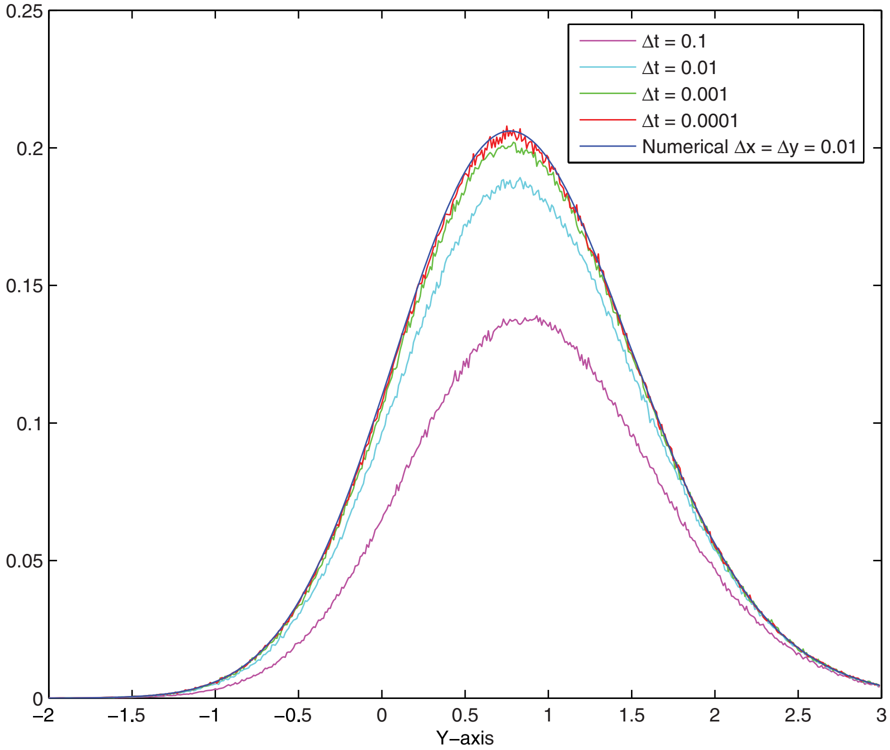


FIG. 9. *The third experiment ($\boldsymbol{a} = (-1, 0)$, correct oblique reflection): y-marginal densities. The numerical solution (blue) is compared to four simulated solutions (with time steps $\Delta t = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$). $n = 10^7$. Resolution: $\Delta x = 0.01$.*

TABLE 3

Survival probability for $\boldsymbol{a} = (-1, 0)$. The third column lists the error between the numerical value of the survival probability $p_{sur} = 0.3722893$ from the solution of the FPE and its estimate n_{sur}/n from the simulation.

Δt	n_{sur}	$p_{sur} - n_{sur}/n$
10^{-1}	2541947	0.1180946
10^{-2}	3399528	0.0323365
10^{-3}	3632622	0.0090271
10^{-4}	3693905	0.0028988

REFERENCES

[1] G. LAMM AND K. SCHULTEN, *Extended Brownian dynamics.* II. *Reactive, nonlinear diffusion*, J. Chem. Phys., 78 (1983), pp. 2713–2734.

[2] R. ZWANZIG, *Diffusion-controlled ligand binding to spheres partially covered by receptors: An effective medium treatment*, Proc. Natl. Acad. Sci. USA, 87 (1990), pp. 5856–5857.

[3] K. TAI, S. D. BOND, H. R. MACMILLAN, N. A. BAKER, M. J. HOLST, AND J. A. MCCAMMON, *Finite element simulations of acetylcholine diffusion in neuromuscular junctions*, Biophys. J., 84 (2003), pp. 2234–2241.

[4] L. BATSILAS, A. M. BEREZHKOVSKII, AND S. Y. SHVARTSMAN, *Stochastic model of autocrine and paracrine signals in cell culture assays*, Biophys. J., 85 (2003), pp. 3659–3665.

[5] A. M. BEREZHKOVSKII, Y. A. MAKHNOVSKII, M. I. MONINE, V. YU. ZITSERMAN, AND S. Y. SHVARTSMAN, *Boundary homogenization for trapping by patchy surfaces*, J. Chem. Phys., 121 (2004), pp. 11390–11394.

[6] M. I. Monine and J. M. Haugh, *Reactions on cell membranes: Comparison of continuum theory and Brownian dynamics simulations*, J. Chem. Phys., 123 (2005), 074908.

[7] Y. Song, Y. Zhang, T. Shen, C. L. Bajaj, J. A. McCammon, and N. A. Baker, *Finite element solution of the steady-state Smoluchowski equation for rate constant calculations*, Biophys. J., 86 (2004), pp. 2017–2029.

[8] R. Erban and J. Chapman, *Reactive boundary conditions for stochastic simulations of reaction-diffusion processes*, Phys. Biol., 4 (2007), pp. 16–28.

[9] S. Andrews and D. Bray, *Stochastic simulation of chemical reactions with spatial resolution and single molecule detail*, Phys. Biol., 1 (2004) pp. 137–151.

[10] W. Feller, *Diffusion processes in one dimension*, Trans. Amer. Math. Soc., 77 (1954), pp. 1–31.

[11] A. V. Skorokhod, *Stochastic equations for diffusion processes in a bounded region*, Theory Probab. Appl., 6 (1961), pp. 264–274.

[12] S. Asmussen, P. Glynn, and J. Pitman, *Discretization error in simulation of one-dimensional reflecting Brownian motion*, Ann. Appl. Probab., 5 (1995), pp. 875–896.

[13] D. Lépingle, *Euler scheme for reflected stochastic differential equations*, Math. Comput. Simulation, 38 (1995), pp. 119–126.

[14] C. Costantini, B. Pacchiarotti, and F. Sartoretto, *Numerical approximation for functionals of reflecting diffusion processes*, SIAM J. Appl. Math., 58 (1998), pp. 73–102.

[15] A. Marchewka and Z. Schuss, *Path integral approach to the Schrödinger current*, Phys. Rev. A, 61 (2000), 052107.

[16] A. Singer and Z. Schuss, *Brownian simulations and unidirectional flux in diffusion*, Phys. Rev. E, 71 (2005), 026115.

[17] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier, *An asymptotic theory of large deviations for Markov jump processes*, SIAM J. Appl. Math., 45 (1985), pp. 1006–1028.

[18] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier, *A singular perturbation approach to first passage times for Markov jump processes*, J. Stat. Phys., 42 (1986), pp. 169–184.

[19] C. Knessl, B. J. Matkowsky, Z. Schuss, and C. Tier, *Boundary behavior of diffusion approximations to Markov jump processes*, J. Stat. Phys., 45 (1986), pp. 245–266.

[20] E. Gobet, *Euler schemes and half-space approximation for the simulation of diffusion in a domain*, ESAIM Probab. Stat., 5 (2001), pp. 261–297.

[21] M. Bossy, E. Gobet, and D. Talay, *A symmetrized Euler scheme for an efficient approximation of reflected diffusions*, J. Appl. Probab., 41 (2004), pp. 877–889.

[22] F. C. Collins and G. E. Kimball, *Diffusion-controlled reaction rates*, J. Colloid Sci., 4 (1949), pp. 425–437.

[23] F. C. Goodrich, *Random walk with semiadsorbing barrier*, J. Chem. Phys., 22 (1954), pp. 588–594.

[24] A. F. Ghoniem and F. S. Sherman, *Grid free simulation of diffusion using random walk methods*, J. Comput. Phys., 61 (1985), pp. 1–37.

[25] S. H. Northrup, M. S. Curvin, S. A. Allison, and J. A. McCammon, *Optimization of Brownian dynamics methods for diffusion-influenced rate constant calculations*, J. Chem. Phys., 84 (1986), pp. 2196–2203.

[26] P. Szymczak and A. J. C. Ladd, *Stochastic boundary conditions to the convection-diffusion equation including chemical reactions at solid surfaces*, Phys. Rev. E, 69 (2004), 036704.

[27] N. J. B. Green, *On the simulation of diffusion processes close to boundaries*, Molec. Phys., 65 (1988), pp. 1399–1408.

[28] P. Szymczak and A. J. C. Ladd, *Boundary conditions for stochastic solutions of the convection-diffusion equation*, Phys. Rev. E, 68 (2003), 036704.

[29] M. Beccaria, G. Curci, and A. Vicere, *Numerical solutions of first-exit-time problems*, Phys. Rev. E, 48 (1993), pp. 1539–1546.

[30] J. Honerkamp, *Stochastic Dynamical Systems: Concepts, Numerical Methods, Data Analysis*, VCH, New York, 1994.

[31] R. Mannella, *Absorbing boundaries and optimal stopping in a stochastic differential equation*, Phys. Lett. A, 254 (1999), pp. 257–262.

[32] R. Mannella, *Integration of stochastic differential equations on a computer*, Internat. J. Modern Phys. C, 13 (2002), pp. 1177–1194.

[33] P. Clifford and N. J. B. Green, *On the simulation of the Smoluchowski boundary condition and the interpolation of Brownian paths*, Molec. Phys., 57 (1986), pp. 123–128.

[34] E. A. J. F. Peters and Th. M. A. O. M. Barenbrug, *Efficient Brownian dynamics simulation of particles near walls.* I. *Reflecting and absorbing walls*, Phys. Rev. E, 66 (2002), 056701.

[35] J. B. Keller and D. W. McLaughlin, *The Feynman integral*, Amer. Math. Monthly, 82 (1975), pp. 451–465.

[36] G. H. Bryan, *Note on a problem in the linear conduction of heat*, Proc. Camb. Phil. Soc., 7 (1891), pp. 246–248.

[37] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, 2nd ed., Oxford University Press, London, 1959.

[38] N. Agmon, *Diffusion with back reaction*, J. Chem. Phys., 81 (1984), pp. 2811–2817.

[39] M. R. von Smolan Smoluchowski, *Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen*, Phys. Zeits., 17 (1916), p. 557–585.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 869–889

A MATHEMATICAL MODEL FOR THE STEADY ACTIVATION OF A SKELETAL MUSCLE[*]

J.-P. GABRIEL[†], L. M. STUDER[‡], D. G. RÜEGG[§], AND M.-A. SCHNETZER[¶]

Abstract. A skeletal muscle is composed of motor units, each consisting of a motoneuron and the muscle fibers it innervates. The input to the motor units is formed of electrical signals coming from higher motor centers and propagated to the motoneurons along a network of nerve fibers. Because of its complexity, this network still escapes actual direct observations. The present model describes the steady state activation of a muscle, i.e., of its motor units. It incorporates the network as an unknown quantity and, given the latter, predicts the input-force relation (activation curve) of the muscle. Conversely, given a suitable activation curve, our model enables the recovery of the network. This step is performed by using experimental data about the activation curve, and the whole activation process of a muscle can then be theoretically investigated. In this way, this approach provides a link between the macroscopic (activation curve) and microscopic (network) levels. From a mathematical viewpoint, solving the preceding inverse problem is equivalent to solving an integral equation of a new type.

Key words. integral equation, muscle, physiology

AMS subject classifications. 45G10, 45D05, 92C30

DOI. 10.1137/05064271X

1. Introduction.

The activation of a muscle is a fascinating phenomenon involving complex and subtle physiological processes. Muscles responsible for voluntary motions are called striated (or skeletal) and are composed of motor units (MUs) consisting of a motoneuron (MN) together with the muscle fibers under its control. Depending on the muscle, the number of MUs can vary from ten up to several thousands. The MNs are located in the spinal chord and are connected to the central nervous system (CNS) through nerve fibers (input fibers) propagating the signal (input) in the form of electrical impulses called action potentials (APs).

When an AP reaches an MN through a (synaptic) contact, it modifies the electrical potential of its membrane, generating a so-called excitatory postsynaptic potential (EPSP). On a given MN, the effects of different APs are supposed to be additive. When the membrane potential reaches a specific threshold value, the MN starts generating APs which are transmitted to the muscle fibers and induce their contractions. As the activity of the input fibers increases, new stronger MUs are recruited (size principle). Additionally already active MUs enhance their forces, a process called frequency modulation [8, 9]. As soon as all MUs are recruited, frequency modulation is the only way for a muscle to increase its force. For simplicity reasons, we consider here only stationary isometric contractions and we assume that the total muscle force

[*]Received by the editors October 14, 2005; accepted for publication (in revised form) September 27, 2007; published electronically January 25, 2008.
http://www.siam.org/journals/siap/68-3/64271.html

[†]Department of Mathematics, University of Fribourg, Rue du Musée 5, 1700 Fribourg, Switzerland (jean-pierre.gabriel@unifr.ch).

[‡]Department of Computer Science and Economics, University of Applied Sciences, av. Max-Huber 6, 3960 Sierre, Switzerland (leo.studer@mus.ch).

[§]Department of Medicine, Division of Physiology, University of Fribourg, Rue du Musée 5, 1700 Fribourg, Switzerland (dieter.ruegg@unifr.ch).

[¶]Department of Mathematics, University of Applied Sciences, bd de Pérolles 82, 1700 Fribourg, Switzerland (marc-adrien.schnetzer@hefr.ch).

is given by the sum of all MU forces. The input-force relation of a muscle or its graph is alternatively called the *activation curve*.

MUs are ordered according to their maximal (tetanic) contraction forces t. The way input fibers activate MNs is extremely complex and the details of a corresponding network are not yet understood. On the level of an MN, this network, also called synaptic weight and denoted $g(t)$ (or sometimes g), is the missing quantity in our approach. Observations provide information about the shape of an activation curve. Our model is then used to recover g from this information (inverse problem). At this point, all aspects of the activation process of a muscle can be predicted.

For an arbitrary but fixed network, let $F(t)$ be the muscle force as a function of the last recruited MU. Clearly $F(t)$ has to be an increasing function of t. It will be seen that $g(t)$ can be deduced directly from $F(t)$ and it is thus sufficient to focus on the last function. $F(t)$ turns out to be a solution of an integral equation of the form

$$F(t) = \int_a^t k(s, F(s), F(t))\, ds, \tag{1}$$

and our task will be to solve (1). The presence of $F(t)$ in the kernel of (1) has striking consequences: on one hand this integral equation is not a classical Volterra equation and thus belongs to an extended type; on the other hand it admits infinitely many discontinuous solutions. Fortunately this equation has a unique continuous solution and this property turns out to be equivalent to increasingness. Since $F(t)$ is increasing in t, (1) admits one and only one physiologically meaningful solution. By using the latter, (1) can be reduced to a classical Volterra equation whose analytical and numerical properties are well known.

Analysis shows that the activation process has the following interesting properties:

(a) Similarly to the notion of the standard normal random variable in probability theory, there exists a standard hyperbolic muscle from which the activation curve of any muscle can be obtained.

(b) Despite the large number of parameters involved in the model, any activation curve is totally determined by a unique number called the activation factor.

(c) Several functionals related to the activation process can be introduced and characterized analytically, e.g., the ratio of the forces due to frequency modulation and recruitment.

All the published models of the MN pool-muscle complex (MNPMC) quantitatively describe the relation between the input to the MN pool and the muscle force [17, 18, 43, 12] or the electromyogram [12]. The models are similar in their structure but differ in the choice of the quantities given a priori. In the literature, all MNPMC models have been reduced to the following three unknowns: (1) the synaptic weight, (2) the MU population, and (3) the activation curve or the electromyogram. Since two of the unknowns can be given and the third one can be deduced from the model, we have three possible configurations: (1) and (2) are given and (3) is computed [17, 12]; (1) and (3) are given and (2) is computed [43]; and—the approach that we propose—the MU population (2) and the activation curve (3) are given and the synaptic weight (1) is determined. The main reason for this choice is that data at the level of the MNs are available for both unknowns (2) and (3) but not for unknown (1). As mentioned above, this configuration also leads to a new and interesting mathematical problem.

1.1. The model. We expose here the mathematical aspects of the model developed in [42], where a thorough discussion of the physiological hypotheses can be found. We focus mainly on equation (1), which is the key to the present investigation.

Since a typical skeletal muscle contains several hundred MUs, it is adequate to represent the MU population by a density function ρ. Choosing the tetanic contraction force t as variable, we get $t \in [t_{min}, t_{max}] \mapsto \rho(t)$, where t_{min} and t_{max} are the tetanic forces of the weakest and strongest MUs and the number of MUs in the pool is given by $\int_{t_{min}}^{t_{max}} \rho(s)\, ds$. All our considerations hold for all integrable and almost everywhere (a.e.) strictly positive functions ρ. We assume that t_{min} and t_{max} are given through ρ.

The global input In to the MNPMC is defined as the sum of all AP frequencies [43]. Each fiber contacts each MN of the pool and its activity induces EPSP conductance changes in the postsynaptic membrane. The EPSPs generated by single input fibers are smaller than 100 μV [15], the activity of these fibers is asynchronous [7], and the voltage threshold for APs is about 12 mV [5]. These three experimental findings imply that the variations of the membrane potential are smaller than 1% of the threshold voltage and are therefore neglected in the model. The EPSP conductances are thus represented by their time averages in the present time-independent approach. Due to the lack of precise information and for simplicity reasons, we assume that the EPSP conductance G_{EPSP} of each MN is proportional to the input:

$$G_{EPSP}(In) = g\, In, \tag{2}$$

where the MN-dependent factor g is the *synaptic weight*. This linearity assumption entails some restrictions on the MN connectivity [42].

The inactive (or subthreshold) MN is modeled with a single compartment and a homogeneous, electrically isolated membrane obeying Ohm's law. The total membrane current i_{tot} is the sum of the capacitive and ionic currents:

$$i_{tot} = C\, \dot{U} + \sum_k G_k\, (U - E_k), \tag{3}$$

where U is the membrane potential, $\dot{U}$ the time derivative of U, G_k the conductance of ion k, E_k its equilibrium potential, and C the membrane capacity. Three types of ionic conductances are distinguished: (a) a transmitter-sensitive conductance G_k^E caused by the synaptic input, (b) a voltage-dependent conductance G_k^U generating APs and E_k^U its equilibrium potential, and (c) a leakage conductance G_L and E_L its equilibrium potential. Clearly, we have $G_{EPSP} = \sum_k G_k^E$. For lack of data about particular MNs that are activated by synaptic input, we rely on data from current injection experiments [24]. The capacitive current is 0 in the steady state and since the membrane is isolated, the total current is equal to the injected current i_{inj}. Consequently, (3) can be rewritten $i_{inj} = \sum_k G_k^E (U - E_k) + \sum_k G_k^U (U - E_k^U) + G_L(U - E_L) = G_{EPSP}\, (U - \frac{\sum_k G_k^E\, E_k}{G_{EPSP}}) + \sum_k G_k^U (U - E_k^U) + G_L\, (U - E_L)$. Introducing the variable $V := U - E_L$, the reversal potential of the EPSP current $E_{EPSP} := \frac{\sum_k G_k^E\, E_k}{G_{EPSP}} - E_L$, and $E_k^V := E_k^U - E_L$, we then get

$$i_{inj} = G_{EPSP}\, (V - E_{EPSP}) + \sum_k G_k^U\, (V - E_k^V) + G_L\, V. \tag{4}$$

The current $i_{EPSP} := G_{EPSP}\, (V - E_{EPSP})$ will be called the *EPSP induced current.*

A subthreshold MN is inactive as long as it receives synaptic inputs without generating APs. During subthreshold depolarizations, there is a small increase of the sodium conductance which tends to depolarize the membrane and a small increase of the potassium conductance which tends to hyperpolarize the membrane. Since the two currents are in opposite directions, they tend to cancel each other (ultimately, this could be quantitatively derived from the Hodgkin–Huxley equations [22]). Therefore, we assume that the voltage-dependent channels are closed in subthreshold MNs, and consequently, the voltage-dependent currents are 0. Moreover, if $i_{inj} = 0$, the membrane potential corresponding to the EPSP is denoted V_{EPSP} and (4) becomes

$$G_{EPSP}\,(V_{EPSP} - E_{EPSP}) + G_L\,V_{EPSP} = 0. \tag{5}$$

According to (2) and (5), the membrane potential as a function of the input is

$$V_{EPSP}(In) = \frac{g\,In\,E_{EPSP}}{G_L + g\,In}. \tag{6}$$

The threshold input In_T is the maximum input an MN can receive in the subthreshold state. It evokes an EPSP equal to the firing threshold voltage V_T and therefore satisfies $V_{EPSP}(In_T) = V_T$. Assuming that all MNs of the pool have the same firing threshold [42], (6) at threshold provides

$$\tilde{g} = \frac{V_T}{(E_{EPSP} - V_T)\,In_T}, \tag{7}$$

where $\tilde{g} := \frac{g}{G_L}$ is called *relative synaptic weight*. Two quantities in (7) are MU-dependent, namely, the relative synaptic weight and the threshold input. According to (7), it is equally adequate to determine either one of them.

Fitting the frequency-injected current relations with a straight line with slope κ [42, 24], the MN firing frequency $\nu(G_{EPSP}, i_{inj})$ as a function of G_{EPSP} and i_{inj} is given by

$$\nu(0, i_{inj}) = \kappa\,(i_{inj} - i_T) + \nu_T \text{ if } i_{inj} \geq i_T \text{ and 0 otherwise,} \tag{8}$$

where ν_T is the *threshold frequency* and i_T the *threshold current*. Estimations of their values have been determined with current injection experiments on MNs [25].

The relation between the frequency evoked by a synaptic input and the injected current is not at all simple since i_{EPSP}, but not i_{inj}, depends on the membrane potential. We look for an injected current $i_{inj}(G_{EPSP})$ which evokes the same frequency as the synaptic input, namely, $\nu(G_{EPSP}, 0) = \nu(0, i_{inj}(G_{EPSP}))$.

Let us suppose that such a current exists for all values of G_{EPSP}: $i_{inj}(G_{EPSP}) = -i_{EPSP}$. Although the membrane potential of active MNs varies, we replace it by a constant virtual potential V_A which is independent of the firing frequency and is supposed to have similar effects as the time-varying membrane potential. In this approach, we set $V_A = V_T$ and we get $i_{inj}(G_{EPSP}) = G_{EPSP}\,(E_{EPSP} - V_T)$. By inserting (2) and (7) into the last equation, we obtain $i_{inj}(G_{EPSP}(In)) = \frac{G_L\,V_T\,In}{In_T}$. Since $In = In_T$ at threshold, the injected threshold current is $i_T = G_L\,V_T$.

The frequency-force relation of an MU during maintained contractions is built by fitting data obtained by injecting long-lasting currents of different intensities into MNs [26]. The MU force is given by

$$f(\nu) = t\,(1 - c\,\exp(-\gamma\,(\nu - \nu_T))), \tag{9}$$

where t is the MU tetanic contraction force, ν_T is its threshold frequency, and γ controls the shape of the curve. The number c determines the fraction of the tetanic force at recruitment according to $f(\nu_T) = t\,(1 - c)$. Since MUs are parameterized by their tetanic contraction forces t, the threshold input $In_T(t)$ is the smallest value necessary to recruit the MU with tetanic force t. Equation (8), $i_T = G_L\,V_T$, and (9) lead to the following representation of the MU transfer function:

$$f(t, In) := t\left(1 - c\,\exp\left(-\alpha\frac{In - In_T(t)}{In_T(t)}\right)\right) \text{ if } In > In_T(t) \text{ and 0 otherwise,} \tag{10}$$

where $\alpha := \gamma\,\kappa\,G_L\,V_T$.

In muscles with parallel fibers, the total muscle force is the sum of the contraction forces of its MUs, and thus its activation curve is $\mathcal{F}(In) = \int_{t_{min}}^{t_{max}} \rho(s)\,f(s, In)\,ds$. Figure 1 depicts the different steps leading from In to $\mathcal{F}(In)$.

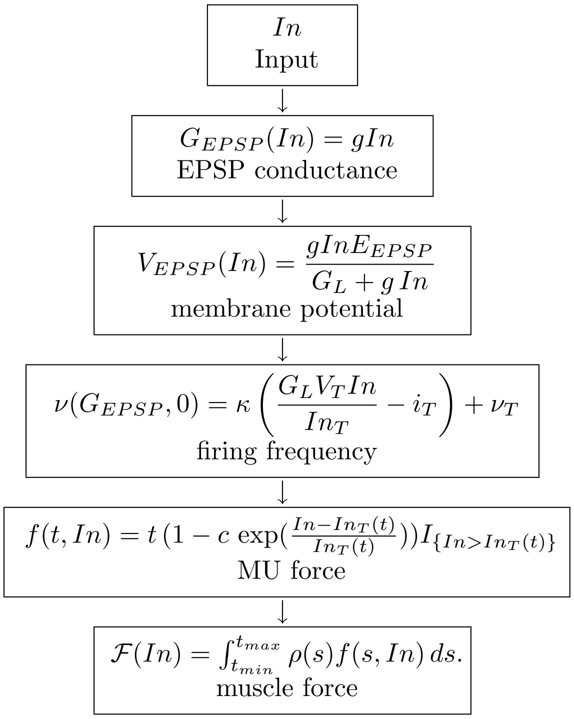

In
Input
$G_{EPSP}(In) = gIn$
EPSP conductance
$V_{EPSP}(In) = \frac{gInE_{EPSP}}{G_L + g\,In}$
membrane potential
$\nu(G_{EPSP}, 0) = \kappa\left(\frac{G_L V_T In}{In_T} - i_T\right) + \nu_T$
firing frequency
$f(t, In) = t\,(1 - c\,\exp(\frac{In - In_T(t)}{In_T(t)}))I_{\{In > In_T(t)\}}$
MU force
$\mathcal{F}(In) = \int_{t_{min}}^{t_{max}} \rho(s) f(s, In)\,ds.$
muscle force

FIG. 1. *Steps for the construction of the activation curve $\mathcal{F}(In)$.*

According to the size principle [21], MUs are recruited according to their tetanic contraction force during muscle activation; i.e., weaker MUs are recruited before stronger MUs. Therefore, the threshold input $In_T(t)$ activates all MUs with tetanic forces smaller than or equal to t. Replacing In by $In_T(t)$, we obtain

$$\mathcal{F}(In_T(t)) = \int_{t_{min}}^{t} \rho(s)\,f(s, In_T(t))\,ds. \tag{11}$$

Human subjects who superimposed ballistic contractions on background activities of different levels [38] provided information to determine the unknown function $In_T(t)$. The data suggest that the force generated by two inputs In_1 and In_2 is the sum of the forces induced by the single inputs. If In_0 denotes the minimal input necessary to recruit the smallest MU of the pool, we get the functional equation $\mathcal{F}(In_0 + In_1 + In_2) = \mathcal{F}(In_0 + In_1) + \mathcal{F}(In_0 + In_2)$. Its unique nonnegative solution with $\mathcal{F}(In_0) = 0$ is (see [1]) $\mathcal{F}(In) = k\,(In - In_0)$. This is the equation of a straight line with slope k. After recruitment, the activation curve is strictly concave. Since no additional force due to newly recruited MUs is available, it is an integral of strictly concave functions given in (10). Consequently, the affine ("linear" in [42]) relation can hold exclusively during recruitment. In a forthcoming paper, we will show that the possible activation curves are not limited to affine functions.

It should also be noted that the value of In corresponding to the end of recruitment is not specified a priori. This is a "free boundary" which will be determined by the model.

One could wonder whether a muscle response, which is not exactly additive, can be reasonably approximated by an additive function or not. The following results [29] shed some light on this question: a function $f : \mathbb{R} \to \mathbb{R}$ is called ϵ-additive if $|f(x+y) - f(x) - f(y)| \le \epsilon$ for all $x, y \in \mathbb{R}$. It can be seen that, if g is additive, any function f fulfilling $|f(x) - g(x)| \le \epsilon$ for all $x \in \mathbb{R}$ is 3ϵ-additive. Conversely, it can be proved that, for any ϵ-additive function f, there exists a unique additive function $g : \mathbb{R} \to \mathbb{R}$ such that $|f(x) - g(x)| \le \epsilon$, given by $g(x) = \lim_{n\to\infty} \frac{f(nx)}{n}$. As a consequence, if a muscle response $\mathcal{F}$ is only ϵ-additive, then there exists an additive function g contained in a band of width 2ϵ around $\mathcal{F}$. Moreover, if $\mathcal{F}$ is either bounded above or measurable, then g is continuous.

Replacing In by $In_T(t)$ as above, we get, for $t \in [t_{min}, t_{max}]$, $\mathcal{F}(In_T(t)) = k\,(In_T(t) - In_0)$, and with (10) and (11),

$$\mathcal{F}(In_T(t)) = \int_{t_{min}}^{t} h(s)\ \left(1 - c\ \exp\left(-\alpha \frac{\mathcal{F}(In_T(t)) - \mathcal{F}(In_T(s))}{\mathcal{F}(In_T(s)) + \Delta}\right)\right)\ ds, \tag{12}$$

where $h(t) = t\,\rho(t)$ is the force density function of the muscle and $\Delta = k\,In_0$.

Equation (12) contains the parameters α, c, Δ, and h [42]. Experimental observations suggest that α [24, 26, 2] and c [27] are MU-independent, and, in the present approach, we assume muscle independence. On the basis of experimental data, α was set to 1.14 and c to 0.9 [42], but the forthcoming general discussion is valid for all $\alpha > 0$, $0 < c < 1$, and $\Delta > 0$. A muscle is thus specified by Δ and h. In_0 cannot be measured experimentally, and in [42] it was assumed to be the same for all muscles. This assumption is however not required here.

2. The integral equation.

2.1. General considerations. Equation (12) is an integral equation for the unknown function $\mathcal{F}(In_T(t))$. By introducing the notation $F(t) = \mathcal{F}(In_T(t))$, (12) takes the form

$$F(t) = \int_{t_{min}}^{t} h(s)\ \left(1 - c\ \exp\left(-\alpha \frac{F(t) - F(s)}{F(s) + \Delta}\right)\right)\ ds, \qquad t \in [t_{min}, t_{max}]. \tag{13}$$

With our assumptions about ρ, the preceding integral has to be understood in the Lebesgue sense and, for obvious physiological reasons, we look for nonnegative solutions.

Equation (13) is not of Volterra type because its kernel

$$k(s, F(s), F(t)) := h(s)\ \left(1 - c\ \exp\left(-\alpha \frac{F(t) - F(s)}{F(s) + \Delta}\right)\right)$$

involves $F(t)$ and not only s, t, and $F(s)$. If (13) has a nonnegative locally bounded solution $F(t)$, $t \in [t_{min}, t_{max}]$, for $\alpha > 0$, $0 < c < 1$, and $\Delta > 0$, it may admit discontinuous solutions. Indeed, as a consequence of the dominated convergence theorem, for every sequence $(t_n)_{n\in\mathbb{N}}$, $t_n \in [t_{min}, t_{max}]$, with $\lim_{n\to\infty} t_n = t$ and $\lim_{n\to\infty} F(t_n) = \gamma$, we have $\lim_{n\to\infty} \int_{t_{min}}^{t_n} k(s, F(s), F(t_n))\, ds = \int_{t_{min}}^{t} k(s, F(s), \gamma)\, ds$. Thus (13) admits a locally bounded solution discontinuous at $t^* \in [t_{min}, t_{max}]$ if and only if the set $A_{t^*} = \{x \in \mathbb{R};\ x = \int_{t_{min}}^{t^*} k(s, F(s), x)\, ds\}$ contains an element $\beta \neq F(t^*)$. Indeed, if the preceding property holds, the function $F^*(t)$ taking the values β at $t = t^*$ and $F(t)$ elsewhere is obviously a discontinuous solution of (13), and the converse is clear. It can be seen, for example, numerically, that A_t contains two elements for t large enough.

2.2. The hyperbolic muscle. The presence of $F(t)$ in the kernel of (13) rules out most of the arguments of the classical theory of integral equations. It is not clear whether this equation admits a solution or not, and this point is important because, in our model, (13) governs the activation of a muscle.

Since ρ is strictly positive a.e., the function $H(t) = \int_{t_{min}}^{t} h(s)\, ds$, $t \in [t_{min}, t_{max}]$, is strictly increasing and hence invertible. Let us note that $H(t)$ is the force of the muscle when all MUs up to level t produce their tetanic force and $H(t_{max}) = F_{max}$.

By introducing $K(a,b) = 1 - c\exp(-\alpha\frac{b-a}{a+\Delta})$, $a, b \in \mathbb{R}_+ = [0, +\infty)$, (13) can be written $F(v) = \int_{t_{min}}^{v} K(F(u), F(v))\, h(u)\, du$. Since H is strictly increasing and absolutely continuous, the change of variable $u = H^{-1}(s)$ leads to $F(v) = \int_{H(t_{min})}^{H(v)} K(F(H^{-1}(s)), F(v))\, ds$ for $v \in [t_{min}, t_{max}]$. Without risk of confusion we write $H(v) = t$ and because $H(t_{min}) = 0$, $H(t_{max}) = F_{max}$, we obtain $F(H^{-1}(t)) = \int_0^t K(F(H^{-1}(s)), F(H^{-1}(t)))\, ds$ for $t \in [0, F_{max}]$. By defining $Y(t) := F(H^{-1}(t))$ and $T = F_{max}$, we get

$$(14) \qquad Y(t) = \int_0^t \left(1 - c\, \exp\left(-\alpha \frac{Y(t) - Y(s)}{Y(s) + \Delta}\right)\right) ds, \qquad t \in [0, T].$$

Since (14) describes the activation of a muscle whose MU density is the hyperbola $\rho(t) = \frac{1}{t}$ $(t > 0)$, it is natural to call it *hyperbolic*. Clearly, such a muscle has no physiological reality since any interval $(0, T]$ contains infinitely many of its MUs. This theoretical muscle is nevertheless interesting. By playing with the notation, we get

$$(15) \qquad F(t) = Y(H(t)), \quad t \in [t_{min}, t_{max}],$$

and we see that the force of an arbitrary muscle can be deduced from that of a hyperbolic one. It is therefore enough to study (14).

2.3. Existence and unicity of a physiological solution. Straightforward computations provide the following theorem.

THEOREM 1. *Let $0 < c < 1$, $\alpha > 0$, $\Delta > 0$, and $a, b \in \mathbb{R}_+ \mapsto K(a,b) = 1 - c\exp(-\alpha\frac{b-a}{a+\Delta})$.*

(a) *$K(a,b)$ together with the partial derivatives $K_a(a,b)$, $K_b(a,b)$, and $K_{bb}(a,b)$ are continuous and bounded.*

(b) *$K_a(a,b) < 0$, $K_b(a,b) > 0$, $K_{bb}(a,b) < 0$, and $K(a,b)$ is thus a concave function in the variable b.*

We first prove the existence and unicity of a nonnegative solution of (14) for small values of T. For $0 < T < \infty$, the set $E_T = \{f : [0,T] \mapsto \mathbb{R};\ f$ is measurable and bounded$\}$, equipped with the metric $d(f,g) := \sup_{t\in[0,T]} |f(t) - g(t)|$, is a complete metric space. Since uniform convergence preserves nonnegativity, $E_T^+ = \{f \in E_T; f \geq 0\}$ is a closed subset of E_T and hence also a complete metric space. Clearly, a solution of (14) is a fixed point of the operator $A_T f(t) = \int_0^t K(f(s), f(t))\, ds$, $t \in [0, T]$, $f \in E_T^+$, which maps E_T^+ into itself. According to Theorem 1, the mean value theorem can be applied to $K(a, b)$ and provides, for any $f, g \in E_T^+$: $d(A_T f, A_T g) \leq 2\, T\, M\, d(f, g)$, where $M := \sup_{a,b \in \mathbb{R}_+}(|K_a(a, b)|, |K_b(a, b)|)$. We conclude that A_T admits a unique fixed point for $T < \frac{1}{2\, M}$ since it is a contraction, and we prove the following theorem.

THEOREM 2. *For $T < \frac{1}{2\, M}$, (14) admits a unique nonnegative solution in E_T^+.*

Physiologically, a solution $Y(t)$ of (14) represents the force developed by a hyperbolic muscle, all of whose MUs, up to level t, are recruited. Therefore, Y has to be a nondecreasing function with $Y(0) = 0$, and any such solution will be called *physiological*. Since $0 < 1 - c \leq K(a, b) \leq 1$ for $0 \leq a \leq b$, we get $0 \leq Y(t) \leq t$ for $t \in [0, T]$, and Y is an element of $S_T := \{f \in E_T;\ f \text{ is nondecreasing, } 0 \leq f(t) \leq t,\ t \in [0, T]\}$. Clearly, $S_T \subset E_T^+$ and every element f of S_T satisfies $f(0) = 0$.

We now prove the existence of a physiological solution of (14) for an arbitrary $T > 0$ and start by recalling the following.

THEOREM OF SCHAUDER (see [41]). *Any nonempty compact and convex subset of a normed space has the fixed point property; i.e., every continuous mapping of such a subset into itself has at least one fixed point.*

THEOREM OF HELLY (see [6]). *Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of uniformly bounded, nondecreasing, and right continuous functions defined over an interval in $\mathbb{R}$. Then there exist a subsequence $(f_{n_k})_{k \in \mathbb{N}}$ and a nondecreasing right continuous function f such that $\lim_{k \to \infty} f_{n_k}(t) = f(t)$, for all continuity points t of f.*

THEOREM 3. *Every sequence in S_T contains a subsequence converging in the mean of order 1 to an element of S_T. Furthermore, S_T is nonempty and convex.*

Proof. For any sequence $(f_n)_{n \in \mathbb{N}}$ in S_T, the functions $t \in [0, T] \longmapsto f_n^+(t) = \lim_{h \downarrow 0} f_n(t + h)$ are also in S_T and right continuous. According to Helly's theorem, there exist a subsequence $(n_k)_{k \in \mathbb{N}}$ and a nondecreasing right continuous function f^+ such that $f_{n_k}(t)$ converges to $f^+(t)$, as $k \to \infty$, for every continuity point t of f^+ in $[0, T]$. Since f^+ is in S_T and the exceptional set is at most countable, the convergence takes place a.e. and hence in measure. Furthermore, the sequence, being uniformly bounded by T, is uniformly integrable, and we conclude to its convergence in the mean of order 1. Because nonemptyness and convexity of S_T are obvious, the proof is complete. □

Since Schauder's theorem requires a normed space, we introduce $E_T^* = E_T$ (mod a.e.) equipped with the L_1 norm. Let us recall that an element of E_T^* is an equivalence class of functions which are equal a.e. to a representative. Hence, S_T^* is the subset of E_T^* such that each equivalence class contains an element of S_T. S_T^* is nonempty and convex. Since every sequence in S_T^* has a corresponding sequence of representatives in S_T, S_T^* is also a compact subset of the normed space E_T^* as a consequence of the first part of Theorem 3.

Let f^* be the equivalence class in S_T^* of an element $f \in S_T$. Since $f = f'$ (a.e.) for elements in S_T implies that $A_T f = A_T f'$ (a.e.), we can define A_T^* over S_T^* with $A_T^* f^* = (A_T f)^*$ because it is independent of the representative.

THEOREM 4. *A_T^* is a continuous mapping of S_T^* into itself.*

Proof. Let f be a nondecreasing representative of f^* in S_T. Since $0 < 1 - c \leq K(a, b) \leq 1$ for $0 \leq a \leq b$, we get $0 \leq A_T f(t) \leq t$ for $t \in [0, T]$. Furthermore, $K(a, b)$ is nonnegative and increasing in the second variable for $a, b \geq 0$, and consequently, for $0 \leq t \leq t' \leq T$, we have $A_T f(t) = \int_0^t K(f(s), f(t))\, ds \leq \int_0^{t'} K(f(s), f(t'))\, ds =$

$A_T f(t')$. Therefore, $A_T f$ is an element of S_T and $A_T^* f^*$ belongs to S_T^* since a.e. equalities are preserved by A_T. To prove the continuity assertion, let us consider a sequence $(f_n^*)_{n\in\mathbb{N}}$ in S_T^* which converges to $\tilde{f}^*$ in L_1. Since S_T^* is closed as a compact subset of a Hausdorff space, we conclude that $\tilde{f}^*$ belongs to S_T^*. Furthermore, convergence in L_1 entails convergence in measure, and, for any σ-finite measure, the latter is equivalent to the statement "every subsequence contains a subsequence converging a.e." By using this property and the boundedness and continuity of $K(.,.)$, we get the desired result via the dominated convergence theorem. □

THEOREM 5. *For every* $T > 0$, (14) *admits a physiological solution.*

Proof. Schauder's theorem provides the existence of a fixed point f^* of A_T^* in S_T^*. For any nondecreasing representative f of the latter, we have $A_T f = f$ (a.e.) in $[0, T]$. The exceptional set E having measure 0, its complementary E^c is dense in $[0, T]$, and for every $t \in E^c$, one can define $f_-(t)$ as the left-hand limit of f over E^c since f is nondecreasing and bounded. The function $\tilde{f}(t) := \lim_{s\uparrow t} f(s)$ is nonnegative and nondecreasing. By using the boundedness and continuity of $K(.,.)$ and the dominated convergence theorem, one can show that $\tilde{f}$ satisfies $A_T\tilde{f}(t) = \tilde{f}(t)$ for every t in $[0, T]$, and the proof is complete. □

We prove that increasingness and continuity are equivalent properties for solutions of (14).

LEMMA 6. *Let* $f : [a, b] \mapsto \mathbb{R}$ *be continuous and* $f(a) = f(b)$. *For every* $\delta > 0$, *there exist* $a \le t < t' \le b$ *with* $0 < t' - t < \delta$ *and* $f(t) = f(t')$.

Proof. The property clearly holds if f is constant. Otherwise, replacing f by $-f$ if necessary, we can assume that f admits a maximum $M \neq f(a)$ at t^* in the open interval (a, b). Choosing $\lambda < M$ sufficiently close to M, the intermediate value theorem for continuous functions implies the existence of $a < t < t^* < t' < b$, such that $f(t) = f(t') = \lambda$ and t, t' are as close as we want to t^*. Let us note that if f is not constant, it is even possible to choose t and t' so that $f(s) \neq f(a)$ for all $s \in [t, t']$. □

THEOREM 7. *A continuous solution of* (14) *starting at* 0 *is strictly increasing.*

Proof. We first show that strict positivity of a continuous solution Y of (14) over an interval $J \subset (0, T]$ entails its injectivity. Assuming the contrary, one can find $a < b$ in J with $Y(a) = Y(b)$. According to the preceding lemma, there exist $t < t'$ in J, as close as we want, with $Y(t) = Y(t') = \lambda > 0$ and, from (14), we deduce $0 = Y(t') - Y(t) = \int_t^{t'} K(Y(s), \lambda)\, ds$. Since $K(\lambda, \lambda) > 0$ and K is continuous over $\mathbb{R}_+^2$, $K(Y(s), \lambda) > 0$ over $[t, t']$ for t, t' sufficiently close to each other. Therefore, the preceding integral is positive and we get a contradiction.

To prove the theorem, we observe that the form of the kernel K implies the existence of a neighborhood of $(0, 0)$ in $\mathbb{R}^2$ over which $K(.,.) > 0$. Since $Y(0) = 0$ and Y is continuous, there exists $0 < t_0 < T$, such that for all $0 \le s \le t \le t_0$, $K(Y(s), Y(t)) > 0$ and hence $Y(t) > 0$ over $(0, t_0]$. Consequently, Y is injective and hence increasing over $[0, t_0]$. A first zero of Y after t_0 would violate its injectivity over an interval of positivity, and the proof is complete. □

The following lemma will be used to discuss the converse of the preceding result. In what follows, we shall use the notation $f_x := \frac{\partial f}{\partial x}$ for a function f.

LEMMA 8. *Let* Y, $Y(0) = 0$, *be a nondecreasing solution of* (14) *defined over some interval* $[0, T]$ *and* $G(a, b) := \int_0^a K(Y(s), b)\, ds$ *for* $(a, b) \in [0, T] \times \mathbb{R}$. *The function* G *has the following properties:* (a) *For every fixed* $a \in [0, T]$, $G(a, b)$ *is a strictly increasing and concave function of* b. (b) $G(a, b)$ *is continuous.* (c) $G_b(a, b)$

is continuous, and the same property holds for $G_a(a,b)$ *if* Y *is continuous.* (d) *If* Y *is continuous, so are* $G(t,Y(t))$ *and* $G_b(t,Y(t))$ *as functions of* t.

Proof. Strict monotonicity in (a) is obvious and concavity follows from the fact that it is preserved by integration. Properties (b) and (c) can be deduced from the dominated convergence theorem and suitable estimations, and (d) is obvious. □

THEOREM 9. *Any nondecreasing solution of* (14) *starting at* 0 *is continuously differentiable.*

Proof. Let $\tilde{Y}$ defined over $[0,T]$ ($\tilde{Y}(0)=0$) be a nondecreasing solution of (14) and Y the function of its right-hand limits. Y is nondecreasing and right continuous, and, by a density argument, continuity of $\tilde{Y}$ and Y are equivalent. Moreover, the dominated convergence theorem provides $Y(t)=\int_0^t K(\tilde{Y}(s),Y(t))\,ds$ and since $\tilde{Y}(t)=Y(t)$ (a.e.), $Y(t)$ is also a solution of (14). Because of the concavity property (a) in Lemma 8, for every $t\in[0,T]$, $G(t,y)=y$ admits at most two solutions, one of them being $Y(t)$ since $G(t,Y(t))=Y(t)$. Moreover, if for a given t the preceding equation has two distinct solutions $y_1(t)<y_2(t)$, then $G_b(t,y_2(t))<1$ and $G_b(t,y_1(t))>1$.

According to Theorem 2, over a sufficiently small right-hand interval of 0, (14) admits a unique nonnegative solution in the space of bounded measurable functions. As a consequence, $Y(t)$ is the unique solution of $G(t,y)=y$ for t small enough. The function G satisfies all regularity conditions of the implicit function theorem [41]. Since $G(0,0)=0$ and $G_b(0,0)=0<1$, the latter entails the existence of a unique continuous function y such that $y(0)=0$ and $G(t,y(t))-y(t)=0$ over a right-hand neighborhood of 0. The preceding unicity property implies $Y\equiv y$ for t small enough and continuity entails $G_b(t,Y(t))<1$ over some right-hand interval of 0. Hence there is an interval $[0,\beta)$ of maximal length β over which $G_b(t,Y(t))<1$.

If Y is not continuous over $[0,\beta)$, then there is a point of discontinuity $t_0\neq 0$ because Y is continuous for small values of t. Y is right continuous and since it is increasing, its left and right-hand limits at t_0 satisfy $Y(t_0^-)<Y(t_0^+)=Y(t_0)$. According to the implicit function theorem, there exists a unique continuous function y_0 defined over an open interval I_{t_0} containing t_0, such that $y_0(t_0)=Y(t_0)$ and $G(t,y_0(t))=y_0(t)$ for $t\in I_{t_0}$. Since $G_b(t_0,Y(t_0))<1$, the continuity of y_0 implies the existence of $t<t_0$ in I_{t_0}, with $G_b(t,y_0(t))<1$ and $Y(t_0^-)<y_0(t)$. Because $Y(t)\leq Y(t_0^-)$, we obtain the configuration $G(t,Y(t))=Y(t)$, $G(t,y_0(t))=y_0(t)$, $G_b(t,Y(t))<1$, $G_b(t,y_0(t))<1$, and $Y(t)\neq y_0(t)$. This is a contradiction of the concavity of $G(t,b)$ in the variable b, and Y is continuous over $[0,\beta)$ and thus identical to $\tilde{Y}$.

The function $G_b(a,b)$ is continuous and, because of (c) in Lemma 8, the continuity of Y implies that of $G_a(a,b)$. Consequently, G is continuously differentiable and, according to another version of the implicit function theorem [10, 11], every locally defined continuous solution y provided by the last theorem is also differentiable. Since Y is continuous over $[0,\beta)$, local unicity implies local identity of y and Y and entails the differentiability of the latter over $[0,\beta)$.

In the following considerations, every interchange of integration and derivation can be justified with the dominated convergence theorem. For $t\in[0,T]$ we introduce $\phi(t):=G_b(t,Y(t))=\int_0^t\frac{\alpha\, c}{Y(s)+\Delta}(\exp(-\alpha\frac{Y(t)-Y(s)}{Y(s)+\Delta}))\,ds$. Writing $'=\frac{d}{dt}$ and using the differentiability of Y, for every $t\in[0,\beta)$ we get

$$\phi'(t)=K_b(Y(t),Y(t))+\frac{1-c}{1-\phi(t)}\int_0^t K_{bb}(Y(s),Y(t))\,ds. \tag{16}$$

We now suppose that $\beta<T$ and claim that $\sup_{t\in[0,\beta)}\phi(t)=m<1$. If not, $m=1$ as a consequence of the definition of β, and by using the mean value theorem of differential calculus, one can choose an increasing sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \uparrow \beta$ as $n \to \infty$ and $\phi'(t_n) \geq 0$ for every $n \in \mathbb{N}$. The first term in (16) is positive and bounded, and the integral is negative and bounded away from 0. Since $0 < c < 1$, the quotient before the integral tends to $+\infty$ as $t \uparrow \beta$. According to (16), $\phi'(t_n) < 0$ for n sufficiently large, and we get a contradiction and conclude that $m < 1$.

Because of the monotonicity of Y, $Y(t) \leq Y(\beta)$ for all $t \in [0, \beta)$, and thus, for any sequence $u_n \uparrow \beta$, we have $\int_0^{u_n} \frac{\alpha\, c}{Y(s)+\Delta}(\exp(-\alpha\frac{Y(\beta)-Y(s)}{Y(s)+\Delta}))\, ds \leq \phi(u_n) \leq m$. Letting $n \to \infty$ in the integral, we get $\phi(\beta) \leq m < 1$ and thus $G_b(\beta, Y(\beta)) < 1$. Applying once more the implicit function theorem, we get a unique continuous function y^+ defined over an open interval I_β containing β with $y^+(\beta) = Y(\beta)$, $G(t, y^+(t)) = y^+(t)$ over I_β, and $G_b(t, y^+(t)) < 1$ for every t sufficiently close to β. If $Y \equiv y^+$ does not hold over a right-hand open neighborhood of β, then one can find a sequence $v_n \downarrow \beta$ as $n \to \infty$, such that $Y(v_n) \neq y^+(v_n)$ for every n. Monotonicity and concavity entail $Y(\beta) \leq Y(v_n) < y^+(v_n)$, and hence $Y(v_n) \to Y(\beta)$ as $n \to \infty$. Concavity again implies that $G_b(v_n, Y(v_n)) > 1$, and thus, letting $n \to \infty$, $G_b(\beta, Y(\beta)) \geq 1$, which is a contradiction. Since $Y \equiv y^+$ on a right-hand open interval of β, $G_b(t, Y(t)) < 1$ for all sufficiently small t to the right of β, which is a contradiction of its maximality. Consequently $\beta = T$ and Y is continuous over $[0, T]$. Moreover, Y is also differentiable, and the relations $Y'(t) = \frac{1-c}{1-\phi(t)} = \frac{1-c}{1-G_b(t,Y(t))}$ show that it is even continuously differentiable and $Y'(t) > 0$ over $[0, T]$. □

We would like to stress that, in Theorem 9, the condition $0 < c < 1$ plays a delicate role. Indeed, let us assume that $c = 1$ and look for a solution of (14) of the form $Y(t) = a\, I_{(t^*,+\infty)}(t)$, where I denotes the indicator function and t^* and a have to be specified. For $t \leq t^*$, (14) is satisfied since both sides are equal to 0. For $t > t^*$, it reduces to $a = t^*\,(1 - \exp(-\frac{\alpha\, a}{\Delta}))$. By choosing t^* so that $\frac{t^*\alpha}{\Delta} > 1$ and the positive solution of the preceding equation for a, we get a discontinuous nondecreasing solution of (14).

Collecting all the preceding results, we get the following theorem.

THEOREM 10. *The following properties are equivalent for a solution Y of* (14) *with $Y(0) = 0$:* (a) *Y is physiological, i.e., nondecreasing,* (b) *Y is strictly increasing,* (c) *Y is continuous, and* (d) *Y is continuously differentiable and Y' is strictly positive.*

It is interesting to note that although (14) admits possibly discontinuous solutions, any physiological solution is automatically continuously differentiable.

THEOREM 11. *Equation* (14) *admits a unique physiological solution.*

This uniqueness result can be deduced from Gronwall's inequality and the fact that, according to the proof of Theorem 9, $\phi(t) < 1$ over $[0, T]$. We propose another argument which will also provide a nice way to get the solution of (14).

Proof. We proved that every nondecreasing solution Y of (14) over $[0, T]$ is strictly increasing and continuously differentiable with a strictly positive and bounded derivative. Consequently, $X = Y^{-1}$ has the same properties as Y and $Y(X(t)) = t$ for every $t \geq 0$. In particular, the derivative of X, denoted x, is continuous. Performing the change of variables $s = X(u)$ and $v = X(t)$ in (14) and rewriting s and t instead of u and v, we get

$$t = \int_0^t x(s)\left(1 - c\,\exp\left(-\alpha\frac{t-s}{s+\Delta}\right)\right) ds. \tag{17}$$

Equation (17) is a linear Volterra equation of the first kind for x, the derivative of X (inverse function of Y). Since the kernel of (17) satisfies all the conditions of Theorem

1.3.5 in [4], this equation admits a unique continuous solution, and thus Theorem 11 is proved. □

It is interesting to note that (17) has an interpretation. Indeed, it is an integral equation for the force density function x of a muscle whose force is t if its MUs up to level t are recruited.

For continuous x, derivation and integration by parts of (17) provide, respectively,

$$x(t) = \frac{1}{1-c} - \frac{\alpha\, c}{1-c}\int_0^t x(s)\frac{1}{s+\Delta}\exp\left(-\alpha\frac{t-s}{s+\Delta}\right) ds \quad \text{and} \tag{18}$$

$$X(t) = \frac{t}{1-c} - \frac{\alpha\, c}{1-c}\int_0^t X(s)\frac{t+\Delta}{(s+\Delta)^2}\exp\left(-\alpha\frac{t-s}{s+\Delta}\right) ds. \tag{19}$$

The last two equations are linear Volterra equations of the second kind, for which theory and numerical treatments are well known [4, 14, 32].

Let us remark that existence and unicity of Y for every $T > 0$ automatically provide a unique solution of (14) defined over $[0,\infty)$.

We would like to point out that Theorem 10 together with the preceding arguments also lead to the existence of an increasing and continuously differentiable solution of (14). However, Schauder's theorem provides the existence of a nondecreasing solution without continuity assumption and thus without Theorem 10. Both approaches are interesting, and methodological diversity is always welcome for the treatment of future investigations.

3. Properties of the physiological solution.

In the following, x will be the unique continuous solution of (17) defined over $[0,+\infty)$ and $X(t) = \int_0^t x(s)\,ds$. As can be seen in (18), x depends on the parameters α, c, and $\Delta = k\,In_0$. Since α and c are fixed and muscle-independent, x, X, and Y will be considered as functions of the two variables $t \geq 0$ and $\Delta > 0$ and written $x(t,\Delta)$, $X(t,\Delta)$, and $Y(t,\Delta)$.

THEOREM 12. *For every fixed $\Delta > 0$, the following properties hold:* (a) $x(t,\Delta)$ *is continuously differentiable in the variable t,* (b) $x(0,\Delta) = \frac{1}{1-c}$, (c) $0 < x(t,\Delta) \leq (1-c)^{-1}$ *for every $t > 0$,* (d) $0 < X(t,\Delta) \leq \frac{t}{1-c}$ *for every $t > 0$,* (e) *as a function of t, $\frac{X(t,\Delta)}{t}$ is decreasing over $(0,\infty)$, and* (f) *as a function of u, $\frac{Y(u,\Delta)}{u}$ is increasing over $(0,\infty)$.*

Proof. Part (a) follows from the facts that (18) is again differentiable in t and that the result involves only continuous functions. Since the integrand in (18) is bounded, letting $t \to 0$, we get (b). The second inequality in (c) follows from (18) and $x(t,\Delta) = X_t(t,\Delta) > 0$. Integrating (c) provides (d). To prove (e), we have to show that $\frac{\partial}{\partial t}(\frac{X(t,\Delta)}{t}) = \frac{1}{t}\,(x(t,\Delta) - \frac{X(t,\Delta)}{t}) \leq 0$, that is, $\frac{X(t,\Delta)}{t} \geq x(t,\Delta)$ for $t > 0$. We divide (19) by t and obtain $\frac{X(t,\Delta)}{t} = \frac{1}{1-c} + \int_0^t K(s,t)\,\frac{s}{t}\frac{t+\Delta}{s+\Delta}\frac{X(s,\Delta)}{s}\,ds = \frac{1}{1-c} + \int_0^t \tilde{K}(s,t)\,\frac{X(s,\Delta)}{s}\,ds$, where $K(s,t) := -\frac{\alpha\, c}{1-c}\frac{1}{s+\Delta}\exp(-\alpha\frac{t-s}{s+\Delta})$ is the kernel of (18). We observe that $\tilde{K}(s,t) = K(s,t)\frac{s}{t}\frac{t+\Delta}{s+\Delta}$. A straightforward computation based on kernel iterations shows that the corresponding resolvents $R(s,t)$ and $R'(s,t)$ also satisfy $R'(s,t) = R(s,t)\frac{s}{t}\frac{t+\Delta}{s+\Delta}$. Obviously for $\Delta > 0$ and $0 < s \leq t$, we have $\frac{s}{t}\frac{t+\Delta}{s+\Delta} < 1$, and according to the general theory [4], we can write $\frac{X(t,\Delta)}{t} = (1-c)^{-1} + (1-c)^{-1}\int_0^t R'(s,t)\,ds = (1-c)^{-1} + (1-c)^{-1}\int_0^t R(s,t)\frac{s}{t}\frac{t+\Delta}{s+\Delta}\,ds \geq (1-c)^{-1} + (1-c)^{-1}\int_0^t R(s,t)\,ds = x(t,\Delta)$. The last inequality clearly holds if $R(s,t) \leq 0$, and it is enough to prove the latter. The resolvent formula [4] provides $-R(s,t) = \frac{\alpha\, c}{1-c}\frac{1}{s+\Delta}\exp(-\alpha\frac{t-s}{s+\Delta}) + \int_s^t R(s,u)\,\frac{\alpha\, c}{1-c}\frac{1}{u+\Delta}\exp(-\alpha\frac{t-u}{u+\Delta})\,du$. Applying the changes of variables $u = v - \Delta$ and then $v = \exp(w)$ and rewriting $t' = \ln(t + \Delta)$, $s' = \ln(s + \Delta)$, for fixed s', $-R'(s', t') = -R(\exp(s') - \Delta, \exp(t') - \Delta)$ is a solution of the convolution-type integral equation $-R'(s', t') = f(t') + \int_{s'}^{t'} R'(s', w)\, h(t' - w)\, dw$, where $f(t') = \frac{\alpha\, c}{1-c} \frac{1}{\exp(s')} \exp(-\alpha \frac{\exp(t') - \exp(s')}{\exp(s')})$ and $h(t') = \frac{\alpha\, c}{1-c} \exp(\alpha - \alpha \exp(t'))$. Since $\frac{f(T)}{f(t)} = \exp(-\alpha\, e^{-s'}\, (e^T - e^t)) \le \exp(-\alpha\, e^{-s}(e^T - e^t)) = \frac{h(T-s)}{h(t-s)}$ for $0 \le s' \le s < T < t$, according to Theorem 6.1 in [32], we conclude to $R'(s', t') \le 0$.

For $t_1 = Y(u_1, \Delta)$ and $t_2 = Y(u_2, \Delta)$, $t_1 < t_2$ implies $u_1 < u_2$. By using (e) and $Y(t, \Delta) = X^{-1}(t, \Delta)$ in the variable t, we get $\frac{u_1}{Y(u_1,\Delta)} = \frac{X(Y(u_1,\Delta),\Delta)}{Y(u_1,\Delta)} = \frac{X(t_1,\Delta)}{t_1} \ge \frac{X(t_2,\Delta)}{t_2} = \frac{X(Y(u_2,\Delta),\Delta)}{Y(u_2,\Delta)} = \frac{u_2}{Y(u_2,\Delta)}$, and (f) is proved. □

THEOREM 13. *For every* $\lambda > 0$, $t \ge 0$, $u \ge 0$, *and* $\Delta > 0$, *we have* (a) $x(t, \lambda\Delta) = x(\frac{t}{\lambda}, \Delta)$, (b) $x(\lambda t, \Delta) = x(t, \frac{\Delta}{\lambda})$, (c) $x(\lambda t, \lambda\Delta) = x(t, \Delta)$, (d) $X(\lambda t, \lambda\Delta) = \lambda X(t, \Delta)$, (e) $Y(\lambda u, \lambda\Delta) = \lambda Y(u, \Delta)$, *and* (f) $Y_\Delta(u, \Delta) + u Y_u(u, \Delta) = Y(u, \Delta)$.

Proof. For every $\lambda > 0$, (17) yields $t = \int_0^t x(s, \Delta)(1 - c \exp(-\alpha \frac{\lambda t - \lambda s}{\lambda s + \lambda\Delta}))\, ds$. By using the change of variable $s = \frac{z}{\lambda}$ and ultimately replacing $\lambda\, t$ by t, we get $t = \int_0^t x(\frac{z}{\lambda}, \Delta)(1 - c \exp(-\alpha \frac{t-z}{z+\lambda\Delta}))\, dz$. However, according to (17), we also have $t = \int_0^t x(z, \lambda\Delta)(1 - c \exp(-\alpha \frac{t-z}{z+\lambda\Delta}))\, dz$. Since (17) admits a unique continuous solution, we conclude that $x(z, \lambda\, \Delta) = x(\frac{z}{\lambda}, \Delta)$ for all possible values of their arguments. This is equivalent to (a) and (b), and writing $z = \lambda\, s$ yields (c). Part (d) follows from $X(\lambda t, \lambda\Delta) = \int_0^{\lambda t} x(\frac{z}{\lambda}, \Delta) dz = \int_0^{\lambda t} x(z, \lambda\Delta) dz = \lambda \int_0^t x(s, \Delta) ds = \lambda X(t, \Delta)$. By using Theorem 13(d) and the fact that, for fixed Δ, Y is the inverse function of X, we can write $\lambda t = Y(X(\lambda t, \lambda\Delta), \lambda\Delta) = Y(\lambda X(t, \Delta), \lambda\Delta) = \lambda Y(X(t, \Delta), \Delta)$. The equality $Y(\lambda X(t, \Delta), \lambda\Delta) = \lambda Y(X(t, \Delta), \Delta)$ is valid for every $t \ge 0$. Because $X(t, \Delta)$ is strictly increasing in t, we substitute $u = X(t, \Delta)$, and (e) is proved. According to Theorem 13(e), $Y(u, \Delta) = \Delta Y(\frac{u}{\Delta}, 1)$, and the differentiability of Y with respect to u entails that with respect to Δ.

For any differentiable function $F(u)$, $\frac{d}{dv} F(v)|_{\lambda u} = \frac{1}{\lambda} \frac{d}{du} F(\lambda u)$, and thus $Y_v(v, 1)|_{\frac{u}{\Delta}} = \Delta Y_u(\frac{u}{\Delta}, 1) = Y_u(u, \Delta)$. By using Theorem 13(e) again, $Y_\Delta(u, \Delta) = \frac{\partial}{\partial\Delta}(\Delta Y(\frac{u}{\Delta}, 1)) = Y(\frac{u}{\Delta}, 1) + \Delta(-\frac{u}{\Delta^2}) Y_v(v, 1)|_{\frac{u}{\Delta}}$ and $Y_\Delta(u, \Delta) = \frac{1}{\Delta} Y(u, \Delta) - \frac{u}{\Delta} Y_u(u, \Delta)$. The last equality is equivalent to (f). The latter, which is the differential equivalent of (e), provides a partial differential equation for Y. Unfortunately, the initial condition is given along a characteristic curve and is equivalent to a solution of (17). □

THEOREM 14. *Let* $x_\infty := (1 - c \int_0^1 \exp(-\alpha \frac{1-s}{s})\, ds)^{-1}$.

(a) *For every* $t \ge 0$, $\lim_{\Delta\uparrow\infty} x(t, \Delta) = \frac{1}{1-c}$.

(b) *For every* $t > 0$, $\lim_{\Delta\downarrow 0} x(t, \Delta) = x_\infty$ *and* $\lim_{\Delta\downarrow 0} x(0, \Delta) = \frac{1}{1-c}$.

(c) *For every* $\Delta > 0$, $\lim_{t\uparrow\infty} x(t, \Delta) = x_\infty$.

Proof. By using the continuity of x, $x(0, \Delta) = \frac{1}{1-c}$, and Theorem 13(a), we get $\frac{1}{1-c} = \lim_{t\downarrow 0} x(t, \Delta) = \lim_{\lambda\uparrow\infty} x(\frac{t}{\lambda}, \Delta) = \lim_{\lambda\uparrow\infty} x(t, \lambda\Delta) = \lim_{\Delta\uparrow\infty} x(t, \Delta)$, and (a) is proved.

It is enough to consider (b) for $t > 0$. For every $0 < t < t^*$ and $\Delta > 0$, according to Theorem 13(c), we have $x(t, \Delta) = x(\frac{t}{t^*} t^*, \frac{t}{t^*} \frac{t^*}{t} \Delta) = x(t^*, \frac{t^*}{t} \Delta)$. Let $\Delta_n \downarrow 0$ and $x(t, \Delta_n) \to \liminf_{\Delta\downarrow 0} x(t, \Delta)$ as $n \uparrow \infty$. The preceding equality entails $\liminf_{\Delta\downarrow 0} x(t, \Delta) = \lim_{n\uparrow\infty} x(t, \Delta_n) = \lim_{n\uparrow\infty} x(t^*, \frac{t^*}{t} \Delta_n) \ge \liminf_{\Delta\downarrow 0} x(t^*, \Delta)$. Since for the same reasons the converse inequality also holds, we conclude that $\liminf_{\Delta\downarrow 0}\ x(t, \Delta) = \liminf_{\Delta\downarrow 0} x(t^*, \Delta)$. Hence, for $t > 0$, $\liminf_{\Delta\downarrow 0} x(t, \Delta)$ is independent of t, and the same property holds for $\limsup_{\Delta\downarrow 0} x(t, \Delta)$. By using (Δ_n) such that $\lim_{n\to\infty} x(t, \Delta_n) = \underline{x}$, (17) and the dominated convergence theorem provide

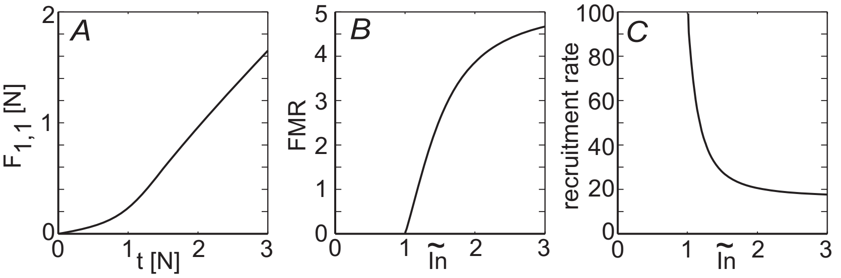


FIG. 2. A: *Force of the standard hyperbolic muscle as a function of the tetanic force of the last recruited MU.* B: *Frequency modulation to recruitment ratio (FMR).* C: *Recruitment rate.*

$t = \int_0^t \underline{x}\,(1 - c\exp(-\alpha\frac{t-s}{s}))\,ds$. Replacing s by $s\,t$ in the preceding integral leads to $\underline{x} = (1 - c\int_0^1 \exp(-\alpha\frac{1-s}{s})\,ds)^{-1} = x_\infty$. Since the same argument holds for $\overline{x} = \limsup_{\Delta\downarrow 0} x(t,\Delta)$, we conclude that $\underline{x} = \overline{x} = x_\infty$, and the proof of (b) is complete.

(c) follows from (b), with $\lim_{t\uparrow\infty} x(t,\Delta) = \lim_{\lambda\downarrow 0} x(\frac{t}{\lambda},\Delta) = \lim_{\lambda\downarrow 0} x(t,\lambda\Delta) = \lim_{\Delta\downarrow 0} x(t,\Delta) = x_\infty$. □

The existence of the limits in the next theorem follows from the monotonicity properties discussed above.

THEOREM 15. *For every* $\Delta > 0$, *we have* (a) $\lim_{t\uparrow\infty} \frac{X(t,\Delta)}{t} = x_\infty$, (b) $\lim_{t\downarrow 0} \frac{X(t,\Delta)}{t} = \frac{1}{1-c}$, (c) $\lim_{u\uparrow\infty} \frac{Y(u,\Delta)}{u} = \frac{1}{x_\infty}$, *and* (d) $\lim_{u\downarrow 0} \frac{Y(u,\Delta)}{u} = 1 - c$.

Proof. (a) follows from Theorem 14(a) and the fact that ordinary convergence implies convergence in the Cesaro sense. Since $X(0,\Delta) = 0$, the definition of the right-hand derivative at 0 provides (b). Part (d) follows from the same argument applied to the inverse function Y. To prove (c), for fixed Δ we substitute $t = Y(u,\Delta)$ in (a) and use $x_\infty = \lim_{t\uparrow\infty} \frac{X(t,\Delta)}{t} = \lim_{u\uparrow\infty} \frac{X(Y(u,\Delta),\Delta)}{Y(u,\Delta)} = \lim_{u\uparrow\infty} \frac{u}{Y(u,\Delta)}$. □

3.1. Representation of the physiological solution. The use of Theorem 13 requires caution with units. For simplicity of notation, we introduce two rules:

- The argument of $F_{(h,\Delta)}$ and $\tilde{F}_{(h,\Delta)}$ is always expressed in Newtons.
- The presence or absence of units attributed to $F_{(h,\Delta)}$ and Y is imposed by the context.

Since α and c are muscle-independent, the solution F of (13),

$$F(t) = \int_{t_{min}}^{t} h(s)\,\left(1 - c\,\exp\left(-\alpha\frac{F(t) - F(s)}{F(s) + \Delta}\right)\right)\,ds, \qquad t \in [t_{min}, t_{max}],$$

depends only on h and Δ and is therefore denoted $F_{(h,\Delta)}$. Recall that a muscle with $h \equiv 1$ was called hyperbolic, and with the new notation we have $F_{(1,\Delta)}(t) = Y(t,\Delta)$ for all $t \geq 0$. By using Theorem 13(e) and (15), we get $F_{(h,\Delta)}(t) = Y(H(t),\Delta) = \Delta Y(\frac{H(t)}{\Delta},1) = \Delta F_{(1,1)}(\frac{H(t)}{\Delta})$. $F_{(1,1)}$ is the solution of (13) for $h \equiv 1$ and $\Delta = 1$, a muscle which will be called *standard hyperbolic* (Figure 2A). Introducing the relative force $\tilde{F}_{(h,\Delta)}(t) := \frac{F_{(h,\Delta)}(t)}{F_{max}}$, we get

$$\tilde{F}_{(h,\Delta)}(t) = \frac{\Delta}{F_{max}} F_{(1,1)}\left(\frac{H(t)}{\Delta}\right) = \frac{\Delta}{F_{max}} F_{(1,1)}\left(\frac{\frac{H(t)}{F_{max}}}{\frac{\Delta}{F_{max}}}\right) = A F_{(1,1)}\left(\frac{\tilde{H}(t)}{A}\right),$$

where $\tilde{H}(t) := \frac{H(t)}{F_{max}}$ and $A := \frac{\Delta}{F_{max}}$, a unit-free number. Thus

$$\tilde{F}_{(h,\Delta)}(t) = AF_{(1,1)}\left(\frac{\tilde{H}(t)}{A}\right) \tag{20}$$

shows that the solution of (13) for an arbitrary muscle can be derived from $F_{(1,1)}$, which is the inverse function of the solution $X(t,1)$ of (19) with $\Delta = 1$: $X(t,1) = \frac{t}{1-c} - \frac{\alpha c}{1-c}\int_0^t X(s,1)\frac{t+1}{(s+1)^2}\exp(-\alpha\frac{t-s}{s+1})\,ds$. The values of $F_{(1,1)}$ can therefore be computed very accurately once and for all and then memorized for subsequent computations (Figure 2A).

The relative synaptic weight. By using $In_T(t) = \frac{F_{(h,\Delta)}(t)}{k} + In_0$, (7), and (20), the relative synaptic weight as a function of the tetanic force t is given by

$$\tilde{g}(t) = \frac{V_T}{(E_{EPSP} - V_T)In_0(F_{(1,1)}(\frac{\tilde{H}(t)}{A}) + 1)}. \tag{21}$$

The range of $\tilde{g}$ is

$$[\tilde{g}(t_{max}), \tilde{g}(t_{min})] = \left[\frac{V_T}{(E_{EPSP} - V_T)In_0(F_{(1,1)}(\frac{1}{A}) + 1)}, \frac{V_T}{(E_{EPSP} - V_T)In_0}\right]$$

and is thus independent of the shape of the MU distribution.

The trace of the muscle in $\tilde{g}(t)$ appears in $\tilde{H}(t)$ and A. According to the preceding results, there always exists a unique relative synaptic weight providing a given affine muscle response during recruitment. Since $F_{(1,1)}(\frac{\tilde{H}(t)}{A})$ is strictly increasing in t, $\tilde{g}(t)$ is strictly decreasing. The integrability of ρ and (21) imply that $\rho(t) \overset{a.e.}{=} \frac{\Delta}{t}\frac{d}{dt}F^{-1}_{(1,1)}(\frac{V_T}{(E_{EPSP}-V_T)In_0\tilde{g}(t)} - 1)$. For fixed values of Δ and In_0 (or equivalently for fixed k and In_0) and once two densities equal a.e. have been identified, there is a one-to-one correspondence between the relative synaptic weight and the MU density function.

4. Activation of the muscle and related functionals. The study of the activation of a muscle is simplified by normalization of the input $\tilde{In} = \frac{In}{In_0}$ and the muscle force $\tilde{\mathcal{F}} = \frac{\mathcal{F}}{F_{max}}$. The curve given by $\tilde{In} \mapsto \tilde{\mathcal{F}}(\tilde{In})$ will be called the *relative activation curve*. During recruitment, the latter is given by $\tilde{\mathcal{F}}(\tilde{In}) = \frac{\Delta}{F_{max}}(\tilde{In} - 1)$ (Figure 3A). The unit-free number $A = \frac{\Delta}{F_{max}}$ in (20) is the slope of the preceding straight line and is called the *activation factor* (denoted S in [42]). It is remarkable that A depends on ρ only through its first moment F_{max}.

Recruitment ratio. The *recruitment ratio* is the fraction Q of the force at the end of recruitment and the maximal force of the muscle. As a ratio of two forces, Q is a unit-free number. By using (20), we can write $Q = AF_{(1,1)}(\frac{H(t_{max})}{\Delta})$. Since $H(t_{max}) = F_{max}$ and $A = \frac{\Delta}{F_{max}}$, the recruitment ratio (Figure 3C) depends on A only since

$$Q(A) = A\,F_{(1,1)}\left(\frac{1}{A}\right). \tag{22}$$

Since $F_{(1,\Delta)}(t) = Y(t,\Delta)$, $Q(A) = \frac{Y(\frac{1}{A},1)}{\frac{1}{A}}$, and because of Theorem 12(f), it is a decreasing function of A. Therefore, by using Theorem 15(c) and (d), for all $A > 0$

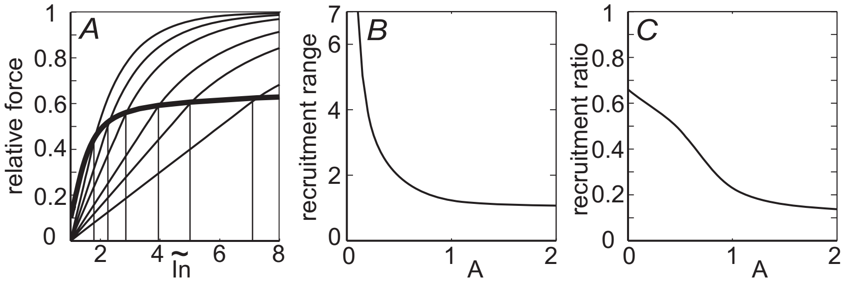


FIG. 3. A: *Relative activation curves and end-recruitment curve as a function of the relative input. They are obtained by projection of the curves in Figure* 4 *in the* $(\tilde{In}, A)$ *plane.* B: *Relative recruitment range* $R(A)$. C: *Recruitment ratio* $Q(A)$.

we have $1 - c = \lim_{u\downarrow 0} \frac{Y(u,1)}{u} \le Q(A) \le \lim_{u\uparrow\infty} \frac{Y(u,1)}{u} = x_\infty^{-1}$. For the estimated values $\alpha = 1.14$ and $c = 0.9$, we get $0.1 \le Q(A) \le 0.66$ for all $A > 0$ and thus for all muscles.

Relative recruitment range. We call *relative recruitment range* R the smallest relative input range within which all MUs are recruited. It is also the factor by which the threshold input of the smallest MU of the pool has to be multiplied in order to recruit all MUs. Since $\tilde{In}_T(t_{min}) = 1$, we have $R = \tilde{In}_T(t_{max}) - 1$. Since A is the slope of the relative activation curve during recruitment (affine range), we have $R = \frac{Q}{A}$. Because of (22), R is a function of A only, given by $R = F_{(1,1)}(\frac{1}{A})$ and thus decreasing (Figure 3B).

If $A \to \infty$, then $R(A) \to 0$ and $Q(A) \to 0.1$, meaning that all MUs tend to have the same threshold and be instantaneously recruited. The muscle force is then close to 10% of the maximal muscle force since, at recruitment, each MU contracts at 10% of its tetanic force ($1 - c = 0.1$). If $A \to 0$, then $R(A) \to +\infty$ and $Q(A) \to 0.66$. The muscle force increases with a slope close to 0. Each recruited MU increases its firing frequency nearly to the tetanic contraction force until the next is recruited. Thus frequency modulation and recruitment develop parallel to each other, and at completion of the MU recruitment, the muscle force approaches 66% of its maximal force.

Relative activation surface. We know that $\mathcal{F}(In) = \int_{t_{min}}^{t_{max}} \rho(s) f(s, In)\, ds$, where $f(s, In) = s\,(1 - c\exp(-\alpha\frac{In - In_T(s)}{In_T(s)}))$ if $In > In_T(s)$ and 0 otherwise. According to (7) and (21), we have

$$\tilde{In}_T(s) = F_{(1,1)}\left(\frac{\tilde{H}(s)}{A}\right) + 1, \quad s \in [t_{min}, t_{max}]. \tag{23}$$

By using the preceding relations and the change of variable $u = \overline{H}(s)$, for $\tilde{In} \ge 1$, we get

$$\tilde{\mathcal{F}}(\tilde{In}) = \int_0^1 I_{\{u \le AF_{(1,1)}^{-1}(\tilde{In}-1)\}} \left(1 - c\,\exp\left(-\alpha \frac{\tilde{In} - (F_{(1,1)}(\frac{u}{A}) + 1)}{F_{(1,1)}(\frac{u}{A}) + 1}\right)\right) du, \tag{24}$$

I_E denoting the indicator function of the set E. For fixed α and c, we see that $\tilde{\mathcal{F}}(\tilde{In})$ is a function of $\tilde{In}$ and A only. Therefore, it can be interpreted as a surface Σ, called the *relative activation surface* (Figure 4) and denoted $\tilde{\mathcal{F}}(A, \tilde{In})$.

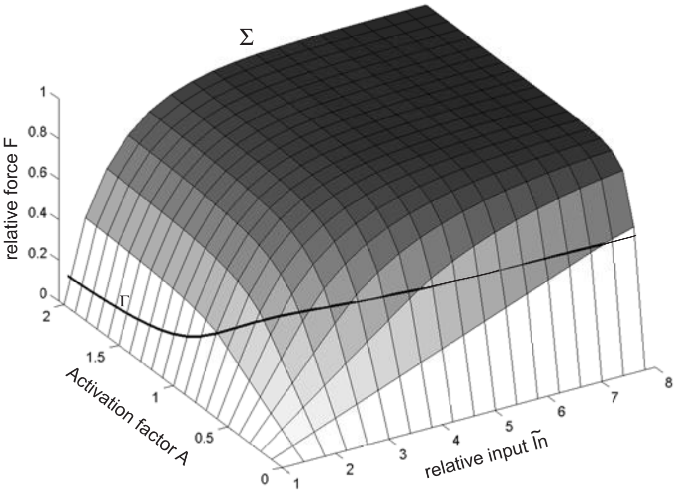


FIG. 4. *The relative activation surface, the relation between the relative input, the activation factor, and the relative force. Vertical sections parallel to the $(\tilde{F}, \tilde{I}n)$ planes are the relative activation curves. The interconnected ends of the relative activation curves give the recruitment boundary curve (bold line).*

Since the function $\hat{F}_{(1,1)}$ is strictly increasing, (24) implies the equivalence of $A_1 < A_2$ and $\tilde{\mathcal{F}}(A_1, \tilde{I}n) < \tilde{\mathcal{F}}(A_2, \tilde{I}n)$ for all $\tilde{I}n > 1$. Consequently, their projections on the (input, force) plane (Figure 3A) never intersect each other for $\tilde{I}n > 1$. The end of the affine part of an activation curve is a point on Σ corresponding to the end of recruitment. The set of these points defines a curve Γ called the *recruitment boundary curve* (bold line in Figure 4). One of the parametric forms of Γ is $0 < A \mapsto (A, 1 + R(A), Q(A))$. We can project Γ on three different planes:

(a) The projection on $(A, \tilde{\mathcal{F}})$ provides the recruitment ratio $Q(A)$ (Figure 3C).

(b) The projection on $(A, \tilde{I}n)$ provides, up to a translation, the relative recruitment range $R(A)$ (Figure 3B).

(c) The projection on $(\tilde{I}n, \tilde{\mathcal{F}})$ provides the end-recruitment curve. Simple computations lead to its representation $\tilde{I}n \mapsto \frac{\tilde{I}n - 1}{F^{-1}_{(1,1)}(\tilde{I}n - 1)}$ (Figure 3A), which, according to Theorem 12(e), is an increasing function.

The vertical sections of Σ parallel to the $(\tilde{I}n, \tilde{\mathcal{F}})$ plane are the relative activation curves, and according to (24), each one of them is determined by A only (Figure 3A).

Relative force contributions due to recruitment and frequency modulation and related functions. The total muscle force is the sum of the contributions due to recruitment and frequency modulation $\mathcal{F}(\tilde{I}n) = \mathcal{F}^{rec}(\tilde{I}n) + \mathcal{F}^{mod}(\tilde{I}n)$. Dividing by F_{max}, we get $\tilde{\mathcal{F}}(\tilde{I}n) = \tilde{\mathcal{F}}^{rec}(\tilde{I}n) + \tilde{\mathcal{F}}^{mod}(\tilde{I}n)$. It is sufficient to compute $\tilde{\mathcal{F}}^{rec}$, which is the fraction $(1 - c)$ of the maximal force produced by all recruited MUs for the input $\tilde{I}n$. Thus $\tilde{\mathcal{F}}^{rec}(\tilde{I}n) = (1 - c)\frac{H(t(\tilde{I}n))}{F_{max}}$, where $t(\tilde{I}n)$ is the tetanic force of the strongest MU recruited by the relative input $\tilde{I}n$. By using (23), we get $\tilde{I}n = F_{(1,1)}(\frac{H(t(\tilde{I}n))}{\Delta}) + 1$, and hence,

$$\tilde{\mathcal{F}}^{rec}(\tilde{I}n) = \begin{cases} (1 - c)\, A\, F^{-1}_{(1,1)}(\tilde{I}n - 1) & \text{if } 1 \le \tilde{I}n \le \tilde{I}n(t_{max}), \\ 1 - c & \text{if } \tilde{I}n > \tilde{I}n(t_{max}). \end{cases} \tag{25}$$

Again we see that the relative muscle forces due to recruitment and to frequency modulation, as functions of $\tilde{In}$, depend only on the activation factor. Straightforward computations provide the unit independent quotients for $1 \leq \tilde{In} \leq In_T(t_{max})$:

$$\frac{\mathcal{F}^{rec}}{\mathcal{F}}(\tilde{In}) = \frac{\tilde{\mathcal{F}}^{rec}}{\tilde{\mathcal{F}}}(\tilde{In}) = (1-c)\frac{F^{-1}_{(1,1)}(\tilde{In}-1)}{\tilde{In}-1},$$

$$\frac{\mathcal{F}^{mod}}{\mathcal{F}^{rec}}(\tilde{In}) = \frac{\tilde{\mathcal{F}}^{mod}}{\tilde{\mathcal{F}}^{rec}}(\tilde{In}) = \frac{1}{1-c}\frac{\tilde{In}-1}{F^{-1}_{(1,1)}(\tilde{In}-1)} - 1.$$

We call the *recruitment rate* the number $100\,\frac{\tilde{\mathcal{F}}^{rec}}{\tilde{\mathcal{F}}}$ (Figure 2C) since it gives, at the end of recruitment, the percentage of the force due to recruitment. The ratio $\frac{\tilde{\mathcal{F}}^{mod}}{\tilde{\mathcal{F}}^{rec}}$ will be called *frequency modulation to recruitment ratio* or simply *FMR* (Figure 2B).

According to Theorem 12(e), $\frac{\tilde{\mathcal{F}}^{rec}}{\tilde{\mathcal{F}}}$ is decreasing and $\frac{\tilde{\mathcal{F}}^{mod}}{\tilde{\mathcal{F}}^{rec}}$ increasing. Both functions are muscle-independent as long as recruitment is not achieved. At the end of recruitment (*e.r.*), we have $\tilde{In} - 1 = R(A)$ and thus

$$\frac{\mathcal{F}^{rec}}{\mathcal{F}}|_{e.r.} = \frac{\tilde{\mathcal{F}}^{rec}}{\tilde{\mathcal{F}}}|_{e.r.} = \frac{(1-c)}{Q(A)}, \qquad \frac{\mathcal{F}^{mod}}{\mathcal{F}^{rec}}|_{e.r.} = \frac{\tilde{\mathcal{F}}^{mod}}{\tilde{\mathcal{F}}^{rec}}|_{e.r.} = \frac{1}{1-c}\,Q(A) - 1.$$

The last two quantities depend only on A, and according to Theorem 12(f), the first one is increasing and the second one is decreasing. For fixed values of Δ, the same properties hold for the variable F_{max} instead of A.

Recruitment gain. The *recruitment gain*, introduced in [28] in the context of H-reflexes, is the "size of threshold differences to recruit additional MUs." The situation is simple in the case of the H-reflex since MUs are activated only once during this reflex and rate modulation is thus nonexistent. The recruitment gain corresponds to the derivative, during recruitment, of the number of active MUs with respect to the relative input. It is given by $Rg(\tilde{In}) = \frac{d}{d\tilde{In}}\int_{t_{min}}^{t(\tilde{In})} \rho(s)\,ds = \rho(t(\tilde{In}))\frac{d}{d\tilde{In}}t(\tilde{In})$. As we have seen before, $t(\tilde{In}) = H^{-1}(\Delta F^{-1}_{(1,1)}(\tilde{In}-1))$, and by differentiating both sides with respect to $\tilde{In}$ and denoting $' = \frac{d}{d\tilde{In}}$, we obtain $\rho(t(\tilde{In}))\,t(\tilde{In})\,t'(\tilde{In}) = \Delta(F^{-1}_{(1,1)})'(\tilde{In}-1)$. Finally, the last three equations lead to

$$Rg(\tilde{In}) = \frac{\Delta(F^{-1}_{(1,1)})'(\tilde{In}-1)}{H^{-1}(\Delta F^{-1}_{(1,1)}(\tilde{In}-1))}.$$

In contrast to the preceding relations derived in this section, $Rg(\tilde{In})$ depends on the particular muscle via H and Δ.

5. Discussion. The relative synaptic weight, which specifies the efficacy of the synaptic input to the individual MNs, cannot be determined with the present experimental techniques. However, every model of the MNPMC requires this quantity, and we present here an approach allowing its computation. A main feature of the model is that it is based on a known behavior of the activation curve during recruitment. Indirect measurements [38] indicate that this function is affine during recruitment, and this turns out to be sufficient to determine the relative synaptic weight. The MNPMC model can now be used to compute various functionals related to the muscle activation. They provide a deeper insight into the processes occurring during muscle activation and also yield values for missing experimental data required for more complex models of the MNPMC. The computed relative synaptic weight has been implemented in a time-dependent model which considers each MU individually [36, 39].

Normalizing the force by the maximal muscle force F_{max} and the input by the threshold input In_0 is the source of several advantages. In this new frame, the slope k of the activation curve during recruitment is transformed into the activation factor $A = \frac{kIn_0}{F_{max}}$. The signature in A of the muscle MU population is F_{max}, the first moment of ρ. One could have expected a more intricate dependence since ρ is an arbitrary nonnegative and integrable function.

The activation factor is the parameter which governs completely the activation process of a muscle in our model. Indeed, the relative activation surface depends on $\tilde{I}n$ and A. It turns out that the relative activation curve depends only on A (even after recruitment is completed), entailing that the recruitment boundary curve Γ, the end-recruitment curve γ, the relative activation curves, the relative recruitment ratio Q, and the relative recruitment range $R(A)$ depend only on A.

It is also quite remarkable that the ratios $\frac{\tilde{\mathcal{F}}^{mod}}{\tilde{\mathcal{F}}^{rec}} = \frac{\mathcal{F}^{mod}}{\mathcal{F}^{rec}}$ and $\frac{\tilde{\mathcal{F}}^{rec}}{\tilde{\mathcal{F}}} = \frac{\mathcal{F}^{rec}}{\mathcal{F}}$, as long as recruitment is not achieved, depend only on the function $F^{-1}_{(1,1)}$ and are thus totally independent of the muscle. Of course, the values of these ratios depend on A at the end of recruitment. However, the relative synaptic weight depends on In_0 and A, and finally, the recruitment gain $Rg(\tilde{I}n)$ depends on Δ and ρ. As in [45], several functionals become muscle-independent or depend only on the activation factor.

The preceding normalization also allows for the comparison of muscles with different strengths, as described in [42] for the first dorsal interosseus, a small hand muscle, and the gastrocnemius, a much stronger leg muscle.

We proved that for fixed values of k and In_0, the MU population density ρ and the relative synaptic weight $\tilde{g}$ are linked by a one-to-one relation. The MU population ρ of a muscle can therefore be recovered from the synaptic weight and conversely is a feature which might be used by the CNS. If the properties of an MU population change by a lesion or a pathological situation such as muscular dystrophy, or simply by disuse or training, an input to the MN pool does not result in the force expected by the CNS. As a consequence, the relative synaptic weight in the MN pool might be readjusted in order to achieve the activation curve required for a properly working motor control. Sensory input from muscle spindles and additional sensors might play a major role in such a feedback system. This hypothesis could be tested in patients with motor diseases, in subjects participating at bed rest and thus concerned with muscle atrophy, or in subjects undergoing a force training.

The activation curve is composed of an affine part, prescribed a priori during recruitment, followed by a nonaffine portion due to frequency modulation only. Our model predicts that the affine part can be maintained up to at most 66% of the maximal muscle force, a situation achievable with a slope approaching 0. For muscle forces above the relative recruitment range, the slope of the activation curve is steadily decreasing. As a consequence, relatively strong inputs are required to adjust high force levels. The activation curve has not yet been investigated in muscles with a large activation factor where the nonaffine range extends over an important domain of the input. However, there is evidence that an affine relationship holds in the working range of the human soleus muscle. Unpublished data (D. G. Ruegg and T. H. Kakebeeke) show that humans are able to voluntarily contract the soleus muscle up to only about 60% of its maximal force, and the activation curve is affine over that range, suggesting that the behavior of the soleus is compatible with our model.

In this way, the CNS might limit the muscle's working field to the affine range, a possibly very useful property. Indeed, motor centers that are hierarchically above the MN pool would be faced with a fixed activation curve. Consequently, a control mechanism at the spinal level would be sufficient to adjust the synaptic weight in order to maintain the activation curve, when changes in the MU population are induced by training, disuse, or disease. Moreover, the whole activation curve is automatically adjusted by the affine part since it depends only on its slope, namely, the activation factor. The verification of this property requires subjects with a modified MU density function.

REFERENCES

[1] J. Aczel, *Vorlesungen über Funktionalgleichungen und ihre Anwendungen*, Birkhäuser Verlag, Basel, 1961.

[2] F. Baldissera and B. Gustafsson, *Firing behavior of a neurone model based on the afterhyperpolarization conductance time course. First interval firing*, Acta Physiol. Scand., 91 (1974), pp. 528–544.

[3] F. D. Bremmer, J. R. Baker, and J. A. Stephens, *Correlation between the discharges of motor units recorded from the same and from different finger muscles in man*, J. Physiol. Lond., 432 (1991), pp. 355–380.

[4] H. Brunner and P. J. Van der Houwen, *The Numerical Solution of Volterra Equations*, North–Holland, Amsterdam, 1986.

[5] W. H. Calvin and P. C. Schwindt, *Steps in production of motoneuron spikes during rhythmic firing*, J. Physiol., 35 (1972), pp. 297–310.

[6] K. L. Chung, *A Course in Probability Theory*, Harcourt, Brace and World, New York, 1968.

[7] A. K. Datta and J. A. Stephens, *Synchronization of motor unit activity during voluntary contraction in man*, J. Physiol. Lond., 422 (1990), pp. 397–419.

[8] J. E. Desmedt and E. Godaux, *Ballistic contractions in man: Characteristic recruitment pattern of single motor units of the tibialis anterior muscle*, J. Physiol. Lond., 264 (1977), pp. 673–693.

[9] J. E. Desmedt and E. Godaux, *Fast motor units are not preferentially activated in rapid voluntary contractions in man*, Nature Lond., 267 (1977), pp. 717–719.

[10] T. M. Flett, *Mathematical Analysis*, McGraw-Hill, New York, 1966.

[11] O. Forster, *Analysis* 2, Friedr. Vieweg & Sohn, Braunschweig, Wiesbaden, Germany, 1996.

[12] A. J. Fuglevand, K. M. Zackowski, K. A. Huey, and R. M. Enoka, *Impairment of neuromuscular propagation during human fatiguing contractions at submaximal forces*, J. Physiol. Lond., 460 (1993), pp. 549–572.

[13] R. A. F. Garnett, M. J. O'Donovan, J. A. Stephens, and A. Taylor, *Motor unit organization of human medial gastrocnemius*, J. Physiol. Lond., 287 (1978), pp. 33–43.

[14] G. Gripenberg, S. O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[15] P. J. Harrison and A. Taylor, *Individual excitatory post-synaptic potentials due to muscle spindle* I*a afferents in cat triceps surae motoneurones*, J. Physiol., 312 (1981), pp. 455–470.

[16] C. J. Heckman and M. D. Binder, *Analysis of effective synaptic currents generated by homonymous* I*a afferent fibers in motoneurons of the cat*, J. Neurophysiol., 60 (1988), pp. 1946–1966.

[17] C. J. Heckman and M. D. Binder, *Computer simulation of the steady-state input-output function of the cat medial gastrocnemius motoneuron pool*, J. Neurophysiol., 65 (1991), pp. 952–967.

[18] C. J. Heckman and M. D. Binder, *Computer simulations of motoneuron firing rate modulation*, J. Neurophysiol., 69 (1993), pp. 1005–1008.

[19] C. J. Heckman and M. D. Binder, *Computer simulations of the effects of different synaptic input systems on motor unit recruitment*, J. Neurophysiol., 70 (1993), pp. 1827–1840.

[20] C. J. Heckman and M. D. Binder, *Computer simulations of the effects of different synaptic input systems on the steady-state input-output structure of the motoneuron pool*, J. Neurophysiol., 71 (1994), pp. 1727–1739.

[21] E. Henneman, *Principles governing distribution of sensory input to motor neurons*, in The Neurosciences: Third Study Program, F. O. Schmitt and F. G. Worden, eds., MIT Press, Cambridge, MA, 1974, pp. 281–291.

[22] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 50–544.

[23] K. Kanosue, M. Yoshida, K. Akazawa, and K. Fujii, *The number of active motor units and their firing rates in voluntary contraction of human brachialis muscle*, Jpn. J. Physiol. Lond., 29 (1979), pp. 427–443.

[24] D. Kernell, *The limits of firing frequency in cat lumbosacral motoneurones possessing different time course of afterhyperpolarization*, Acta Physiol. Scand., 65 (1965), pp. 87–100.

[25] D. Kernell, *High-frequency repetitive firing of cat lumbosacral motoneurones stimulated by long-lasting injected currents*, Acta Physiol. Scand., 65 (1965), pp. 74–86.

[26] D. Kernell, *Functional properties of spinal motoneurons and graduation of muscle force*, in Motor Control Mechanisms in Health and Disease, J. E. Desmedt, ed., Raven Press, New York, 1983, pp. 213–226.

[27] D. Kernell, *Rhythmic properties of motoneurones innervating muscle fibres of different speed in m. gastrocnemius medialis of the cat*, Brain Res., 160 (1979), pp. 159–162.

[28] D. Kernell and H. Hultborn, *Synaptic effects on recruitment gain: A mechanism of importance for the input-output relations of motoneurone pools?*, Brain Res., 507 (1990), pp. 176–179.

[29] M. Kuczma, *An introduction to the theory of functional equations and inequalities*, in Panstwowe Wydawnictwo Naukowe, Uniwersytet Slaski, Warzawa, 1985, pp. 424–426.

[30] C. G. Kukulka and P. Clamann, *Comparison of the recruitment and discharge properties of motor units in human brachial biceps and adductor pollicis during isometric contractions*, Brain Res., 219 (1981), pp. 45–55.

[31] G. W. H. Mantel and R. N. Lemon, *Cross-correlation reveals facilitation of single motor units in thenar muscles by single corticospinal neurons in the conscious monkey*, Neurosci. Lett., 77 (1987), pp. 113–118.

[32] R. K. Miller, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

[33] H. S. Milner-Brown, R. B. Stein, and R. Yemm, *The orderly recruitment of human motor units during voluntary isometric contractions*, J. Physiol. Lond., 230 (1973), pp. 359–370.

[34] H. S. Milner-Brown, R. B. Stein, and R. Yemm, *Changes in firing rate of human motor units during linearly changing voluntary contractions*, J. Physiol. Lond., 230 (1973), pp. 371–390.

[35] A. W. Monster and H. Chan, *Isometric force production by motor units of extensor digitorum communis muscle in man*, J. Neurophysiol., 40 (1977), pp. 1432–1443.

[36] R. M. Nussbaumer, D. G. Rüegg, L. M. Studer, and J.-P. Gabriel, *Computer simulation of the motoneuron pool–muscle complex.* I. *Input system and motoneuron pool*, Biol. Cybern., 86 (2002), pp. 317–333.

[37] S. S. Palmer and E. E. Fetz, *Effects of single intracortical microstimuli in motor cortex on activity of identified forearm motor units in behaving monkeys*, J. Neurophysiol., 54 (1985), pp. 1194–1212.

[38] D. G. Ruegg and F. Bongioanni, *Superposition of ballistic on steady contractions in man*, Exp. Brain Res., 77 (1989), pp. 412–420.

[39] M.-A. Schnetzer, D. G. Ruegg, R. Baltensperger, and J.-P. Gabriel, *Three-dimensional model of the muscle structure and the surface EMG*, in Proceedings of the 5th European Conference of the European Society of Mathematical and Theoretical Biology (ESMTB) on Mathematical Modelling & Computing in Biology and Medicine (Milan, Italy, 2002), Esculapio, Bologna, Italy, 2003.

[40] P. J. Slot and T. Sinkjaer, *Simulations of the alpha motoneuron pool electromyogram reflex at different preactivation levels in man*, Biol. Cybern., 70 (1994), pp. 351–358.

[41] D. R. Smart, *Fixed Point Theorems*, Cambridge University Press, Cambridge, UK, 1974.

[42] L. M. Studer, D. G. Ruegg, and J. P. Gabriel, *A model for steady isometric muscle activation*, Biol. Cybern., 80 (1999), pp. 339–355.

[43] A. A. M. Tax and J. J. Denier van der Gon, *A model for neural control of gradation of muscle force*, Biol. Cybern., 65 (1991), pp. 227–234.

[44] R. J. H. Wilmink and P. J. Slot, *Modeling of the H-reflex facilitation during ramp and hold contractions*, J. Comput. Neurosci., 3 (1996), pp. 337–346.

[45] F. E. Zajac, *Muscle and tendon: Properties, models, scaling, and application to biomechanics and motor control*, Crit. Rev. Biomed. Eng., 17 (1989), pp. 359–411.

SIAM J. Appl. Math.
Vol. 68, No. 3, pp. 890–905

RANDOM COVERING OF MULTIPLE ONE-DIMENSIONAL DOMAINS WITH AN APPLICATION TO DNA SEQUENCING*

MICHAEL C. WENDL†

Abstract. Classical results for randomly covering a one-dimensional domain are generalized to multiple domains. The density function for the number of gaps is derived in the context of Bell's polynomials. Limiting forms are determined as well. The multiple domain configuration is a good model for DNA sequencing scenarios in which the target is fragmented, e.g., filtered DNA libraries and macronuclear genomes. Large-scale sequencing efforts are now starting to focus on such projects. Fragmentation effects are most prominent for small targets but vanish for very large targets. Here, the current model converges with classical theory. Pyrosequencing has been suggested as a viable, much cheaper alternative for large filtered projects. However, our model indicates that a recently demonstrated microscale Sanger reaction will likely be far more effective.

Key words. coverage, probabilistic modeling, sequence redundancy

AMS subject classifications. 05A15, 60D05, 92D20

DOI. 10.1137/06065979X

1. Introduction.

The basic problem of covering a one-dimensional domain with randomly placed segments has been studied extensively for over a century [40]. The probability density function (PDF) for gaps has been described for the unit circle [41], as have moments of coverage [38] and asymptotic characteristics [18, 23, 39]. The real line has been similarly examined [13].

A notable result emerging from this work is Stevens's theorem [40, 41]. Suppose there are i trials, which are not independent of each other. If the probability for any specific σ of these trials to have outcome Γ is $f(\sigma)$, regardless of the outcomes for the remaining $i-\sigma$ trials, then the PDF for γ instances of Γ is

$$P(i,\Gamma=\gamma) \;=\; C_{i,\gamma} \sum_{\sigma=0}^{i-\gamma} C_{i-\gamma,\sigma}\;(-1)^{\sigma}\; f(\sigma+\gamma)\,, \tag{1.1}$$

where $C_{i,\gamma}$ are the binomial coefficients. The relationship between (1.1) and the standard inclusion-exclusion approach is readily demonstrated [16, 28]. In his paper, Stevens [41] went on to show, via straightforward geometric arguments, that the number of gaps in coverage for a circular domain of unit circumference is governed by the kernel probability $f(\sigma) = (1-\sigma\,\varphi)_+^{i-1}$, where i is the number of random covering segments, φ is the fractional length of each segment, and $(t)_+ = \max(0,t)$.

The eminent mathematician-scientist Harold Jeffreys offered one of the more colorful interpretations of the one-dimensional problem [14]. He imagined a bicyclist passing through an intersection strewn with tacks. Upon emerging, the rider wants to know if any tacks have been picked up in her tire but can only glance down at short random intervals as she moves down the road. If each glance covers a fraction φ of the tire circumference, what is the probability that the whole tire has been examined

*Received by the editors May 12, 2006; accepted for publication October 5, 2007; published electronically January 25, 2008. This work was partially supported by a grant from the National Human Genome Research Institute (HG003079).

http://www.siam.org/journals/siap/68-3/65979.html

†Genome Sequencing Center, Washington University, St. Louis, MO 63108 (mwendl@wustl.edu).

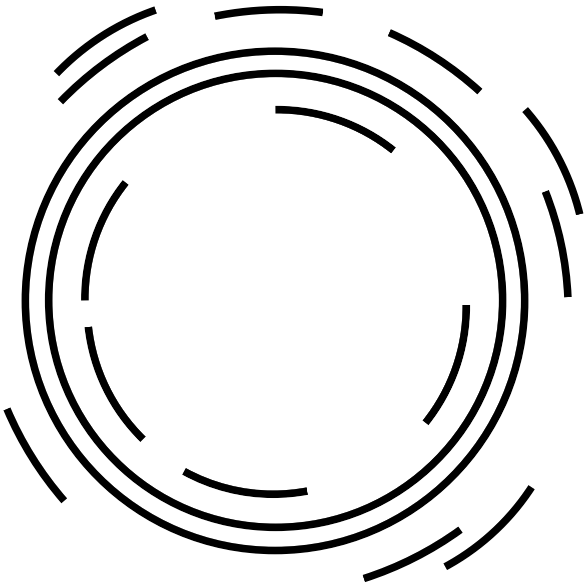


FIG. 1.1. *Schematic of the shotgun covering process for double-stranded DNA. The target is shown as two concentric circles, representing equivalent lengths of complimentary nucleotide sequence. Covering fragments for the outer and inner strands are depicted as arcs lying outside and inside the circles, respectively. The circularized configuration is characteristic of numerous DNA molecules, for example, bacterial chromosomes, genomic plasmids, mitochondrial DNA, and cloned fragments (e.g., bacterial artificial chromosomes (BACs), fosmids, plasmid subclones, etc.).*

after i glances? According to (1.1), the solution is $P\,(i, \Gamma = 0)$, the instance in which there are no observational gaps.

Aside from its mathematical significance [17] and the perhaps surprising relevance to cycling, the one-dimensional configuration is a useful abstraction for a variety of physical problems. Here, we are interested in its singularly important role in modeling DNA processing, particularly the various random "shotgun" strategies for sequencing and mapping [32]. DNA is a nonbranching biopolymer, usually of high molecular weight. Biochemical limitations preclude direct experimental characterization of DNA molecules, for example, at the level of chromosomes or whole genomes. However, smaller fragments can readily be resolved by methods such as Sanger's chain-termination sequencing reaction [37]. On a conceptual level, the shotgun approach consists of nothing more than randomly oversampling a library of suitably small fragments with the objective of covering a larger domain of interest [1, 11]. A number of models have been devised for shotgun processing [29, 35], and these have been shown to be special cases of Stevens's equation (our (1.1)) [46].

One of the factors that has not yet been adequately considered is target multiplicity, of which perhaps the simplest illustration is the double-stranded configuration of DNA itself. For example, if random covering fragments are derived from phage clones, the actual strand from which any particular fragment originates is not known (Figure 1.1). Existing theories tacitly assume that all fragments come from the same strand, meaning that strand information is essentially lost. However, computational algorithms that assemble these fragments find overlaps by checking both the actual sequence and its reverse complement. (The latter is a fragment's representation on the opposite strand deduced from Watson–Crick base-pairing rules.) In effect, strand information is restored ex post facto. Biologists consider closure on both strands, independent of one another, to be important for fully resolving a desired DNA sequence [8, 21]. In particular, this allows detection of strand-specific anomalies [15].

The issue of stranding implies that we are actually dealing with the random, simultaneous covering of two domains, as shown in Figure 1.1. As with the standard independently and identically distributed (IID) assumption for spatial distribution of the fragments, one could presume fragments to be IID over the domains as well. Incidentally, the problem depicted in Figure 1.1 is really the correct analogue to Jeffreys's bicyclist, where the rider glances randomly at both the back and front tires; Jeffreys's original interpretation applies strictly to unicycles!

Quite a few scenarios in DNA sequencing actually demand the generalization of this idea to larger numbers of domains. For example, the most recalcitrant genomes are not directly amenable to the conventional whole genome shotgun method. Biologists have developed various filtering techniques to identify and remove those regions that cannot be resolved using current technology [47]. The filtering process cleaves the original target into thousands of smaller domains (essentially the individual genes). The required shotgun procedure is then performed on the collective set of domains. A similar multiple-target scenario arises with organisms having macronuclear genomes that consist of thousands of small chromosomal structures [12, 33].

Another area of growing interest is the so-called *metagenomic* project, where a community of microbe types is extracted and sequenced directly, i.e., without the need for culturing [10, 20, 30, 36, 43]. Here, there may be as many as 10^5 distinct microbial DNA targets [2]. Strictly speaking, a number of more traditional scenarios are also multiple-target problems—the whole genome shotgun method applied to multichromosome organisms (tens of domains) [42], whole genome reads projected over large-insert libraries (hundreds to tens of thousands of domains, depending upon the project) [34], and clone mapping projects (tens of domains) [24].

Existing theoretical work is limited, either focusing on relevant expected values [45] or assuming distribution of segments according to an expected value argument [31]. Here, we report a more comprehensive characterization of the multiple-domain problem based on the PDF and its limiting forms. These results are applied to a number of unresolved issues in DNA sequencing.

Let us take Γ as the random variable representing the number of gaps in a problem having an arbitrary number of domains N, over which i covering segments have been randomly placed. Here, Γ is the total number of gaps, i.e., the sum of the numbers of gaps over all the individual target domains. This interpretation corresponds to the biological motivations of the problem, especially that of characterizing closure probabilities for the overall DNA target. Each target domain has a length $\Delta > 0$, while each covering segment has a length Λ, where $0 < \Lambda < \Delta$. Consequently, the total target size is $N\Delta$. We do not consider the configuration $\Lambda = \Delta$, as it reduces to the classical occupancy problem [16, 28], which does not exhibit the type of gaps we are interested in here. Segment placements are presumed to be IID within each domain as well as over the N domains. Relevant dimensionless groups are the domain ratio $\varphi = \Lambda/\Delta$, the target ratio φ/N, and the covering redundancy $\rho = i\varphi/N$. The latter can be thought of as the number of segments covering the average target position.

2. The multiple-domain kernel probability.

The kernel probability $f(\sigma)$ in (1.1) is the probability of gaps after σ specific segments, regardless of the gap status associated with all other remaining segments. It can be derived by a straightforward geometric argument for $N = 1$, as shown by Stevens [41]. For $N > 1$, we must also consider the combinatorial aspect of how these σ particular segments can be partitioned over the various N domains. The multiple-domain kernel function is then given by the following constructs.

LEMMA 2.1. *Let x_k be the probability of realizing any given configuration of the i segments over the N domains, where σ specific segments are partitioned in a certain permutation over k of the domains. Then*

$$x_k \; = \; N^{-i} \prod_{\beta=1}^{k} (N - \beta + 1)\,,$$

where $k \in \{1, 2, \ldots, \sigma\}$.

Proof. Probability x_k is a function of two components—one representing the σ segments partitioned in a specified way over the k domains and the other for the remaining $i - \sigma$ nonpartitioned segments, which can be distributed in any manner over all N domains. These components are clearly independent of one another so that x_k is simply the product of the two individual probabilities.

Consider a particular permutation of the σ partitioned segments. Let any k of these segments be placed on any k distinct domains. The remaining $\sigma - k$ segments in the partition must then each fall on a specific domain to satisfy the given permutation. The probability of this event is

$$\frac{N-0}{N} \times \frac{N-1}{N} \times \frac{N-2}{N} \times \cdots \times \frac{N-k+1}{N} \times \left(\frac{1}{N}\right)^{\sigma-k} .$$

There are $N^{i-\sigma}$ ways of placing the remaining $i - \sigma$ nonpartition segments over all N domains. Each of these arrangements is equiprobable, giving a probability of $N^{-(i-\sigma)}$ for any specific one. The product of these two expressions yields

$$\left(\frac{1}{N}\right)^{i-\sigma} \left(\frac{1}{N}\right)^{\sigma-k} \prod_{\beta=1}^{k} \left(\frac{N-\beta+1}{N}\right),$$

from which Lemma 2.1 follows directly. $\square$

LEMMA 2.2. *Let $\Pi(\sigma)$ be the geometric probability for gaps appearing after the σ partitioned segments described in Lemma 2.1 for the union of all configurations of the remaining $i - \sigma$ segments over the N domains. $\Pi(\sigma)$ connotes the partition $g_1^{\nu_1} g_2^{\nu_2} \cdots g_\sigma^{\nu_\sigma}$, where the number of islands hosting the given partitioned segments is $k = \nu_1 + \nu_2 + \cdots + \nu_\sigma$. The probability is*

$$\Pi(\sigma) \; = \; (N - \sigma\varphi)^{i-\sigma} \prod_{\beta=1}^{\sigma} (1 - \beta\varphi)_+^{(\beta-1)\,\nu_\beta} .$$

Proof. This expression can be demonstrated via systematic enumeration, according to how the $i - \sigma$ nonpartition segments can be distributed among all N domains. Each arrangement is mutually exclusive of all others, so the union is evaluated by simple summation. The geometric probability is formulated along the lines of Stevens's argument [41] for each distinct arrangement of segments. That is, a domain of interest hosts $1 \le \pi \le \sigma$ of the partitioned segments, along with $0 \le j \le i - \sigma$ of the nonpartitioned segments. Stevens's term, localized for this domain, can be written in the form $(1 - \pi\,\varphi)_+^{\pi+j-1}$.

A partition can be represented in the general form $g_1^{\nu_1} g_2^{\nu_2} \cdots g_\sigma^{\nu_\sigma}$, where $\sigma = 1\nu_1 + 2\nu_2 + \cdots + \sigma\nu_\sigma$ specific segments are partitioned over $k = \nu_1 + \nu_2 + \cdots + \nu_\sigma$ domains of interest. That is, each g_π represents one of these domains, which hosts π of the σ segments. This partitional notation implies a left-to-right arrangement of the k domains, though the ordering is simply a convenience. Now, consider all configurations where j of the $i - \sigma$ nonpartitioned segments are "mixed in" with the partitioned segments; i.e., they reside in some manner on the same k domains where the partitioned segments lie. Then, the remaining $i - \sigma - j$ segments must lie on the other $N - k$ domains.

We must first quantify the joint geometric probabilities for all the different ways that the j nonpartitioned segments can be distributed over the k domains. For example, if all j reside on the leftmost g_1 domain, we have a geometric probability of the form

$$\underbrace{(1 - 1\,\varphi)_+^{1+j-1}\,(1 - 1\,\varphi)_+^{1+0-1} \cdots (1 - 1\,\varphi)_+^{1+0-1}}_{\nu_1 \text{ such terms representing all ``} g_1 \text{'' domains}}\,(1 - 2\,\varphi)_+^{2+0-1} \cdots (1 - \sigma\,\varphi)_+^{\sigma+0-1} \; .$$

Note that all j nonpartition segments are assigned to the leftmost term representing the leftmost g_1 domain. This expression would also have a coefficient $C_{i-\sigma,j}$ to account for the number of ways to pick j segments from the collection of $i - \sigma$ nonpartition segments. Also, there is an implied multiplier of unity, indicating that there is only one way to assign the j segments to this single domain.

In a similar fashion, we then consider all the remaining ways of distributing the j nonpartitioned segments over the k domains, formulating the appropriate joint geometric probability in each case. In addition to the $C_{i-\sigma,j}$ coefficient for the number of ways of picking the j segments, we have the appropriate k-nomial coefficient $j!/(j_1! j_2! \cdots j_k!)$, where $j_1 + j_2 + \cdots + j_k = j$, to account for the number of ways of assigning the j segments to the k domains. All of these joint geometric probabilities are summed, and the result is multiplied by $(N-k)^{i-\sigma-j}$, which represents the number of ways the $i - \sigma - j$ remaining nonpartitioned segments could be assigned to the other $N - k$ domains. The resulting summation S_j is such that one can factor out the product

$$(1 - 1\,\varphi)_+^{1-1} \cdots (1 - \sigma\,\varphi)_+^{\sigma-1} \;=\; \prod_{\beta=1}^{\sigma} (1 - \beta\varphi)_+^{(\beta-1)\,\nu_\beta} \;=\; \mathfrak{P} \, .$$

Note that the $(\,)_+$ limiters continue to govern the positivity of $\mathfrak{P}$, and because $\mathfrak{P}$ is a multiplicative factor, these limiters are superfluous for the remaining Stevens-type terms in the summation. Consequently, the overall expression can be written

$$S_j \;=\; (N-k)^{i-\sigma-j} \cdot C_{i-\sigma,j} \cdot \mathfrak{P} \cdot \Big[(1 - 1\,\varphi)^j \;+\; \cdots \;+\; (1 - \sigma\,\varphi)^j \Big] \; ,$$

where the $(\,)_+$ notation has been dropped from the summed terms. These terms collapse via the multinomial theorem, and, after some algebraic manipulation, this expression can be written

$$S_j \;=\; (N-k)^{i-\sigma-j} \cdot C_{i-\sigma,j} \cdot \mathfrak{P} \cdot \Big[\nu_1\,(1 - 1\,\varphi) \;+\; \cdots \;+\; \nu_\sigma\,(1 - \sigma\,\varphi) \Big]^j \; .$$

By factoring further, S_j reduces to

$$S_j \;=\; (N-k)^{i-\sigma-j} \cdot C_{i-\sigma,j} \cdot \mathfrak{P} \cdot (k - \sigma\varphi)^j \; .$$

Considering the union of all cases for $j \in \{0, 1, \ldots, i - \sigma\}$, we find

$$\Pi(\sigma) \;=\; \mathfrak{P} \cdot \sum_{j=0}^{i-\sigma} C_{i-\sigma,j} \cdot (N-k)^{(i-\sigma)-j} \cdot (k - \sigma\varphi)^j \;,$$

which itself collapses via the binomial theorem into Lemma 2.2. □

Theorem 2.3. *The kernel probability for the multiple-domain covering problem can be cast in terms of Bell's exponential polynomials* [3, 9]

$$f(\sigma) \;=\; \sum_{\Pi(\sigma)} \frac{\sigma!\, x_k}{\nu_1!\, \nu_2! \,\cdots\, \nu_\sigma!} \left(\frac{g_1}{1!}\right)^{\nu_1} \left(\frac{g_2}{2!}\right)^{\nu_2} \!\cdots \left(\frac{g_\sigma}{\sigma!}\right)^{\nu_\sigma} \;,$$

where the x_k are probabilities given by Lemma 2.1 *and the set partitions $g_1^{\nu_1} g_2^{\nu_2} \cdots g_\sigma^{\nu_\sigma}$ represent the probabilities given by Lemma* 2.2.

Proof. The kernel probability for σ specific segments can be found by considering the union of all the ways to both partition these σ specific segments and assign the remaining $i - \sigma$ nonspecific segments over the N domains. That is, $f(\sigma)$ is the summation over all the segment configurations of $P_c P_g$, which are probability of the configuration itself and the geometric gap probability associated with the configuration, respectively.

The probabilities for all configurations having a common permutation of the σ partitioned segments are identical, as discussed in Lemma 2.1. Thus, x_k is the factored configurational probability that multiplies the corresponding sum of geometric probabilities given by Lemma 2.2. Finally, the coefficients in Bell's polynomials account for how many ways a given partition can be permuted. □

As an example, consider the partitioning of exactly four segments. Theorem 2.3 takes the form

$$f(4) \;=\; 1\, g_4\, x_1 \;+\; \left(4\, g_3\, g_1 \;+\; 3\, g_2^2\right) x_2 \;+\; 6\, g_2\, g_1^2\, x_3 \;+\; 1\, g_1^4\, x_4 \,. \tag{2.1}$$

The g_4 partition indicates that all four segments lie on one domain, $g_3\, g_1$ specifies three segments on one domain with the fourth on another, etc. Polynomial coefficients indicate the number of ways each partition can be permuted. Stevens's problem for a single domain [41] can be thought of as a special case of Theorem 2.3 in which only the first term of each polynomial σ is retained. That is, $f(\sigma) = 1 g_\sigma x_1$, where $x_1 = 1$ and g_σ is Stevens's geometric probability.

3. Asymptotic limiting cases. Although the results in the previous section are exact, there are well-known limitations for evaluating Bell's polynomials as σ becomes large. Complications arise primarily as a result of growing N because terms are needed up to x_N in each polynomial. Here, we report two asymptotic limiting cases that are of biological relevance.

First asymptotic limit. Consider the case of a finite, fixed target size, but where the target itself becomes progressively more fragmented. In the limit, the size of the individual target domains approaches the covering segment length. This scenario implies that φ/N can be held at certain constant values such that $\varphi \to 1$. Here, Stevens's solution shows that for any given domain

$$P(i = 1, \Gamma = \gamma) = \begin{cases} 0 & : \quad \gamma = 0, \\ 1 & : \quad \gamma = 1, \end{cases} \qquad \text{and} \qquad P(i \geq 2, \Gamma = \gamma) \to \begin{cases} 1 & : \quad \gamma = 0, \\ 0 & : \quad \gamma = 1 \end{cases}$$

as well as $P\,(i, \Gamma \geq 2) = 0$. Consequently, finding the distribution of gaps among the target domains is asymptotically equivalent to the problem of determining the distribution of domains having exactly one segment each. The latter is a variation of the classical occupancy problem [16, 28].

Within the combinatorial context of our analysis thus far, the most straightforward solution for this case is given by the following construct.

THEOREM 3.1. *The probability distribution P_1 for the first asymptotic case $\varphi \to 1$ is*

$$P_1\,(i, \Gamma = \gamma) \;=\; N^{-i} \sum_{\Pi(i)} \frac{i!}{\nu_1!\,\nu_2!\,\cdots\,\nu_i!\,(1!)^{\nu_1}\,(2!)^{\nu_2}\cdots(i!)^{\nu_i}}\;\delta_{\gamma,\nu_1}(N)_k\,,$$

where δ is the Kronecker delta, $(N)_k = N\,(N{-}1)\cdots(N{-}k{+}1)$, and $\Gamma \in \{0, 1, 2, \ldots, N\}$.

Proof. Bell's coefficient (the quotient) gives the number of ways to realize a given partition of i segments over $k = \nu_1 + \nu_2 + \cdots + \nu_\sigma$ specific domains, where ν_1 of these domains have exactly one segment each. There are $(N)_k$ permutations of a k-bin arrangement so that the product of the two values is the total number of ways to find the ν_1 domains of interest among the N total domains. The enumeration is summed over all partitions, tallying only those where $\nu_1 = \gamma$ via the Kronecker delta. There are N^i possible configurations of the i segments over the N domains. Assuming each is equally likely, the probability is the quotient of the above enumeration and N^i. □

One can actually find the distribution of domains having any desired number of segments by simply adjusting the second subscript on δ appropriately. Other combinatorial results for this problem are available as well [28]. Unfortunately, such solutions are not terribly practical from a computational standpoint for reasons already discussed above. A more amenable form comes directly from probability theory [16]:

$$P_1\,(i, \Gamma = \gamma) \;=\; \frac{(-1)^\gamma\,N!\,i!}{\gamma!\,N^i} \sum_{j=\gamma}^{\min(N,i)} \frac{(-1)^j\,(N-j)^{i-j}}{(j-\gamma)!\;(N-j)!\;(i-j)!}\,. \tag{3.1}$$

Computing each PDF using (3.1) requires evaluation of $(N+1)(N+2)/2$ terms for the usual case of interest, i.e., $i \geq N$.

Second asymptotic limit. We can extend the previous concept to the scenario where the original target is infinitely large. Specifically, the target is fragmented into $N \to \infty$ domains such that $\varphi \to 1$ (thus $\varphi/N \to 0$). This variation once again has a direct analogue in the corresponding occupancy problem, which converges to a Poisson process if the rate remains bounded. That is, the distribution for the second asymptotic case approaches

$$P_2\,(i, \Gamma = \gamma) \;=\; \frac{\exp\,(-\lambda)\,\lambda^\gamma}{\gamma!}\,, \tag{3.2}$$

where the rate is $\lambda = i \exp\,(-i/N)$ [16].

4. Discussion.

Shotgun DNA sequencing is a comparatively new research tool to the biomedical sciences [1, 11], yet it is already responsible for numerous discoveries of fundamental importance. Sequencing projects continue to be guided almost exclusively by the collective empirical experience of the sequencing community, with little to no input from theoretical foundations. This is largely attributable to the fact that the latter have not evolved sufficiently to treat many of the biologically relevant phenomena of sequencing.

Lander and Waterman developed what most biologists consider to be the standard model almost two decades ago [29]. This theory provides the expected value of gaps based on the assumption of a single, monolithic DNA target. In fact, subsequent work has focused almost entirely on the monolithic target [35, 46]. Yet, many sequencing scenarios now routinely involve fragmented targets, as outlined above. Wendl and Barbazuk [45] have extended the standard theory to multiple domains of linear molecules, but their model still considers only expected values. Here, boundary conditions (i.e., "edge-effects") appreciably influence sequencing progress. However, expected value analysis is of little use for circular configurations. Indeed, it is fairly straightforward to show that the expected number of gaps for a fragmented target is identical to the Lander–Waterman expression if the target is large enough. This suggests that the influences of fragmentation may be more subtle.

Here, we illustrate a few applications of the theory for which fragmentation is known, or at least presumed, to play a role. Unless otherwise stated, a covering segment length of 1,000 is assumed. This value represents the current limit of what can be resolved by the chain-termination sequencing reaction [37] on a sustained basis. Each calculation uses 700 digits behind the decimal point. We will also abbreviate DNA nucleotide bases and base-pairs as "nuc" and "bp," respectively.

Trends under increasing fragmentation. How does the gap census evolve as targets become progressively more fragmented? This question has some practical bearing on DNA sequencing projects, especially for sample-related calculations that depend upon the variance [22].

First, let us examine fragmentation in the mild to moderate range. In general, the PDF becomes increasingly diffuse, but the degree to which this occurs is highly dependent upon the parameters. Consider, for example, the case $\varphi/N = 1/80$, which corresponds to the shotgun sequencing of a standard double-stranded 40,000 bp fosmid clone [26]. The total target size is 80,000 nuc. Figure 4.1 shows gap PDFs for three configurations—a single 80,000 nuc target, four 20,000 nuc target domains, and eighty 1,000 nuc domains. The last configuration corresponds to the first asymptotic limit in (3.1). Functions are shown for light redundancy $\rho = 1$ and moderate redundancy $\rho = 4$. Note that we calculate ρ based on the total number of nucleotides. Consequently, our definition differs by a factor of two from the casual, more common usage of ρ, which does not consider stranding.

Graphical resolution shows no difference between the single monolithic target $N = 1$ and a mildly fragmented target consisting of $N = 4$ domains. Differences for these two configurations would be noticeable only for smaller target sizes, e.g., those characteristic of single gene islands (see below). Consequently, it appears that the issue of stranding we described above can be neglected in most cases of biological interest. It also appears that mild fragmentation will not have a significant effect for most projects, especially since read lengths will often be somewhat less than 1000 bp. Conversely, the first asymptotic case (3.1) does differ to some degree. For example, the PDF is clearly more diffuse at $\rho = 1$, although this configuration shows a significant trend toward convergence at $\rho = 4$.

At the opposite extreme is the scenario where a very large target is fragmented into numerous very small domains. Such is the case, for example, when filtering schemes are used to extract and sequence the gene complement of large genomes [47] or when sequencing certain genomes that are inherently highly fragmented [12, 33]. In particular, filtering something like the maize genome gives a target size on the order of 5×10^8 bp with about 2×10^5 islands [47]. In these scenarios, one finds complete

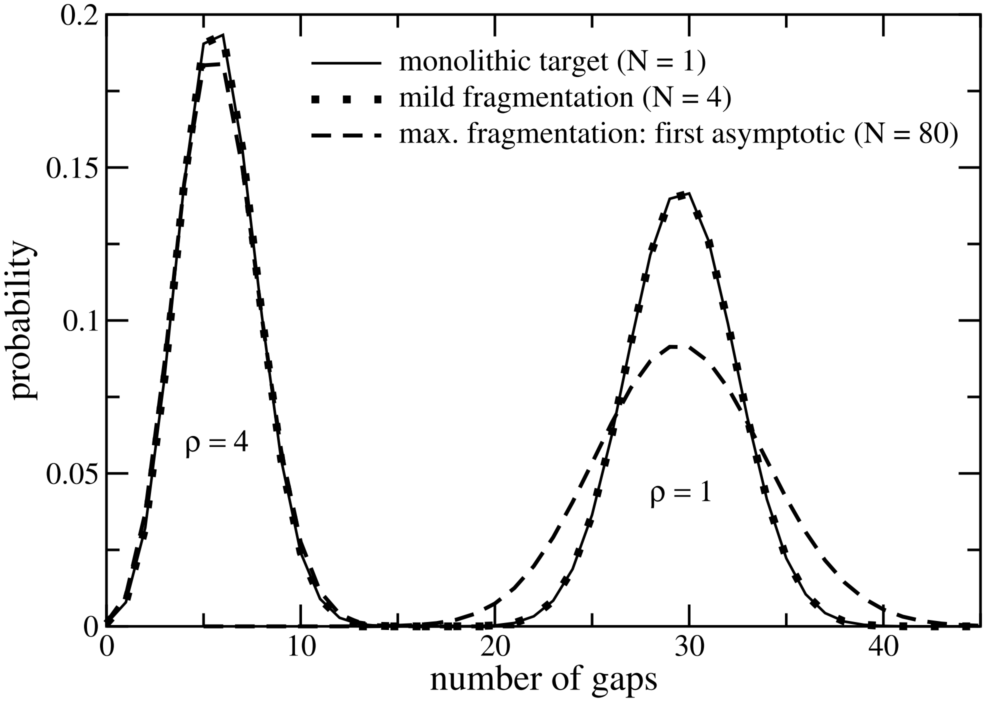


Fig. 4.1. *Gap density functions at* 1*-fold and* 4*-fold redundancies for a fosmid-class target at various levels of fragmentation.*

convergence with the classical theory for monolithic targets; i.e., the influence of fragmentation vanishes. The Poisson rate in the second asymptotic limit (3.2) is equivalent to $\lambda = i \exp(-\rho)$ since $\varphi \to 1$. The binomial PDF for monolithic targets, e.g., equation (9) in [46], associates a Bernoulli probability of $\exp(-\rho)$ with i trials. Consequently, it converges to this same expression for large targets.

In summary, these results indicate that φ is a strong variable when φ/N is relatively large. Conversely, the process is essentially independent of φ as φ/N becomes vanishingly small.

The stopping problem. One of the more practical questions that arises in all sequencing projects is when to halt the random phase of processing. This is the so-called *stopping problem*. Appreciable economic consequences revolve around this decision because subsequent directed procedures are at least an order of magnitude more expensive per unit sequence recovered. Incidentally, the nonrandom continuation is required because non-IID characteristics become apparent at higher redundancies, making it difficult to achieve a $\gamma = 0$ "base-perfect" sequence using the random method alone.

Although early work established $\rho = 10$ as the nominal "full shotgun" stopping point [48], the matter is still debated. For example, Bouck et al. [6] advise $\rho \leq 6$, while Blakesley et al. [4] recommend $\rho \geq 8$. Draft sequences have been reported for as low as $\rho \approx 3.5$ [49] and as high as $\rho \approx 17.5$ [19]. Although one finds sequence redundancies distributed somewhat uniformly between 5-fold and 15-fold, they do not correlate with any gross genome features, e.g., genome size [44]. In short, available data indicate that there is no commonly accepted system by which stopping points should be chosen.

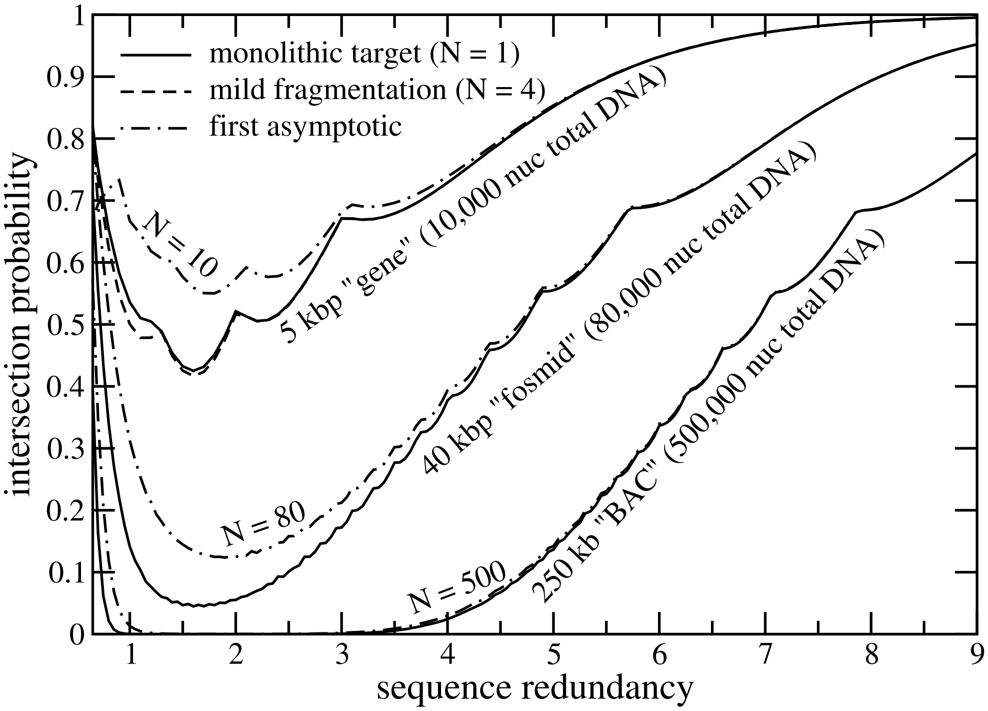


Fig. 4.2. *Intersection probabilities for three classes of DNA target.*

Wendl [44] proposed a quantity called the intersection probability $P_\cap$ to quantify stopping points probabilistically. The idea is based on the simple observation that pairs of density functions separated by intervals of $\rho = 1$ share increasing fractions of their event spaces as ρ grows. $P_\cap$ is the tail probability calculated from the overlap. This implies that any differences in two identical but independent projects become more attributable to stochastic variation rather than the disparity in their redundancies.

Figure 4.2 shows the intersection probability plotted as a function of redundancy for three classes of genomic targets—a 5 kbp gene-sized island, a 40 kbp fosmid clone, and a 250 kbp BAC clone. The trends confirm a number of earlier observations. First, fragmentation tends to be important only in scenarios where the total target size is small. For example, the effect of mild fragmentation ($N = 4$) is apparent only for the 5 kbp gene island at low redundancy. Likewise, the state of maximal fragmentation (the first asymptotic) rapidly converges to the classical result. In particular, it is almost indistinguishable from the $N = 1$ curve for the BAC. It appears that the classical model could be used to calculate $P_\cap$ in many cases. For smaller targets, the effect of fragmentation, though mild, suggests halting the random phase earlier than for the equivalent nonfragmented project.

The second trend echoes what previous work has found with respect to $P_\cap$ for genome coverage: larger projects should be sequenced to higher redundancies [44]. In other words, successive PDFs spaced 1-fold units of redundancy apart are more isolated from one another in larger projects. This clearly applies whether a genomic target is fragmented or not. Results shown in Figure 4.2 advise substantially lower redundancies for large-insert clone targets than what is often specified at the present time. For example, the $P_\cap = 0.3$ threshold implies limiting redundancies to something less than 4 and 6 for fosmids and BACs, respectively. This is most in line with the empirical findings of Bouck et al. [6].

Cost optimization of filtered projects. With the completion of the human genome project [25], many large-scale efforts are now shifting to even more recalcitrant, repeat-laden genomes, e.g., maize. Current thinking holds that standard whole genome methods are unlikely to succeed here, which has prompted proposals for various forms of genomic filtering [47]. Filtration schemes produce numerous low-copy, sequenceable islands, thereby circumventing many of the well-known problems of genomic assembly. Maximizing read length is not nearly as critical in these scenarios, and this has led to growing debate surrounding proper choice of read length [7].

One of the more relevant aspects on which to judge this question would certainly be cost, specifically overall project cost. Traditional Sanger sequencing instruments, e.g., the ABI 3730xl, provide "full-length" reads but at relatively high per-read expenditures [25]. Conversely, newer platforms such as the 454 GS-20 pyrosequencer yield short, much lower-cost reads [27]. Let us assume that high-throughput labs can generate Sanger 3730xl reads and GS-20 pyrosequencing reads for about \$0.44 and \$0.013 each, respectively. Instruments based on a recent demonstration of dramatically scaled-down Sanger reactions [5] will be future contenders as well. Such devices should easily be able to sequence for about one-tenth the cost of the current generation of Sanger-based machines.

Figure 4.3 shows overall project cost for sequencing a filtered maize library as a function of the amount of random sequence. The latter is quantified by the intersection probability rather than standard redundancy because it provides for scale-free comparisons [44]. Results are all generated using the second asymptotic limit in (3.2). We specify read lengths of 719, 556, and 110, respectively, for traditional Sanger reads [47], next-generation Sanger reads [5], and pyrosequencing reads [27]. We also examine three thresholds for detecting overlaps [29]: 30, 40, and 50 bp. These correspond, respectively, to minimum, typical, and highly conservative values [47].

The plot shows a number of notable trends. First, the current landscape clearly favors pyrosequencing over traditional Sanger sequencing, although not by as much as we might intuitively expect based on the > 30-fold difference in unit costs. The pyrosequencing technique yields a larger project in the sense of φ/N, meaning that comparatively higher redundancies are needed to obtain given milestones of $P_\cap$. This effect was shown in Figure 4.2 and is discussed extensively in [44].

The economics of pyrosequencing reads are quite sensitive to how much overlap is required for detection, whereas this sensitivity is not much of a factor for Sanger data. This is somewhat intuitive, since a given overlap consumes a larger fraction of read length for the former. Conversely, traditional Sanger reads are tied more strongly to $P_\cap$ than pyrosequencing reads. $P_\cap$ is a free parameter, and its choice will have a larger bearing on total project cost for Sanger data.

The next-generation Sanger data show the more surprising results. We have assumed a plausible unit cost of just over 4 cents per read—roughly 3-fold higher than that for pyrosequencing. Yet, the overall project cost predictions are significantly lower. Moreover, these data show little sensitivity to $P_\cap$ and are almost constant with respect to the overlap detection threshold. Various commentators have advocated decreasing unit costs or increasing read length, but it is clearly their combination that is most important.

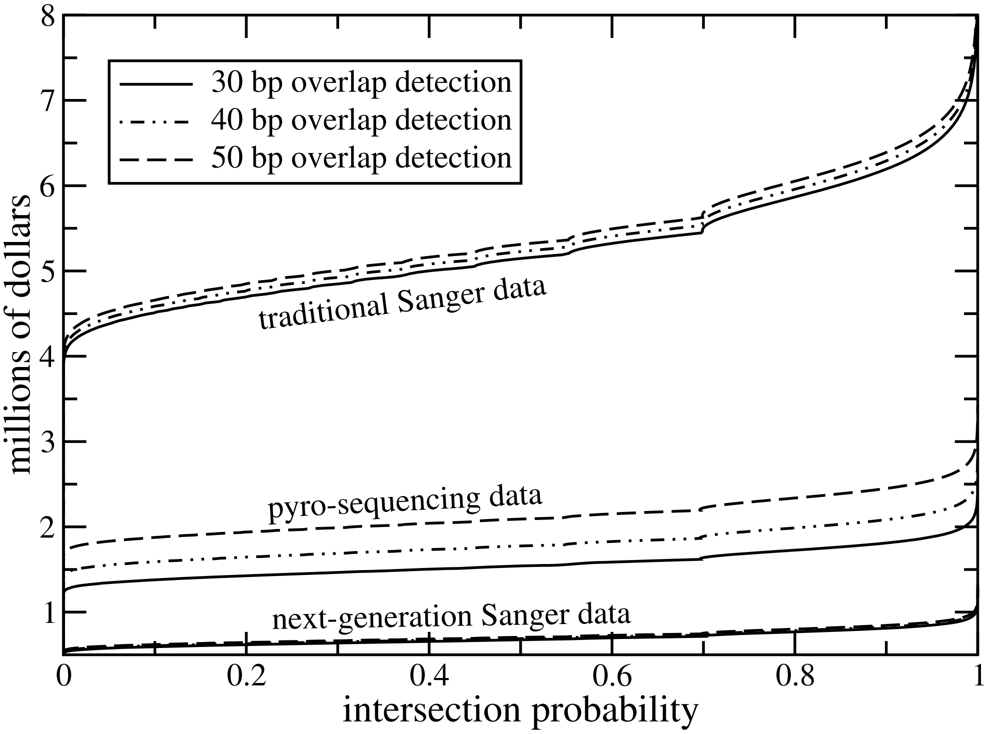


FIG. 4.3. *Project cost estimates for sequencing a filtered* 540 *Mb maize library using three different methods of data generation.*

5. Closing remarks. The problem of randomly covering multiple domains is sufficiently interesting from a mathematical perspective but also provides a good model for many scenarios of DNA sequencing. Investigators have yet to devise a suitable a priori model of cloning bias, so results for Sanger-generated data should be considered in this context. Pyrosequencing does not utilize bacterial cloning, so its predictions are probably somewhat more realistic.

Commercial instruments that eventually implement the miniaturized Sanger paradigm [5] will almost surely realize unit costs even lower than what we have presumed here. It is also likely that read lengths could be extended further. With respect to overall cost of de novo sequencing projects, our results suggest that this approach is much more promising than other techniques that emphasize read cost at the expense of read length.

Acknowledgments. The author gratefully acknowledges discussions with B. Barbazuk (Donald Danforth Plant Science Center) and J. Wallis (Washington University) regarding various aspects of this work and suggestions from one of the anonymous referees that improved the presentation.

REFERENCES

[1] S. ANDERSON, *Shotgun DNA sequencing using cloned DNase* I*–generated fragments*, Nucleic Acids Research, 9 (1981), pp. 3015–3027.

[2] F. E. Angly, B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer, *The marine viromes of four oceanic regions*, PLoS Biology, 4 (2006), article e368.

[3] E. T. Bell, *Exponential polynomials*, Ann. of Math., 35 (1934), pp. 258–277.

[4] R. W. Blakesley, N. F. Hansen, J. C. Mullikin, P. J. Thomas, J. C. McDowell, B. Maskeri, A. C. Young, B. Benjamin, S. Y. Brooks, B. I. Coleman, J. Gupta, S. L. Ho, E. M. Karlins, Q. L. Maduro, S. Stantripop, C. Tsurgeon, J. L. Vogt, M. A. Walker, C. A. Masiello, X. B. Guan, G. G. Bouffard, and E. D. Green, *An intermediate grade of finished genomic sequence suitable for comparative analyses*, Genome Research, 14 (2004), pp. 2235–2244.

[5] R. G. Blazej, P. Kumaresan, and R. A. Mathies, *Microfabricated bioprocessor for integrated nanoliter–scale Sanger DNA sequencing*, Proc. Nat. Acad. Sci. USA, 103 (2006), pp. 7240–7245.

[6] J. Bouck, W. Miller, J. H. Gorrell, D. Muzny, and R. A. Gibbs, *Analysis of the quality and utility of random shotgun sequencing at low redundancies*, Genome Research, 8 (1998), pp. 1074–1084.

[7] M. Chaisson, P. Pevzner, and H. Tang, *Fragment assembly with short reads*, Bioinformatics, 20 (2004), pp. 2067–2074.

[8] S. L. Chissoe, M. A. Marra, L. Hillier, R. Brinkman, R. K. Wilson, and R. H. Waterston, *Representation of cloned genomic sequences in two sequencing vectors: Correlation of DNA sequence and subclone distribution*, Nucleic Acids Research, 25 (1997), pp. 2960–2966.

[9] L. Comtet, *Advanced Combinatorics*, Reidel Publishing, Dordrecht, The Netherlands, 1974.

[10] A. I. Culley, A. S. Lang, and C. A. Suttle, *Metagenomic analysis of coastal RNA virus communities*, Science, 312 (2006), pp. 1795–1798.

[11] P. L. Deininger, *Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis*, Analytical Biochemistry, 129 (1983), pp. 216–223.

[12] T. G. Doak, A. R. O. Cavalcanti, N. A. Stover, D. M. Dunn, R. Weiss, G. Herrick, and L. F. Landweber, *Sequencing the Oxytricha trifallax macronuclear genome: A pilot project*, Trends in Genetics, 19 (2003), pp. 603–607.

[13] C. Domb, *The problem of random intervals on a line*, Proceedings of the Cambridge Philosophical Society, 43 (1947), pp. 329–341.

[14] C. Domb, *On Hammersley's method for one–dimensional covering problems*, in Disorder in Physical Systems, G. R. Grimmett and D. J. A. Welsh, eds., Oxford University Press, Oxford, UK, 1990, pp. 33–53.

[15] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, *Base–calling of automated sequencer traces using* Phred. *I. Accuracy assessment*, Genome Research, 8 (1998), pp. 175–185.

[16] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., John Wiley & Sons, New York, 1968.

[17] R. A. Fisher, *On the similarity of the distributions found for the test of significance in harmonic analysis and in Stevens' problem in geometrical probability*, Annals of Eugenics, 10 (1940), pp. 14–17.

[18] L. Flatto and A. G. Konheim, *The random division of an interval and the random covering of a circle*, SIAM Rev., 4 (1962), pp. 211–222.

[19] J. E. Galagan, S. E. Calvo, K. A. Borkovich, E. U. Selker, N. D. Read, D. Jaffe, W. Fitzhugh, L. J. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S. Wang, C. B. Nielsen, J. Butler, M. Endrizzi, D. Qui, P. Ianakiev, D. Bell-Pedersen, M. A. Nelson, M. Werner-Washburne, C. P. Selitrennikoff, J. A. Kinsey, E. L. Braun, A. Zelter, U. Schulte, G. O. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg, A. Roy, K. Foley, J. Naylor, N. Stange-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvysselis, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R. L. Metzenberg, D. D. Perkins, S. Kroken, C. Cogoni, G. Macino, D. Catcheside, W. Li, R. J. Pratt, S. A. Osmani, C. P. C. Desouza, L. Glass, M. J. Orbach, J. A. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seiler, J. Dunlap, A. Radford, R. Aramayo, D. O. Natvig, L. A. Alex, G. Mannhaupt, D. J. Ebbole, M. Freitag, I. Paulsen, M. S. Sachs, E. S. Lander, C. Nusbaum, and B. Birren, *The genome sequence of the filamentous fungus Neurospora crassa*, Nature, 422 (2003), pp. 859–868.

[20] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson, *Metagenomic*

analysis of the human distal gut microbiome, Science, 312 (2006), pp. 1355–1359.

[21] P. Green, *Against a whole–genome shotgun*, Genome Research, 7 (1997), pp. 410–417.

[22] P. G. Hoel, *Introduction to Mathematical Statistics*, John Wiley & Sons, New York, 1947.

[23] J. Hüsler, *Random coverage of the circle and asymptotic distributions*, J. Appl. Probab., 19 (1982), pp. 578–587.

[24] International Human Genome Mapping Consortium, *A physical map of the human genome*, Nature, 409 (2001), pp. 934–941.

[25] International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome*, Nature, 431 (2004), pp. 931–945.

[26] U.-J. Kim, H. Shizuya, P. J. de Jong, B. Birren, and M. I. Simon, *Stable propagation of cosmid sized human DNA inserts in an F–factor based vector*, Nucleic Acids Research, 20 (1992), pp. 1083–1085.

[27] J. Kling, *The search for a sequencing thoroughbred*, Nature Biotechnology, 23 (2005), pp. 1333–1335.

[28] V. F. Kolchin, B. A. Sevastyanov, and V. P. Christyakov, *Random Allocations*, John Wiley & Sons, New York, 1978.

[29] E. S. Lander and M. S. Waterman, *Genomic mapping by fingerprinting random clones: A mathematical analysis*, Genomics, 2 (1988), pp. 231–239.

[30] H. G. Martín, N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz, *Metagenomic analysis of two enhanced biological phosphorus removal EBPR sludge communities*, Nature Biotechnology, 24 (2006), pp. 1263–1269.

[31] J. Moriarty, J. R. Marchesi, and A. Metcalfe, *Bounds on the distribution of the number of gaps when circles and lines are covered by fragments: Theory and practical application to genomic and metagenomic projects*, BMC Bioinformatics, 8 (2007), article 70.

[32] G. Myers, *Whole–genome DNA sequencing*, Comput. Sci. Engrg., 1 (1999), pp. 33–43.

[33] D. M. Prescott, J. D. Prescott, and R. M. Prescott, *Coding properties of macronuclear DNA molecules in Sterkiella nova (Oxytricha nova)*, Protist, 153 (2002), pp. 71–77.

[34] Rat Genome Sequencing Project Consortium, *Genome sequence of the Brown Norway rat yields insights into mammalian evolution*, Nature, 428 (2004), pp. 493–521.

[35] J. C. Roach, *Random subcloning*, Genome Research, 5 (1995), pp. 464–473.

[36] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter, *The* Sorcerer II *global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific*, PLoS Biology, 5 (2007), article e77.

[37] F. Sanger, S. Nicklen, and A. R. Coulson, *DNA sequencing with chain–terminating inhibitors*, Proc. Natl. Acad. Sci. USA, 74 (1977), pp. 5463–5467.

[38] A. F. Siegel, *Random arcs on the circle*, J. Appl. Probab., 15 (1978), pp. 774–789.

[39] A. F. Siegel, *Asymptotic coverage distributions on the circle*, Ann. Probab., 7 (1979), pp. 651–661.

[40] H. Solomon, *Geometric Probability*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 28, SIAM, Philadelphia, 1978.

[41] W. L. Stevens, *Solution to a geometrical problem in probability*, Ann. Eugenics, 9 (1939), pp. 315–320.

[42] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. Q. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. H. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, C. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. M. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. M. Ge, F. C. Gong, Z. P. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. X. Ke, K. A. Ketchum, Z. W. Lai, Y. D.

Lei, Z. Y. Li, J. Y. Li, Y. Liang, X. Y. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. X. Shue, J. T. Sun, Z. Y. Wang, A. H. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. L. Xiao, C. H. Yan, A. Yao, J. Ye, M. Zhan, W. Q. Zhang, H. Y. Zhang, Q. Zhao, L. S. Zheng, F. Zhong, W. Y. Zhong, S. P. C. Zhu, S. Y. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. J. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Y. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. J. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Y. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. H. Zhu, *The sequence of the human genome*, Science, 291 (2001), pp. 1304–1351.

[43] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith, *Environmental genome shotgun sequencing of the Sargasso Sea*, Science, 304 (2004), pp. 66–74.

[44] M. C. Wendl, *Occupancy modeling of coverage distribution for whole genome shotgun DNA sequencing*, Bull. Math. Biol., 68 (2006), pp. 179–196.

[45] M. C. Wendl and W. B. Barbazuk, *Extension of Lander–Waterman theory for sequencing filtered DNA libraries*, BMC Bioinformatics, 6 (2005), article 245.

[46] M. C. Wendl and R. H. Waterston, *Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing*, Genome Research, 12 (2002), pp. 1943–1949.

[47] C. A. Whitelaw, W. B. Barbazuk, G. Pertea, A. P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J. L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M. A. Budiman, A. Resnick, S. van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C. M. Fraser, and J. Quackenbush, *Enrichment of gene–coding sequences in maize by genome filtration*, Science, 302 (2003), pp. 2118–2120.

[48] R. Wilson, R. Ainscough, K. Anderson, C. Baynes, M. Berks, J. Burton, M. Connell, J. Bonfield, T. Copsey, J. Cooper, A. Coulson, M. Craxton, S. Dear, Z. Du, R. Durbin, A. Favello, A. Fraser, L. Fulton, A. Gardner, P. Green, T. Hawkins, L. Hillier, M. Jier, L. Johnston, M. Jones, J. Kershaw, J. Kirsten, N. Laisster, P. Latreille, C. Lloyd, B. Mortimore, M. Ocallaghan, J. Parsons, C. Percy, L. Rifken, A. Roopra, D. Saunders, R. Shownkeen, M. Sims, N. Smaldon, A. Smith, M. Smith, E. Sonnhammer, R. Staden, J. Sulston, J. Thierry-Mieg, K. Thomas, M. Vaudin, K. Vaughan, R. Waterston, A. Watson, L. Weinstock, J. Wilkinson-Sproat, and P. Wohldman, 2.2 *Mb of contiguous nucleotide sequence from chromosome* III *of C. elegans*, Nature, 368 (1994), pp. 32–38.

[49] J. Yu, S. Hu, J. Wang, G. K. S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W.

Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang, *A draft sequence of the rice genome (Oryza sativa L. ssp. indica)*, Science, 296 (2002), pp. 79–92.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 907–924

THE FACTORIZATION METHOD FOR ELECTRICAL IMPEDANCE TOMOGRAPHY IN THE HALF-SPACE*

MARTIN HANKE† AND BIRGIT SCHAPPEL†

Abstract. We consider the inverse problem of electrical impedance tomography in a conducting half-space, given electrostatic measurements on its boundary, i.e., a hyperplane. We first provide a rigorous weak analysis of the corresponding forward problem and then develop a numerical algorithm to solve an associated inverse problem. This inverse problem consists of the reconstruction of certain inclusions within the half-space which have a different conductivity than the background. To solve the inverse problem we employ the so-called factorization method of Kirsch, which so far has only been considered for the impedance tomography problem in bounded domains. Our analysis of the forward problem makes use of a Liouville-type argument which says that a harmonic function in the entire two-dimensional plane must be a constant if some weighted L^2-norm of this function is bounded.

Key words. electrical impedance tomography, half-space problem, factorization method

AMS subject classifications. 65N21, 35R30, 35J20, 35D05, 31A05

DOI. 10.1137/06067064X

1. Introduction.

Electrical impedance tomography (EIT) is a technique to recover information of the interior of a conducting object from electrostatic measurements taken on its boundary. In mathematical terms, this amounts to recovering information about the spatially varying (nonnegative) conductivity σ in the elliptic partial differential equation

$$\nabla \cdot \sigma \nabla u = 0 \qquad \text{in } B \tag{1.1}$$

from Neumann and Dirichlet boundary values of all stationary electric potentials u. This inverse boundary value problem goes back to Calderón [8] who considered (1.1) in a *bounded* domain B, provoking substantial interest in the medical imaging community.

In geoelectric applications, on the other hand, the domain B and its boundary are typically very large compared to the small fraction of its boundary where data can be measured. Therefore it makes sense to reconsider the inverse boundary value problem for (1.1) in *unbounded* domains B with *unbounded* boundary ∂B, with the half-space being the most obvious and prominent example. Another application for this model problem concerns the automatic recognition of gesture input for interactive displays, called touchless interaction, which has recently been considered by van Berkel and Lionheart [26]. Finally, in its original medical context, the half-space problem may serve as an appropriate model for certain mammography systems (cf., e.g., [2, 16, 22]) where measurements are taken on only a small portion of the patient's skin.

For the half-space B, Druskin [11] has shown that the conductivity can be reconstructed from the knowledge of the boundary data on a subdomain $\Gamma \subset \partial B$, provided that B can be subdivided into a finite set of domains with piecewise smooth boundaries and constant conductivities, respectively. In this paper we are concerned with

*Received by the editors September 25, 2006; accepted for publication (in revised form) June 19, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/siap/68-4/67064.html

†Institut für Mathematik, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany (hanke@math.uni-mainz.de, schappel@math.uni-mainz.de).

numerical algorithms to reconstruct the conductivity σ or partial information about it. In general, reconstruction methods can be divided into iterative and direct methods (we refer to Borcea [4, 5] for a relatively recent survey with the focus on bounded domains), but concerning unbounded domains B we are aware only of previous algorithms based on linearization with or without an outer iteration; cf. Lukaschewitsch, Maass, and Pidcock [20, 21], Mueller, Isaacson, and Newell [22], and the references therein. Iterative methods require the repeated solution of forward problems in each iteration, i.e., differential equations, which tends to be extremely time-consuming. We therefore present a noniterative reconstruction algorithm which can be used to detect abrupt local deviations of the conductivity from a homogenous background conductivity.

Our method is a variation of the so-called *factorization method* which goes back to an idea of Kirsch [18] in the context of inverse scattering and has been applied successfully to the impedance tomography problem by Brühl [6, 7]. In these and subsequent papers all authors have formulated the problem in either bounded domains or all of $\mathbb{R}^n$, thus avoiding the difficulties that arise with domains with unbounded boundaries. Here we employ a general framework developed by Gebauer [12] to adapt this method to the case of the half-space

$$B = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \cdot e_n > 0\},$$

with $e_n \in \mathbb{R}^n$ a given unit vector, the inner normal vector on ∂B. Most results will be presented for the case $n = 3$, but at the end of this paper we will give a short summary of the two-dimensional case.

For our approach we assume a constant background conductivity $\sigma_{\mathbf{1}} = \mathbf{1}$, where $\mathbf{1}$ is the function identically 1, and consider conductivities of the form

$$\sigma(x) = \begin{cases} \kappa(x), & x \in \overline{\Omega}, \\ 1, & x \in \mathbb{R}^3_+ \setminus \overline{\Omega}, \end{cases} \tag{1.2}$$

where $\Omega \subset B$ is a finite collection of separated and bounded domains with sufficiently smooth boundary $\Sigma = \partial\Omega$, and for which $\mathbb{R}^n_+ \setminus \overline{\Omega}$ is connected. Below we will denote by ν the normal of Σ pointing into Ω.

The positive conductivity $\kappa \in L^\infty(\Omega)$ is assumed to be significantly higher or lower than the background conductivity; i.e., there exists $\varepsilon > 0$ such that

$$\kappa(x) \geq 1 + \varepsilon \quad \text{or} \quad \varepsilon \leq \kappa(x) \leq 1 - \varepsilon \qquad \text{for } x \in \overline{\Omega}. \tag{1.3}$$

By means of the factorization method we provide an explicit characterization of the inclusions Ω in terms of the (local) Neumann–Dirichlet operator Λ_σ which maps Neumann boundary values of a potential u in (1.1) to its Dirichlet boundary values.

We should mention that in principle it should be possible to relax the assumption that the background conductivity is constant. However, the numerical implementation of our method will then become much more difficult, as the algorithm requires the availability of the associated Neumann function.

The paper is organized as follows: We first introduce appropriate function spaces to deal with the forward problem (1.1) in the half-space $B = \mathbb{R}^3_+$, and then clarify our notion of weak solutions of (1.1) and their existence. The inverse problem and some preliminary statements will be specified in section 3. Then, in sections 4 and 5 we prove the characterization of inclusions from the knowledge of Λ_σ. In section 6 we comment on our numerical algorithm and present some reconstructions based on simulated data. To conclude, we briefly comment in section 7 on the necessary modifications of our theory in two space dimensions.

In an appendix we establish a Liouville-type result for harmonic functions in the plane which we use to show that certain apparently different function spaces over $\mathbb{R}^3_+$ which have been introduced in the literature, and which are relevant for our problem, are essentially the same.

2. Function spaces and weak solutions of the forward problem. The forward problem associated with impedance tomography in the half-space is the Neumann problem

$$\nabla \cdot \sigma \nabla u = 0 \quad \text{in } \mathbb{R}^3_+, \qquad -\sigma \frac{\partial u}{\partial e_3} = f \quad \text{on } \mathbb{R}^2, \tag{2.1}$$

together with an appropriate growth condition near infinity. Problem (2.1) represents the physical process of injecting a current f into the upper half-space $B = \mathbb{R}^3_+$ from its boundary. In this section the conductivity σ is assumed to be bounded and strictly positive in $\mathbb{R}^3_+$. In (2.1) and in the remainder of this paper, we always identify the boundary of $\mathbb{R}^3_+$ with $\mathbb{R}^2$, with straightforward abuse of notation.

Care has to be taken concerning the correct behavior of $u(x)$ for $|x| \to \infty$ which is reflected by the choice of appropriate function spaces for the solution u. For example, physically relevant solutions of problem (2.1)—like the fundamental solution of the Laplace equation—need not belong to the Sobolev space $H^1(\mathbb{R}^3_+)$.

Example 2.1 (see [21]). For $\sigma = \mathbf{1}$ and $f(y) = (1+|y|^2)^{-3/2}$, a solution of (2.1) is given by $u(x) = |x + e_3|^{-1}$. It is easy to see that u does not belong to $L^2(\mathbb{R}^3_+)$; however, the gradient of u is square integrable on $\mathbb{R}^3_+$.

To construct a suitable function space we recall the following familiar definitions and notation. For a (possibly unbounded) domain $G \subset \mathbb{R}^3$ we take $C_0^\infty(G)$ to be the set of all functions $u \in C^\infty(G)$ with compact support $\operatorname{supp} u$, and let

$$C_0^\infty(\overline{G}) = \{u|_G : u \in C_0^\infty(\mathbb{R}^3)\}.$$

Furthermore, $\mathcal{D}'(G)$ is the set of distributions, i.e., the continuous linear functionals on $C_0^\infty(G)$.

In view of the physical setting (and in accordance with Example 2.1) it appears appropriate to restrict our attention to solutions of (2.1) with finite energy, which means that the H^1-seminorm of u is finite. Note that this seminorm is actually a norm on $C_0^\infty(\overline{\mathbb{R}^3_+})$ because constant functions do not belong to this set. We write $H(\mathbb{R}^3_+)$ for the closure of $C_0^\infty(\overline{\mathbb{R}^3_+})$ with respect to this norm, denoted subsequently by $\|\cdot\|_{H(\mathbb{R}^3_+)}$. According to Boulmezaoud [3], this space coincides with the weighted Sobolev space

$$\{u \in \mathcal{D}'(\mathbb{R}^3_+) : (1+|x|^2)^{-1/2} u \in L^2(\mathbb{R}^3_+),\ \nabla u \in L^2(\mathbb{R}^3_+)^3\}. \tag{2.2}$$

Obviously, we have $H(\mathbb{R}^3_+) \subset H^1_{\text{loc}}(\mathbb{R}^3_+)$, and for every bounded domain $G \subset \mathbb{R}^3_+$ the restriction operator $u \mapsto u|_G$ is continuous as a mapping from $H(\mathbb{R}^3_+) \to H^1(G)$. We point out here that for the two-dimensional case the analogous completion of $C_0^\infty(\overline{\mathbb{R}^2_+})$ with respect to the H^1-seminorm does not yield a space of distributions (cf. Deny and Lions [10]), and we refer to section 7 for the modifications which are necessary in two space dimensions.

It has been shown by Janßen [17] that every function $u \in H(\mathbb{R}^3_+)$ has a trace in

$$L^{2,1}(\mathbb{R}^2) = \{g : (1+|y|^2)^{-1/2} g \in L^2(\mathbb{R}^2)\},$$

and that the trace operator is continuous with respect to the norm

$$\|g\|^2_{L^{2,1}(\mathbb{R}^2)} = \int_{\mathbb{R}^2} (1+|y|^2)^{-1} g^2(y)\,\mathrm{d}y$$

of $L^{2,1}(\mathbb{R}^2)$. Note that the dual space of $L^{2,1}(\mathbb{R}^2)$ can be identified with

$$L^{2,-1}(\mathbb{R}^2) = \{f : (1+|y|^2)^{1/2} f \in L^2(\mathbb{R}^2)\}$$

using $L^2(\mathbb{R}^2)$ as a pivot space in the Gelfand triple. The associated norm of $L^{2,-1}(\mathbb{R}^2)$ is denoted by $\|\cdot\|_{L^{2,-1}(\mathbb{R}^2)}$, the dual pairing between $L^{2,-1}(\mathbb{R}^2)$ and $L^{2,1}(\mathbb{R}^2)$ by

$$\langle f, g\rangle_{\mathbb{R}^2} = \int_{\mathbb{R}^2} f(y)g(y)\,\mathrm{d}y.$$

Now we return to the Neumann problem (2.1) for $u \in H(\mathbb{R}^3_+)$. The corresponding weak formulation follows in the usual way by making use of Green's formula for $u, v \in H(\mathbb{R}^3_+)$ established in [3]: Find $u \in H(\mathbb{R}^3_+)$ such that

$$\int_{\mathbb{R}^3_+} \sigma\nabla u\cdot\nabla v\,\mathrm{d}x = \int_{\mathbb{R}^2} fv\,\mathrm{d}y \qquad \text{for all } v \in H(\mathbb{R}^3_+). \tag{2.3}$$

Problem (2.3) is well defined for every $f \in L^{2,-1}(\mathbb{R}^2)$, and a standard application of the Lax–Milgram lemma establishes existence of a unique solution $u \in H(\mathbb{R}^3_+)$ of (2.3) with

$$\|u\|_{H}(\mathbb{R}^3_+) \le c\|f\|_{L^{2,-1}(\mathbb{R}^2)} \tag{2.4}$$

for some constant $c > 0$ depending only on the conductivity σ. We call u the weak solution of problem (2.1).

Example 2.2. If $\sigma = \mathbf{1}$, i.e., if we consider the Laplace equation, then

$$u(x) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{f(y)}{|x-y|}\,\mathrm{d}y, \qquad x \in \mathbb{R}^3_+, \tag{2.5}$$

is the physically relevant classical solution of problem (2.1) provided that f is continuous and that there exists a positive and monotonic function $\varepsilon \in L^1(\mathbb{R}^+)$ such that $|f(y)| \le \varepsilon(|y|)$; see, e.g., Dautray and Lions [9, Chapter II]. In particular, for $f(y) = (1+|y|^2)^{-3/2}$ this yields the function u of Example 2.1. For arbitrary $f \in L^{2,-1}(\mathbb{R}^2)$ the integral representation (2.5) defines the weak solution $u \in H(\mathbb{R}^3_+)$, as is most easily seen by using the Kelvin transformation; see [25] for further details.

Remark 2.3. In principle one can alternatively start with $C^\infty(\mathbb{R}^3_+)$ instead of $C_0^\infty(\overline{\mathbb{R}^3_+})$ and consider for some $\alpha \ge 0$ the completion $W^1_\alpha(\mathbb{R}^3_+)$ of

$$\{u \in C^\infty(\mathbb{R}^3_+) : (1+|x|^2)^{-1/2-\alpha/2}u \in L^2(\mathbb{R}^3_+),\ \nabla u \in L^2(\mathbb{R}^3_+)^3\}$$

with respect to the norm

$$\|u\|^2_{W^1_\alpha(\mathbb{R}^3_+)} = \int_{\mathbb{R}^3_+} (1+|x|^2)^{-1-\alpha}u^2(x)\,\mathrm{d}x + \int_{\mathbb{R}^3_+} |\nabla u(x)|^2\,\mathrm{d}x; \tag{2.6}$$

cf. [17] and [21]. With an argument due to Hanouzet [15, Théorème I.1] it can be shown that for $\alpha = 0$ this space coincides with $H(\mathbb{R}^3_+)$.

Because of a Poincaré-type inequality established in [21, Theorem 2.10], the spaces $W^1_\alpha(\mathbb{R}^3_+)$ with $\alpha > 1/2$ are all the same and contain $H(\mathbb{R}^3_+)$ as a closed subspace; we denote this space by $H_{1/2+}(\mathbb{R}^3_+)$ and remark that $H(\mathbb{R}^3_+)$ is a proper subspace of $H_{1/2+}(\mathbb{R}^3_+)$ because the latter includes the constants. As we will prove in the appendix, we have, in fact,

$$H_{1/2+}(\mathbb{R}^3_+) = H(\mathbb{R}^3_+) \oplus \operatorname{span}\{\mathbf{1}\}; \tag{2.7}$$

i.e., $H_{1/2+}(\mathbb{R}^3_+)$ is made up only from $H(\mathbb{R}^3_+)$ and the constants.

Yet another variant, considered in [21], is to start with C^∞-functions in $\mathbb{R}^3_+$ which are vanishing for $|x| \to \infty$. The completion of this space with respect to the norm (2.6) always yields the space $H(\mathbb{R}^3_+)$ no matter what value of $\alpha \geq 0$ is used [25, Appendix A].

Thus, both of the aforementioned variants lead essentially to the same notion of a weak solution of problem (2.1), for the constants always belong to the null space of the differential operator under consideration.

3. Basic properties of the inverse problem. Now we are going to specify somewhat further the impedance tomography problem we consider in this paper. We shall assume throughout that the conductivity σ has the form given in (1.2), (1.3) and recall that, by virtue of (2.4), we have a well-defined bounded linear operator from $L^{2,-1}(\mathbb{R}^2)$ into $H(\mathbb{R}^3_+)$ which maps a given boundary current $f \in L^{2,-1}(\mathbb{R}^2)$ onto the induced potential $u_\sigma \in H(\mathbb{R}^3_+)$. By passing on to the trace of u_σ on $\mathbb{R}^2$ we obtain the *Neumann–Dirichlet operator*

$$\Lambda^g_\sigma : L^{2,-1}(\mathbb{R}^2) \to L^{2,1}(\mathbb{R}^2), \qquad f \mapsto u_\sigma|_{\mathbb{R}^2}.$$

Here, the superscript g stands for *global*, because for practical purposes it is often sufficient to restrict the attention to currents f supported on some bounded subset $\Gamma \subset \mathbb{R}^2$, and also to confine oneself to taking measurements of u_σ only on Γ. This gives rise to the so-called *local Neumann–Dirichlet operator*

$$\Lambda^\ell_\sigma : L^2(\Gamma) \to L^2(\Gamma), \qquad f|_\Gamma \mapsto u_\sigma|_\Gamma.$$

It is easy to check that there holds

$$\Lambda^\ell_\sigma = P\Lambda^g_\sigma P', \tag{3.1}$$

where P is the projection

$$P : L^{2,1}(\mathbb{R}^2) \to L^2(\Gamma), \qquad g \mapsto g|_\Gamma,$$

and

$$P' : L^2(\Gamma) \to L^{2,-1}(\mathbb{R}^2), \qquad f \mapsto \begin{cases} f(y), & y \in \Gamma, \\ 0, & y \in \mathbb{R}^2 \setminus \Gamma, \end{cases}$$

is the dual operator of P.

Our inverse problem is now the following:

> *Let the conductivity σ be of the form* (1.2) *with κ as in* (1.3)*, and let Λ^g_σ—or Λ^ℓ_σ for some bounded and relatively open subset $\Gamma \subset \mathbb{R}^2$, respectively—be given. How can we reconstruct the support of κ, i.e., the discontinuities of σ?*

Before we proceed to derive a constructive answer to this question, we list some elementary properties of the operators Λ_σ^g and Λ_σ^ℓ.

LEMMA 3.1.

(i) $\Lambda_\sigma^\ell : L^2(\Gamma) \to L^2(\Gamma)$ *is compact, self-adjoint, and positive.*

(ii) *Let* $f \in L^{2,-1}(\mathbb{R}^2)$*, and let* $\sigma_1, \sigma_2 \in L^\infty(\mathbb{R}^3_+)$ *with* $\sigma_1, \sigma_2 \geq \varepsilon$ *almost everywhere in* $\mathbb{R}^3_+$*. Then, if* $\sigma_1 \leq \sigma_2$*, we have*

$$\langle f, \Lambda_{\sigma_1}^g f\rangle_{\mathbb{R}^2} \geq \langle f, \Lambda_{\sigma_2}^g f\rangle_{\mathbb{R}^2}.$$

Proof. Consider some bounded domain $G \subset \mathbb{R}^3_+$ with $\Gamma \subset \partial G$. Then, as mentioned before, the operator which restricts $u \in H(\mathbb{R}^3_+)$ to $u|_G \in H^1(G)$ is bounded, and the trace operator from $H^1(G)$ to $L^2(\Gamma)$ is compact. Hence, Λ_σ^ℓ is compact.

Let $0 \neq f, \tilde{f} \in L^2(\Gamma)$, and $u_\sigma, \tilde{u}_\sigma \in H(\mathbb{R}^3_+)$ be the solutions of (2.3) with $f' = P'f$ and $\tilde{f}' = P'\tilde{f}$, respectively. Then by virtue of (3.1) we have

$$\langle f, \Lambda_\sigma^\ell \tilde{f}\rangle_{L^2(\Gamma)} = \langle f', \Lambda_\sigma^g \tilde{f}'\rangle_{\mathbb{R}^2} = \int_{\mathbb{R}^3_+} \sigma \nabla u_\sigma \cdot \nabla \tilde{u}_\sigma \,\mathrm{d}x = \langle \tilde{f}, \Lambda_\sigma^\ell f\rangle_{L^2(\Gamma)}.$$

Thus Λ_σ^ℓ is self-adjoint. With $f = \tilde{f}$ we obtain, using (1.3), that

$$\langle f, \Lambda_\sigma^\ell f\rangle_{L^2(\Gamma)} = \int_{\mathbb{R}^3_+} \sigma |\nabla u_\sigma|^2 \,\mathrm{d}x \geq \varepsilon \int_{\mathbb{R}^3_+} |\nabla u_\sigma|^2 \,\mathrm{d}x = \varepsilon \|u_\sigma\|^2_{H(\mathbb{R}^3_+)},$$

and hence, Λ_σ^ℓ is positive.

Now, let $f \in L^{2,-1}(\mathbb{R}^2)$ be given, and let $u_{\sigma_1}, u_{\sigma_2} \in H(\mathbb{R}^3_+)$ be the weak solutions of (2.1) for the two conductivities σ_1 and σ_2, respectively. From (2.3) it follows that u_{σ_1} is the unique minimizer in $H(\mathbb{R}^3_+)$ of the quadratic energy functional

$$\frac{1}{2}\int_{\mathbb{R}^3_+} \sigma_1 \,|\nabla u|^2 \,dx - \langle f, u\rangle_{\mathbb{R}^2}$$

with minimum value $-\frac{1}{2}\langle f, u_{\sigma_1}\rangle_{\mathbb{R}^2}$. Therefore

$$\begin{aligned} -\frac{1}{2}\langle f, \Lambda_{\sigma_1}^g f\rangle_{\mathbb{R}^2} &= \frac{1}{2}\int_{\mathbb{R}^3_+} \sigma_1 \,|\nabla u_{\sigma_1}|^2 \,\mathrm{d}x - \langle f, u_{\sigma_1}\rangle_{\mathbb{R}^2} \\ &\leq \frac{1}{2}\int_{\mathbb{R}^3_+} \sigma_1 \,|\nabla u_{\sigma_2}|^2 \,\mathrm{d}x - \langle f, u_{\sigma_2}\rangle_{\mathbb{R}^2} \\ &\leq \frac{1}{2}\int_{\mathbb{R}^3_+} \sigma_2 \,|\nabla u_{\sigma_2}|^2 \,\mathrm{d}x - \langle f, u_{\sigma_2}\rangle_{\mathbb{R}^2} = -\frac{1}{2}\langle f, \Lambda_{\sigma_2}^g f\rangle_{\mathbb{R}^2}, \end{aligned} \tag{3.2}$$

which was to be shown. □

Our approach to the solution of the inverse problem is based on a comparison of the measured Neumann–Dirichlet operator Λ_σ^g or Λ_σ^ℓ with the reference operator $\Lambda_{\mathbf{1}}^g$ or $\Lambda_{\mathbf{1}}^\ell$, respectively, corresponding to the homogeneous background with conductivity $\mathbf{1}$. From Lemma 3.1 we immediately conclude the following.

COROLLARY 3.2. *Under the assumptions* (1.2), (1.3), $\Lambda_\sigma^g - \Lambda_{\mathbf{1}}^g$ *as well as* $\Lambda_\sigma^\ell - \Lambda_{\mathbf{1}}^\ell$ *are self-adjoint and positive (resp., negative) if* $\kappa \leq 1 - \varepsilon$ *(resp.,* $\kappa \geq 1 + \varepsilon$*).*

Proof. An adaptation of the proof of Lemma 3.1(i) establishes that Λ_σ^g and $\Lambda_{\mathbf{1}}^g$ are self-adjoint. For the remainder of the proof we consider only the case where $\kappa \leq 1 - \varepsilon$

for some $\varepsilon > 0$; the other case is treated similarly. For this situation we obtain from Lemma 3.1 that

$$\langle f, (\Lambda_\sigma^g - \Lambda_{\mathbf{1}}^g) f\rangle_{\mathbb{R}^2} = \langle f, \Lambda_\sigma^g f\rangle_{\mathbb{R}^2} - \langle f, \Lambda_{\mathbf{1}}^g f\rangle_{\mathbb{R}^2} \geq 0 \tag{3.3}$$

for every $f \in L^{2,-1}(\mathbb{R}^2)$; moreover, strict inequality holds in (3.2), and thus in (3.3), if the two potentials u_σ and $u_{\mathbf{1}}$ occurring in the proof of Lemma 3.1 are different.

Thus, assuming equality in (3.3) for some $f \in L^{2,-1}(\mathbb{R}^2)$, we can conclude that

$$0 = \langle f, \Lambda_\sigma^g f\rangle_{\mathbb{R}^2} - \langle f, \Lambda_{\mathbf{1}}^g f\rangle_{\mathbb{R}^2} = \int_\Omega (\kappa - 1)|\nabla u_{\mathbf{1}}|^2 \,\mathrm{d}x,$$

and hence, $u_{\mathbf{1}}$ is constant in Ω. Since $u_{\mathbf{1}}$ is harmonic in $\mathbb{R}^3_+$, this implies that it is constant in the entire half-space. It follows that $f = 0$, which proves that $\Lambda_\sigma^g - \Lambda_{\mathbf{1}}^g$ is positive.

For the local Neumann–Dirichlet operators we consider $f \in L^2(\Gamma)$, and set $f' = P'f$. By virtue of (3.1) we obtain

$$\langle f, (\Lambda_\sigma^\ell - \Lambda_{\mathbf{1}}^\ell) f\rangle_{L^2(\Gamma)} = \langle f', (\Lambda_\sigma^g - \Lambda_{\mathbf{1}}^g) f'\rangle_{\mathbb{R}^2},$$

where the latter is positive according to the first part of this proof, unless $f = 0$. Therefore $\Lambda_\sigma^\ell - \Lambda_{\mathbf{1}}^\ell$ is also a positive operator. □

4. The framework for the factorization method. In what follows our notation will no longer make explicit whether we are talking about local or global measurements; i.e., we write Λ_σ for either Λ_σ^g or Λ_σ^ℓ. Furthermore, we denote by $T = \mathbb{R}^2$ or $T = \Gamma$ the domain, on which measurements shall be taken. In accordance with this notation, we let $H(T)$ be either $L^{2,1}(\mathbb{R}^2)$ or $L^2(\Gamma)$, respectively.

To simplify the presentation we will assume throughout that Ω consists of only one connected component. Our theory extends to the general case, and whenever necessary we will point out the appropriate modifications for this more general situation (see also [24]).

We want to apply the general framework of Gebauer and therefore adopt his notation from [12] in what follows. We first introduce, similar to $H(B) = H(\mathbb{R}^3_+)$, a function space $H(Q)$ on $Q = B \setminus \overline{\Omega}$ by closing $C_0^\infty(\overline{Q})$ with respect to the H^1-seminorm, which will be denoted by $\|\cdot\|_{H(Q)}$. The space $H(Q)$ has properties similar to those of $H(B)$. In particular, there is a continuous trace operator $\gamma_{Q\to T}$ from $H(Q)$ to $H(T)$, and $H(Q)$ is continuously embedded in $H^1(G \setminus \overline{\Omega})$ for any bounded neighborhood $G \subset \mathbb{R}^3_+$ of Ω. For $u \in H(Q)$ we can thus define a normalized trace operator

$$\gamma_{Q\to\Sigma} v = v - \frac{\mathbf{1}}{|\Sigma|}\int_\Sigma v \,\mathrm{d}o, \qquad v \in H(Q). \tag{4.1}$$

Here, $|\Sigma|$ is the volume of the surface Σ, and $\gamma_{Q\to\Sigma}$ is a bounded and surjective operator from $H(Q)$ onto

$$H(\Sigma) = \left\{ v \in H^{1/2}(\Sigma) : \int_\Sigma v \,\mathrm{d}o = 0 \right\}.$$

In accordance with $H(\Sigma)$ we also introduce the function space

$$H(\Omega) = \left\{ w \in H^1(\Omega) : \int_\Sigma w \,\mathrm{d}o = 0 \right\},$$

which, again, can be equipped with the H^1-seminorm, so that the usual trace operator $\gamma_{\Omega\to\Sigma}$ maps $H(\Omega)$ continuously onto $H(\Sigma)$. We mention that the need for a Poincaré-type inequality is the reason to enforce vanishing means over Σ for elements from $H(\Omega)$.[1]

The framework of Gebauer also requires a linkage between the spaces $H(B)$, $H(Q)$, and $H(\Omega)$. In particular, we need to define "restriction operators" $E_Q : H(B) \to H(Q)$ and $E_\Omega : H(B) \to H(\Omega)$. In fact, we can take the natural restriction for E_Q, i.e., $E_Q u = u|_Q$, but we need to be more careful in the definition of E_Ω: Similarly to (4.1), we let

$$E_\Omega u = u|_\Omega - \frac{\mathbf{1}}{|\Sigma|}\int_\Sigma u \,\mathrm{d}o, \qquad u \in H(B), \tag{4.2}$$

such that the compatibility condition $\gamma_{Q\to\Sigma}E_Q = \gamma_{\Omega\to\Sigma}E_\Omega$ holds true.

Classical extension operators

$$\gamma^-_{Q\to\Sigma} : H(\Sigma) \to H(Q) \qquad \text{and} \qquad \gamma^-_{\Omega\to\Sigma} : H(\Sigma) \to H(\Omega)$$

yield continuous right inverses of the two "trace operators." Note that $\gamma_{\Omega\to\Sigma}$ has a continuous extension to the classical trace operator $\hat{\gamma}_{\Omega\to\Sigma} : H^1(\Omega) \to H^{1/2}(\Sigma)$, and likewise, $\gamma^-_{\Omega\to\Sigma}$ has a continuous extension to a right inverse $\hat{\gamma}^-_{\Omega\to\Sigma} : H^{1/2}(\Sigma) \to H^1(\Omega)$ of $\hat{\gamma}_{\Omega\to\Sigma}$ by setting $\hat{\gamma}^-_{\Omega\to\Sigma}\mathbf{1} = \mathbf{1}$.

In addition we need to construct continuous right inverses E^-_Q and E^-_Ω of E_Q and E_Ω, respectively. To this end we set

$$E^-_\Omega w = \begin{cases} w & \text{on } \Omega, \\ \gamma^-_{Q\to\Sigma}\gamma_{\Omega\to\Sigma} w & \text{on } Q, \end{cases} \qquad \text{and} \qquad E^-_Q v = \begin{cases} \hat{\gamma}^-_{\Omega\to\Sigma} v|_\Sigma & \text{on } \Omega, \\ v & \text{on } Q. \end{cases}$$

It follows, e.g., from Renardy and Rogers [23, Lemma 6.85], that these piecewise defined functions belong to $H^1_{\mathrm{loc}}(\mathbb{R}^3_+)$, and that E^-_Ω and E^-_Q are continuous operators. Moreover, we obviously have the compatibility requirement that $E_Q E^-_\Omega w = 0$ whenever $\gamma_{\Omega\to\Sigma} w = 0$. The corresponding requirement for the case that $\gamma_{Q\to\Sigma} v = 0$ for $v \in H(Q)$ is slightly more complicated: In this case we necessarily have that $v|_\Sigma$ is constant over Σ, and hence, $\hat{\gamma}^-_{\Omega\to\Sigma} v|_\Sigma$ is a constant also. This shows that the restriction of $E^-_Q v$ to Ω is a constant function, and hence $E_\Omega E^-_Q v = 0$ by virtue of (4.2).

Finally, given

$$\psi \in H'(\Sigma) = \left\{ \psi \in H^{-1/2}(\Sigma) : \int_\Sigma \psi \,\mathrm{d}o = 0 \right\},$$

the variational problem

$$\int_Q \nabla v \cdot \nabla w \,\mathrm{d}x = \int_\Sigma \psi w \,\mathrm{d}o \qquad \text{for all } w \in H(Q) \tag{4.3}$$

has a unique solution $v \in H(Q)$, and this solution can be used to introduce the operator

$$L : H'(\Sigma) \to H(T), \qquad \psi \mapsto v|_T, \tag{4.4}$$

[1] When Ω consists of more than one connected component, the elements of $H(\Sigma)$ and $H(\Omega)$ need to have a vanishing mean over each connected component of Σ. The trace operator (4.1) then needs to be modified accordingly, i.e., by subtracting from v different constants on the different components of Σ. A similar comment applies to the restriction operator E_Ω of (4.2).

which will play a fundamental role in what follows. As in section 2, it can be shown that v is the physically relevant (weak) solution of the exterior Neumann problem

$$\Delta v = 0 \quad \text{in } Q, \qquad \frac{\partial v}{\partial \nu} = \psi \quad \text{on } \Sigma, \qquad \frac{\partial v}{\partial e_3} = 0 \quad \text{on } \mathbb{R}^2. \tag{4.5}$$

Now we can formulate our first main result, using the notation $\mathcal{R}(A)$ to denote the range space of some operator A.

THEOREM 4.1. *Under the assumptions* (1.2), (1.3), *there holds*

$$\mathcal{R}\big(|\Lambda_\sigma - \Lambda_{\mathbf{1}}|^{1/2}\big) = \mathcal{R}(L),$$

where L is given by (4.4).

Proof. The assertion is an immediate consequence of Theorem 3.1 in [12]. Except for the straightforward discussion of the bilinear forms occurring in [12], we have already verified all the assumptions of this theorem. Making use of the standard identification of $H'(\Sigma)$ with the dual space of $H(\Sigma)$, employing $L^2(\Sigma)$ as pivot space in the Gelfand triple, it is also obvious that the operator L of (4.4) is nothing but a reformulation of the operator L defined in [12]. □

We mention that the operator L of (4.4) and its dual operator appear naturally in a factorization of the difference of the two measurement operators,

$$\Lambda_\sigma - \Lambda_{\mathbf{1}} = LFL'$$

(cf. [12]), hence the name of the factorization method. Within the framework of Gebauer, an explicit derivation of this factorization and the operator F, in particular, is not necessary. In fact, a specification of F requires the introduction of some additional diffraction problems, similar to the ones in [6, 7]: Since we never need to return to this operator, we omit the details here, but rather refer the reader to [25] or the aforementioned papers for the details.

5. The range test.

The range identity of Theorem 4.1 can be exploited to characterize the set Ω, since the range of L is easy to describe.

THEOREM 5.1. *Let $z \in \mathbb{R}^3_+$ be arbitrarily chosen. Then, for every $d \in \mathbb{R}^3 \setminus \{0\}$ the function*

$$g_{z,d}(x) = \frac{d \cdot (x-z)}{|x-z|^3}, \qquad x \in T, \tag{5.1}$$

belongs to $\mathcal{R}(L)$ if and only if $z \in \Omega$.

Proof. We first observe that $g_{z,d} = u_{z,d}|_T$, where

$$u_{z,d}(x) = \frac{1}{2}\, d \cdot \nabla_z \left(\frac{1}{|x-z|} + \frac{1}{|x-z'|} \right), \qquad x \in \mathbb{R}^3_+ \setminus \{z\},$$

is the superposition of two dipole potentials in z and z'. Here, $z' = z - 2(z \cdot e_3)e_3$ is the reflection of z with respect to the plane $\mathbb{R}^2$. Therefore, $u_{z,d}$ is a harmonic function in $\mathbb{R}^3_+ \setminus \{z\}$ with zero flux across $\mathbb{R}^2$. Moreover, $u_{z,d}$ belongs to $H(Q)$ if and only if $z \in \Omega$. Therefore, if $z \in \Omega$ and $\psi = \psi_{z,d}$ is the flux of $u_{z,d}$ across Σ into Ω, then $u_{z,d}$ is the solution of the exterior Neumann problem (4.5). Note that $\psi_{z,d}$ belongs to $H^{-1/2}(\Sigma)$; see, e.g., Girault and Raviart [13, Theorem 2.5]. Finally, we have for

$z \in \Omega$ that

$$\begin{aligned}\int_\Sigma \psi_{z,d}\,\mathrm{d}o_x &= \frac{1}{2}\, d\cdot\nabla_z \left(\int_\Sigma \frac{\partial}{\partial\nu_x}\frac{1}{|x-z|}\,\mathrm{d}o_x + \int_\Sigma \frac{\partial}{\partial\nu_x}\frac{1}{|x-z'|}\,\mathrm{d}o_x \right)\\ &= \frac{1}{2}\, d\cdot\nabla_z\big(4\pi\mathbf{1}(z)\big) = 0\end{aligned}$$

(see, e.g., [19, Example 6.16]), which shows that $\psi_{z,d} \in H'(\Sigma)$. We therefore have proved that $L\psi_{z,d} = g_{z,d}$, i.e., that $g_{z,d} \in \mathcal{R}(L)$ for $z \in \Omega$.

Now let $z \in \mathbb{R}^3_+ \setminus \Omega$, and assume that $g_{z,d} \in \mathcal{R}(L)$, i.e., that $g_{z,d} = L\psi$ for some $\psi \in H'(\Sigma)$. This is equivalent to the statement that $g_{z,d} = v|_T$, where $v \in H(Q)$ is the weak solution of (4.5). Thus, $u_{z,d}$ and v are two harmonic functions in $\mathbb{R}^3_+ \setminus (\{z\} \cup \overline{\Omega})$ which share the same Cauchy data on $\mathbb{R}^2$. By the uniqueness of the Cauchy problem (see, e.g., [9, Chapter II]) the two functions must be the same in $\mathbb{R}^3_+ \setminus (\{z\} \cup \overline{\Omega})$. This, however, contradicts the fact that $u_{z,d}$ has a singularity in z and, hence, does not belong to $H(Q)$. Therefore we have shown that $g_{z,d} \notin \mathcal{R}(L)$ whenever $z \in \mathbb{R}^3_+ \setminus \Omega$. □

As a corollary of Theorems 4.1 and 5.1 we obtain the following useful range test to decide whether some point $z \in \mathbb{R}^3_+$ belongs to Ω or not.

Corollary 5.2. *A point $z \in \mathbb{R}^3_+$ belongs to Ω if and only if the function $g_{z,d}$ of Theorem* 5.1 *belongs to the range of $|\Lambda_\sigma - \Lambda_{\mathbf{1}}|^{1/2}$.*

6. Numerical results. We now present a numerical realization of the range test of Corollary 5.2 for simulated data in three space dimensions. Data are given on $T = \Gamma = [0,2]^2$, shown as the somewhat darker area of the bounding plane in the subsequent figures. In all examples to follow, data have been generated by a boundary element method, with the conductivity within the inclusion being set to $\kappa = 0.5$. Modifications of κ have a negligible effect on the reconstructions, provided that (1.3) is satisfied for any small ε; this has been demonstrated convincingly in [7] for bounded domains in two space dimensions.

A very detailed discussion of the general approach for implementing the range test can be found in [7, 14], so here we focus mainly on the differences that are important for this half-space problem.

The first major difference is the fact that data are given on a two-dimensional interval rather than a one-dimensional interval. We have found it convenient to use tensor products of piecewise constant Haar wavelets (with vanishing mean over Γ) as current patterns and to expand the simulated potentials in the same orthogonal basis. The data we use thus correspond to the Galerkin projection of $\Lambda_\sigma - \Lambda_{\mathbf{1}}$ onto the space of the particular current patterns. All our computations use the corresponding first 1023 basis functions, which are far more than is required for the resolution of our reconstructions due to the inevitable presence of noise in the data.

Figure 6.1 reveals a second major difference from the results in [7, 14], which appears to be a characteristic property of the factorization method in three space dimensions. The eigenvalues of $\Lambda_\sigma - \Lambda_{\mathbf{1}}$ do not obey a strict geometric decay; rather, they tend to come in clusters of increasing size. Note that, in theory, the function $g_{z,d}$ belongs to the range of $|\Lambda_\sigma - \Lambda_{\mathbf{1}}|^{1/2}$ if and only if the corresponding Picard series

$$\sum_{j=1}^{\infty} \frac{\langle g_{z,d}, v_j\rangle^2_{L^2(\Gamma)}}{|\lambda_j|} \tag{6.1}$$

converges; here v_j, $j \in \mathbb{N}$, are the orthonormal eigenfunctions of $\Lambda_\sigma - \Lambda_{\mathbf{1}}$, and λ_j are the associated eigenvalues. In [7, 14] we have estimated the geometric decay of the

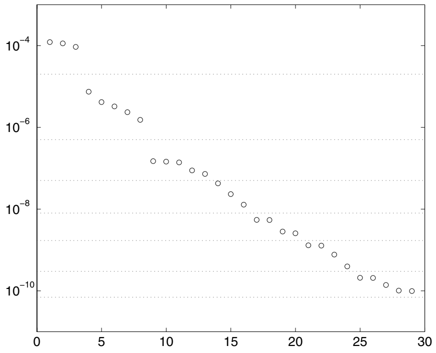


FIG. 6.1. *First test case: eigenvalues of* $\Lambda_\sigma - \Lambda_{\mathbf{1}}$.

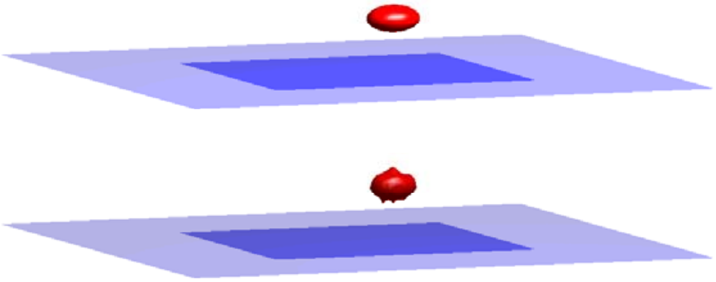


FIG. 6.2. *First test case: an ellipsoidal object (top) and its reconstruction (bottom).*

individual terms of this series to decide whether we believe that (6.1) converges or not. Here, instead, we have decided to average the eigenvalue clusters and investigate the root convergence factor of the geometric decay of the associated partial sums. The eigenvalue plot in Figure 6.1 (and similarly in Figure 6.4) contains dotted lines to indicate the eigenvalues that were considered to be clustered. The clustering has always been performed manually and is optimized to some extent to improve the quality of the reconstructions. Eigenvalue clusters below 10^{-10} have been ignored (except for section 6.4).

6.1. First test case. In the first example, which we have already mentioned, the object to be reconstructed is an ellipsoid with center in $P = (1.2, 0.8, 0.4)$ as shown in Figure 6.2. Its semiaxes are aligned with the coordinate axes and have radii $r_1 = 0.2$, $r_2 = 0.15$, and $r_3 = 0.1$. This isosurface plot is based on a certain average of the root convergence factors obtained from nine different dipole moments d_k, $k = 1, \dots, 9$. (We refer the reader to [25] for further details.) We emphasize, as this might be difficult to see, that the reconstruction is at the correct place and has about the right size. It is only the boundary which is not accurate. Alternatively, we have also evaluated the series (6.1) for the respective range of eigenvalues and have used this function of z for a surface plot, as was done, e.g., by Kirsch in [18]. However, this gave somewhat inferior reconstructions.

6.2. Second test case. Our second example (see Figures 6.3 and 6.4) consists of two objects. One is an ellipsoid with center in $P = (0.4, 0.4, 0.4)$ and radii

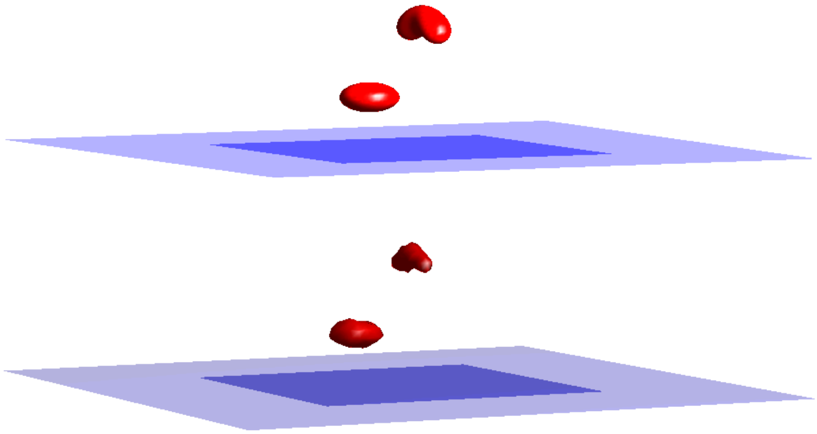


FIG. 6.3. *Second test case: two objects (top) and their reconstruction (bottom).*

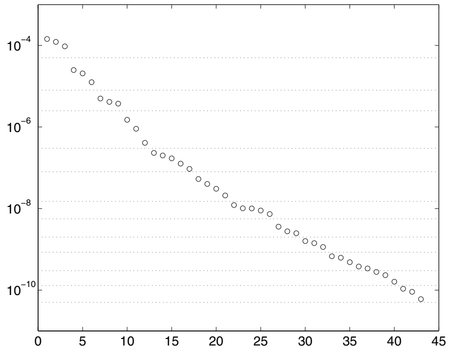


FIG. 6.4. *Eigenvalues for the second test case.*

$r_1 = r_2 = 0.2$ and $r_3 = 0.1$, respectively; again, the semiaxes are aligned with the coordinate axes. The other object has the shape of a kidney and is located around the point $Q = (1.2, 1.2, 0.8)$. The corresponding reconstructions are again at the correct locations. Note that the nonconvexity of the kidney is still well depicted, although it is a little farther away from Γ. On the other hand, its reconstruction is somewhat too small. If the nonconvex boundary is turned upwards, however, the reconstruction is qualitatively worse.

6.3. Third test case. The third test case is similar to the previous one, but now the ellipsoidal object is moved off to the side; i.e., its orthogonal projection onto $\mathbb{R}^2$ is outside of Γ; see Figure 6.5. More precisely, the ellipsoid of the second test case now has its center at $R = (-0.2, -0.2, 0.4)$. Our method reconstructs both objects at their true locations, but the reconstruction of the ellipsoid exhibits typical shady artifacts, similar to two-dimensional reconstructions shown in [14].

6.4. Fourth test case. For the next experiment we return to the ellipsoid from our first example, and increase its vertical distance to the plane. Figure 6.6 shows the reconstructions for three snapshots. As one expects, the quality deteriorates with

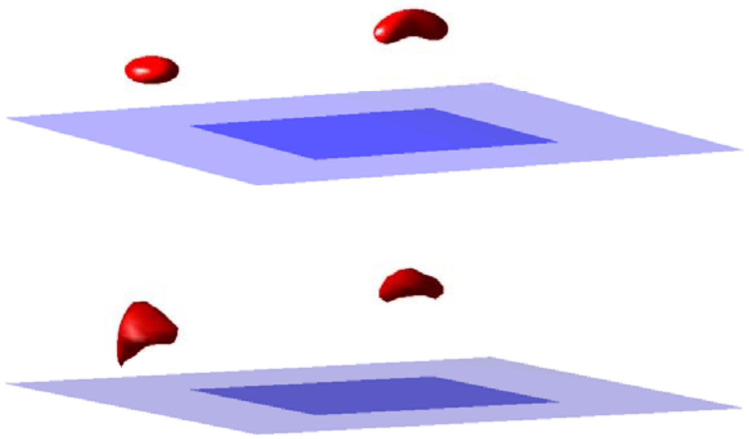
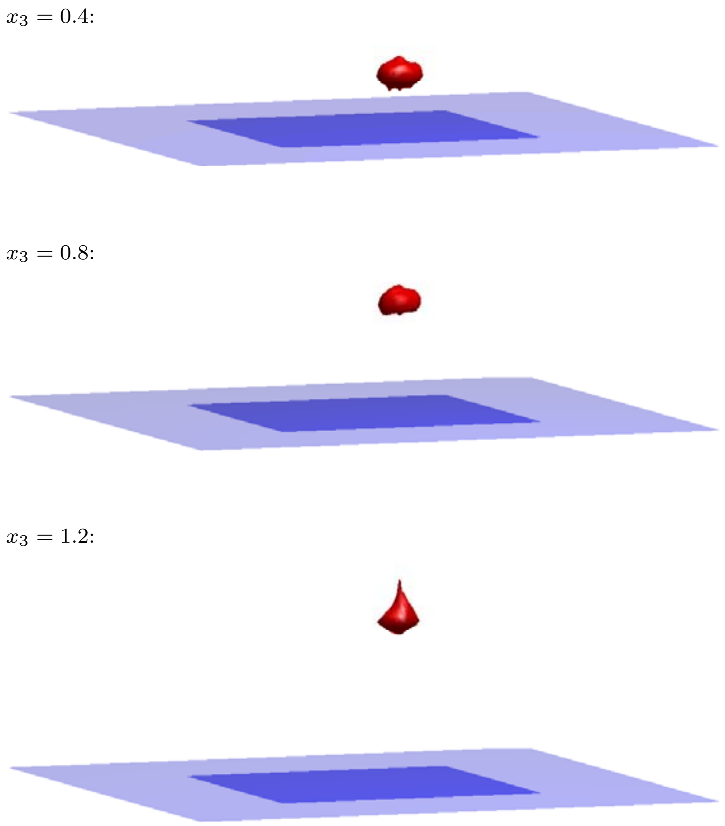


FIG. 6.5. *Third test case: two objects (top), one being off to the side, and their reconstruction (bottom).*



$x_3 = 0.4$:

$x_3 = 0.8$:

$x_3 = 1.2$:

FIG. 6.6. *Fourth test case: reconstructions of ellipsoids with increasing vertical heights.*

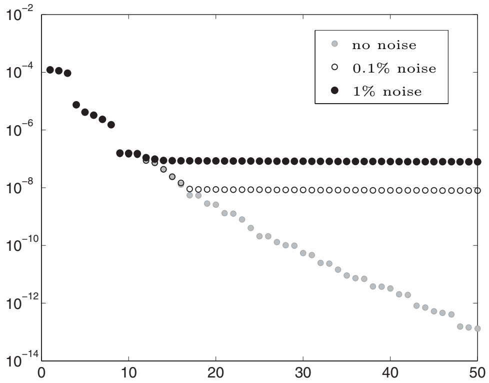
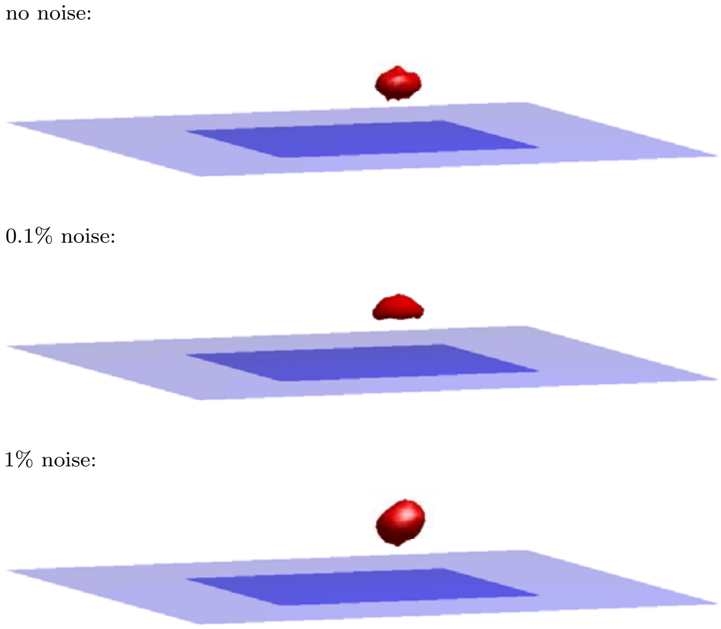


FIG. 6.7. *Eigenvalues in the presence of noise.*



no noise:

0.1% noise:

1% noise:

FIG. 6.8. *Reconstructions in the presence of noise.*

increasing distance x_3, measured at the center of the ellipsoid; see Figure 6.6. For these reconstructions we have used a slightly larger range of eigenvalues, going down to 10^{-12}.

6.5. Fifth test case. In a final study, we investigate the influence of noise on our reconstructions. To this end we superpose the data of our first test case (cf. Figure 6.2) with 0.1% and 1% noise, respectively. (These noise levels refer to the L^2-norms of the noise over the L^2-norm of the exact data.) Figure 6.7 shows the resulting eigenvalues of $\Lambda_\sigma - \Lambda_{\mathbf{1}}$. It is easy to see how the eigenvalues level off in the presence of noise, from which we can easily determine which eigenvalues can reliably be used to perform the range test. Figure 6.8 shows the corresponding reconstructions, which are quite reasonable even with 1% noise (bottom reconstruction).

7. The two-dimensional case. In this section we briefly comment on the modifications of our theory in two space dimensions; as a general reference we refer the reader to [25]. In two dimensions, solutions of the boundary value problem

$$\nabla \cdot \sigma \nabla u = 0 \quad \text{in } \mathbb{R}^2_+, \qquad -\sigma \frac{\partial u}{\partial e_2} = f \quad \text{on } \mathbb{R}, \tag{7.1}$$

are unique (up to additive constants) within the space $H^{1,0+}(\mathbb{R}^2_+)$ which is obtained by closing either $C^\infty(\mathbb{R}^2_+)$ or $C_0^\infty(\mathbb{R}^2_+)$ with respect to the inner product (2.6) for any $\alpha > 0$ (replacing the integrals by integrals over $\mathbb{R}^2_+$, of course). These spaces all contain the same functions, independent of the choice of $\alpha > 0$, including in particular the constant functions. We can get rid of these constants by turning to the quotient space $H(\mathbb{R}^2_+) = H^{1,0+}(\mathbb{R}^2_+)/\operatorname{span}\{\mathbf{1}\}$, for which we can use the H^1-seminorm as an equivalent norm.

Investigating the weak formulation of (7.1), the existence of a solution in $H^{1,0+}(\mathbb{R}^2_+)$ is guaranteed provided that the imposed current f belongs to

$$L^{2,-1-\alpha}_\diamond(\mathbb{R}) = \left\{ f : (1+|y|^2)^{1/2+\alpha/2} f \in L^2(\mathbb{R}) : \int_T f \,\mathrm{d}y = 0 \right\}$$

for some $\alpha > 0$; note that the normalization condition $\int_T f \,\mathrm{d}y = 0$ has not been required in the three-dimensional case.

Since the solution u of (7.1) is unique only up to additive constants, it is necessary to normalize the trace of u to set up a well-defined associated Neumann-to-Dirichlet operator. Accordingly, the general framework developed in section 4 requires some obvious changes for two space dimensions; in particular, a similar normalization is required in the definition of the operator L of (4.4). With these modifications, however, the result of Theorem 4.1 remains true, and a valid test function to be used in Theorem 5.1 (again, up to a suitable additive constant) is given by

$$g_{z,d}(x) = \frac{d \cdot (x-z)}{|x-z|^2}, \qquad x \in T. \tag{7.2}$$

We refer the reader to [25] for several numerical reconstructions in two space dimensions; preliminary results had been published in [14] and [24].

Appendix. In this appendix we prove that the weighted Sobolev space $H_{1/2+}(\mathbb{R}^3_+)$ introduced in Remark 2.3 is the direct sum

$$H_{1/2+}(\mathbb{R}^3_+) = H(\mathbb{R}^3_+) \oplus \operatorname{span}\{\mathbf{1}\}.$$

In the proof of this result we use the following Liouville-type theorem on bounded harmonic functions in the entire space, which appears to be of independent interest.

THEOREM A.1. *Every harmonic function u over $\mathbb{R}^3$ which satisfies*

$$\int_{\mathbb{R}^3} \frac{|u(x)|^2}{(1+|x|^2)^{5/2}} \,\mathrm{d}x < \infty \tag{A.1}$$

is a constant.

Proof. Our proof makes use of an appropriate modification of the argument given in Axler, Bourdon, and Ramey [1], which starts with the mean-value property of harmonic functions, to write

$$\begin{aligned}|u(x^*)-u(0)| &= \frac{3}{4\pi r^3}\left|\int_{B_r(x^*)} u(x)\,\mathrm{d}x - \int_{B_r(0)} u(x)\,\mathrm{d}x\right| \\ &= \frac{3}{4\pi r^3}\left|\int_{D_r} s(x)u(x)\,\mathrm{d}x\right|\end{aligned}$$

for any fixed $x^* \in \mathbb{R}^3$. In this equation $B_r(y)$ denotes the ball of radius r around y, $D_r = B_r(x) \cup B_r(0) \setminus (B_r(x) \cap B_r(0))$ is the symmetric difference of the two balls, and s is a sign function that attains the two values ± 1 in the respective components of D_r. We denote $|x^*|$ by r^* and restrict r to be larger than $r_0 \geq 2r^* + 1$ in what follows. Then D_r is contained in the annulus

$$A_r = \{x \in \mathbb{R}^3 : r - r^* < |x| < r + r^*\},$$

and we can estimate

$$|u(x^*)-u(0)| \leq \frac{3}{4\pi r^3}\int_{A_r} |u(x)|\,\mathrm{d}x \leq c\int_{A_r} \frac{|u(x)|}{(1+|x|^2)^{3/2}}\,\mathrm{d}x,$$

where, from now on, we use c to denote a generic positive constant, depending only on x^*. Integrating the above inequality from $r = r_0$ to some $R > r_0$, we obtain

$$\begin{aligned}|u(x^*)-u(0)| &\leq \frac{c}{R-r_0}\int_{r_0}^{R}\int_{A_r} \frac{|u(x)|}{(1+|x|^2)^{3/2}}\,\mathrm{d}x\,\mathrm{d}r \\ &\leq \frac{2r^*c}{R-r_0}\int_{r_0-r^*<|x|<R+r^*} \frac{|u(x)|}{(1+|x|^2)^{3/2}}\,\mathrm{d}x.\end{aligned}$$

Thus, the Cauchy–Schwarz inequality yields

$$\begin{aligned}|u(x^*)-u(0)|^2 &\leq \frac{c}{(R-r_0)^2}\int_{|x|>r_0-r^*} \frac{|u(x)|^2}{(1+|x|^2)^{5/2}}\,\mathrm{d}x \int_{|x|<R+r^*} \frac{1}{(1+|x|^2)^{1/2}}\,\mathrm{d}x \\ &\leq \frac{c}{(R-r_0)^2}\int_{|x|>r_0-r^*} \frac{|u(x)|^2}{(1+|x|^2)^{5/2}}\,\mathrm{d}x \int_0^{R+r^*} (1+r^2)^{1/2}\,\mathrm{d}r \\ &\leq c\left(\frac{R+r^*+1}{R-r_0}\right)^2 \int_{|x|>r_0-r^*} \frac{|u(x)|^2}{(1+|x|^2)^{5/2}}\,\mathrm{d}x.\end{aligned}$$

Now, if R is sufficiently large, then we can choose $r_0 = R/2$ and thus obtain

$$|u(x^*)-u(0)|^2 \leq c\int_{|x|>R/2-r^*} \frac{|u(x)|^2}{(1+|x|^2)^{5/2}}\,\mathrm{d}x = o(1)$$

as $R \to \infty$. It follows that $u(x^*) = u(0)$, i.e., that u is a constant. □

We mention that this result is sharp in that all polynomials u in x of exact degree one are harmonic in $\mathbb{R}^3$ and satisfy (A.1) for any exponent in the denominator bigger than 5/2.

Now we turn to verify (2.7). Let $w \in H_{1/2+}(\mathbb{R}^3_+)$, and consider the variational problem

$$\int_{\mathbb{R}^3_+} \nabla w_0(x) \cdot \nabla v(x)\,\mathrm{d}x = \int_{\mathbb{R}^3_+} \nabla w(x) \cdot \nabla v(x)\,\mathrm{d}x \qquad \text{for all } v \in H(\mathbb{R}^3_+).$$

This problem has a unique solution $w_0 \in H(\mathbb{R}^3_+)$, and it follows that $u = w - w_0 \in H_{1/2+}(\mathbb{R}^3_+)$ satisfies

$$\int_{\mathbb{R}^3_+} \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x = 0$$

for all $v \in C_0^\infty(\overline{\mathbb{R}^3_+})$, and hence, according to Weyl's lemma, u is a harmonic function in $\mathbb{R}^3_+$ with vanishing Neumann boundary values on the boundary of this half-space. Thus, u can be extended by reflection to an even harmonic function $\tilde{u}$ over the entire space $\mathbb{R}^3$; cf., e.g., [1]. As $u \in H_{1/2+}(\mathbb{R}^3_+)$ and hence has finite norm (2.6) for any $\alpha > 1/2$, it follows that $\tilde{u}$ satisfies (A.1). Thus $\tilde{u}$ and u are constant functions by virtue of Theorem A.1, and we have shown that any function $w \in H_{1/2+}(\mathbb{R}^3_+)$ can be decomposed in a unique way as $w = w_0 + c$, where $w_0 \in H(\mathbb{R}^3_+)$ and c is some constant. This proves (2.7).

REFERENCES

[1] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, 2nd ed., Springer, New York, 2001.

[2] M. Azzouz, M. Hanke, C. Oesterlein, and K. Schilcher, *The factorization method for electrical impedance tomography data from a new planar device*, Int. J. Biomed. Imaging, 2007 (2007), Article ID 83016.

[3] T. Z. Boulmezaoud, *On the Laplace operator and on the vector potential problems in the half-space: An approach using weighted spaces*, Math. Methods Appl. Sci., 26 (2003), pp. 633–669.

[4] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[5] L. Borcea, *Addendum to "Electrical impedance tomography,"* Inverse Problems, 19 (2003), pp. 997–998.

[6] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[7] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[8] A. P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Application to Continuum Physics, W. H. Meyer and M. A. Raupp, eds., Soc. Brasil Mat., Rio de Janeiro, 1980, pp. 65–73; reprinted in Comput. Appl. Math., 25 (2006), pp. 133–138.

[9] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Volume* 1: *Physical Origins and Classical Methods*, Springer, New York, 2000.

[10] J. Deny and J.-L. Lions, *Les espaces du type de Beppo Levi*, Ann. Inst. Fourier (Grenoble), 5 (1954), pp. 305–370.

[11] V. Druskin, *On the uniqueness of inverse problems from incomplete boundary data*, SIAM J. Appl. Math., 58 (1998), pp. 1591–1603.

[12] B. Gebauer, *The factorization method for real elliptic problems*, Z. Anal. Anwend., 25 (2006), pp. 81–102.

[13] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.

[14] M. Hanke and M. Brühl, *Recent progress in electrical impedance tomography*, Inverse Problems, 19 (2003), pp. 65–90.

[15] B. Hanouzet, *Espaces de Sobolev avec poids—application au problème de Dirichlet dans un demi espace*, Rend. Sem. Mat. Univ. Padova, 46 (1971), pp. 227–272.

[16] T. A. Hope and S. Iles, *Technology review: The use of electrical impedance scanning in the detection of breast cancer*, Breast Cancer Res., 6 (2004), pp. 69–74.

[17] R. Janßen, *Elliptic problems on unbounded domains*, SIAM J. Math. Anal., 17 (1986), pp. 1370–1389.

[18] A. Kirsch, *Characterization of the shape of the scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[19] R. Kress, *Linear Integral Equations*, 2nd ed., Springer, New York, 1999.

[20] M. Lukaschewitsch, *Inversion of Geoelectric Boundary Data, a Non-Linear Ill-Posed Problem*, Dissertation, Universität Potsdam, Potsdam, Germany, 1999.

[21] M. Lukaschewitsch, P. Maass, and M. Pidcock, *Tikhonov regularization for electrical impedance tomography on unbounded domains*, Inverse Problems, 19 (2003), pp. 585–610.

[22] J. L. Mueller, D. Isaacson, and J. C. Newell, *Reconstruction of conductivity changes due to ventilation and perfusion from EIT data collected on a rectangular electrode array*, Physiol. Meas., 22 (2001), pp. 97–106.

[23] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer, New York, 1993.

[24] B. Schappel, *Electrical impedance tomography in the half space: Locating obstacles by electrostatic measurements on the boundary*, in Proceedings of the 3rd World Congress on Industrial Process Tomography, Banff, AB, Canada, VCIPT, 2003, pp. 788–793.

[25] B. Schappel, *Die Faktorisierungsmethode für die elektrische Impedanztomographie im Halbraum*, Dissertation, Johannes Gutenberg-Universität Mainz, Mainz, Germany, 2005; available online from http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hebis:77-7427.

[26] C. van Berkel and W. R. B. Lionheart, *Reconstruction of a grounded object in an electrostatic halfspace with an indicator function*, Inverse Probl. Sci. Eng., 15 (2007), pp. 585–600.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 925–946

A VARIATIONAL APPROACH TO REMOVING MULTIPLICATIVE NOISE[*]

GILLES AUBERT[†] AND JEAN-FRANÇOIS AUJOL[‡]

Abstract. This paper focuses on the problem of multiplicative noise removal. We draw our inspiration from the modeling of speckle noise. By using a MAP estimator, we can derive a functional whose minimizer corresponds to the denoised image we want to recover. Although the functional is not convex, we prove the existence of a minimizer and we show the capability of our model on some numerical examples. We study the associated evolution problem, for which we derive existence and uniqueness results for the solution. We prove the convergence of an implicit scheme to compute the solution.

Key words. calculus of variation, functional analysis, BV, variational approach, multiplicative noise, speckle noise, image restoration

AMS subject classifications. 68U10, 94A08, 49J40, 35A15, 35B45, 35B50

DOI. 10.1137/060671814

1. Introduction.

Image denoising is a widely studied problem in the applied mathematics community. We refer the reader to [4, 14] and references therein for an overview of the subject. Most of the literature deals with the additive noise model: given an original image u, it is assumed that it has been corrupted by some additive noise v. The problem is then to recover u from the data $f = u + v$. Many approaches have been proposed. Among the most famous are wavelet approaches [17], stochastic approaches [21], and variational approaches [37, 30].

In this paper, we are concerned with a different denoising problem. The assumption is that the original image u has been corrupted by some multiplicative noise v: the goal is then to recover u from the data $f = uv$. Multiplicative noise occurs as soon as one deals with active imaging systems: laser images, microscope images, synthetic aperture radar (SAR) images, etc. As far as we know, the only variational approach devoted to multiplicative noise is the one by Rudin, Lions, and Osher [36] as used, for instance, in [33, 28, 29, 38]. The goal of this paper is to go further and to propose a functional that is well adapted to removing multiplicative noise. Inspired from the modeling of active imaging systems, this functional is

$$E(u) = \int |Du| + \int \left(\log u + \frac{f}{u} \right),$$

where f is the original corrupted image and $\int |Du|$ stands for the total variation of u.

From a mathematical point of view, part of the difficulty comes from the fact that, contrary to the additive case, the proposed model is nonconvex, which causes uniqueness problems, as well as the issue of convergence of the algorithms. Another mathematical issue comes from the fact that we deal with a linear growth functional.

[*] Received by the editors October 9, 2006; accepted for publication (in revised form) July 16, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/siap/68-4/67181.html

[†] Laboratoire J. A. Dieudonné, UMR CNRS 6621, University of Nice-Sophia Antipolis, 06108 Nice Cedex 2, France (gaubert@math.unice.fr).

[‡] CMLA, ENS Cachan, CNRS, PRES UniverSud, 94235 Cachan, France (Jean-Francois.Aujol@cmla.ens-cachan.fr).

The natural space in which we compute a solution is BV, the space of functions with bounded variations. But contrary to what happens with classical Sobolev spaces, the minimum of the functional does not verify an associated Euler–Lagrange equation (see [3] and [2] where this problem is studied) but a differential inclusion involving the subdifferential of the energy.

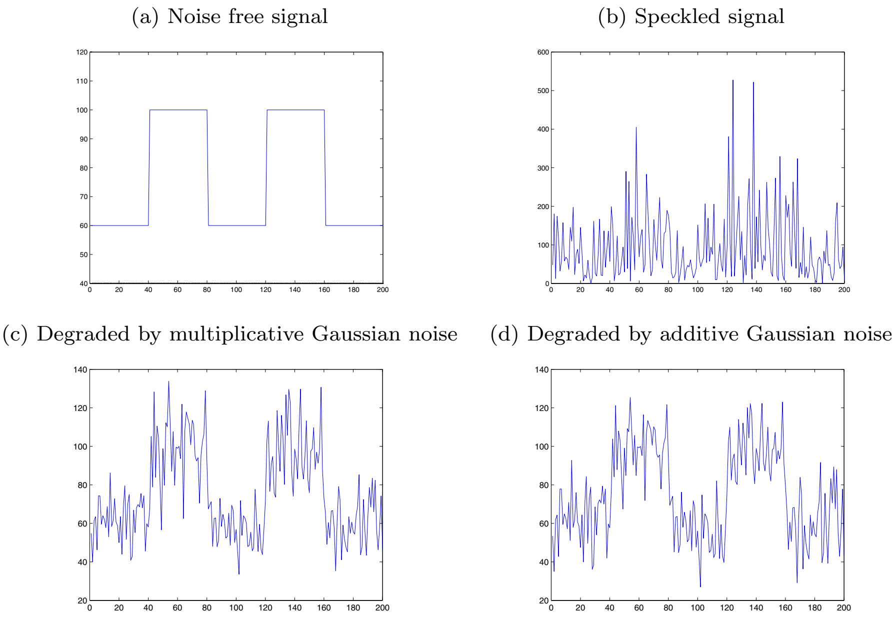
The paper is organized as follows. We draw our inspiration from the modeling of active imaging systems, which we describe to the reader in section 2. We use the classical MAP estimator to derive a new model to denoise nontextured SAR images in section 3. We then consider this model from a variational point of view in section 4 and we carry out the mathematical analysis of the functional in the continuous setting. In section 5 we illustrate our model by displaying some numerical examples. We also compare it with other models. Then in section 6 we study the evolution equation associated to the problem. To prove the existence and the uniqueness of a solution to the evolution problem we first consider a semi-implicit discretization scheme and then we let the discretization time step go to zero. The proofs are rather technical, and we give them in the appendix.

2. Speckle noise modeling. SAR images are strongly corrupted by a noise called speckle. A radar sends a coherent wave which is reflected on the ground and then registered by the radar sensor [26, 31]. If the coherent wave is reflected on a coarse surface (compared to the radar wavelength), then the image processed by the radar is degraded by a noise with large amplitude: this gives a speckled aspect to the image, and this is the reason such a noise is called speckle [24]. To illustrate the difficulty of speckle noise removal, Figure 1 shows a 1-dimensional (1D) noise free signal and the corresponding speckled signal (the noise free signal has been multiplied by a speckle noise of mean 1). It can be seen that almost all the information has disappeared (notice in particular that the vertical scale goes from 40 to 120 for the noise free signal presented in Figure 1(a), whereas it goes from 0 to 600 on the speckled signal presented in Figure 1(b)). As a comparison, Figure 1(c) shows the 1D signal of Figure 1(a) once it has been multiplied by a Gaussian noise of mean 1 and standard deviation 0.2 (as used, for instance, in [36]), and Figure 1(d) shows the 1D signal of Figure 1(a) with the addition of a Gaussian noise of zero mean and standard deviation $\sigma = 15$ (notice that for both Figures 1(c) and (d), the vertical scale goes from 20 to 140).

If we denote by I the image intensity considered as a random variable, then I follows a negative exponential law. The density function is $g_I(x) = \frac{1}{\mu_I} e^{-\frac{x}{\mu_I}} \mathbf{1}_{\{x \geq 0\}}$, where μ_I is both the mean and the standard deviation of I. In general the image is obtained as the summation of L different images (this is very classical with satellite images). If we assume that the variables I_k, $1 \leq k \leq L$, are independent and have the same mean μ_I, then the intensity $J = \frac{1}{L} \sum_{k=1}^{L} I_k$ follows a gamma law, with density function $g_J(x) = \left(\frac{L}{\mu_I}\right)^L \frac{1}{\Gamma(L)} x^{L-1} \exp\left(-\frac{Lx}{\mu_I}\right) \mathbf{1}_{\{x \geq 0\}}$, where $\Gamma(L) = (L-1)!$. Moreover, μ_I is the mean of J, and $\frac{\mu_I}{\sqrt{L}}$ is its standard deviation.

The classical modeling [41] for SAR images is $I = RS$, where I is the intensity of the observed image, R the reflectance of the scene (which is to be recovered), and S the speckle noise. S is assumed to follow a gamma law with mean equal to 1: $g_S(s) = \frac{L^L}{\Gamma(L)} s^{L-1} \exp(-Ls) \mathbf{1}_{\{s \geq 0\}}$. In the rest of the paper, we will assume that the image to recover has been corrupted by some multiplicative gamma noise.

Speckle removal methods have been proposed in the literature. There are geometric filters, such as the Crimmins filter [15] based on the application of convex hull


(a) Noise free signal

(b) Speckled signal

(c) Degraded by multiplicative Gaussian noise

(d) Degraded by additive Gaussian noise

Fig. 1. *Speckle noise in one dimension: notice that the vertical scale is not the same on the different images (scale between* 40 *and* 120 *on* (a)*,* 0 *and* 600 *on* (b)*,* 20 *and* 140 *on* (c)*,* 20 *and* 140 *on* (d)*).* (a) $1D$ *signal* f*;* (b) f *degraded by speckle noise of mean* 1*;* (c) f *degraded by a multiplicative Gaussian noise* ($\sigma = 0.2$)*;* (d) f *degraded by an additive Gaussian noise* ($\sigma = 15$)*. Speckle noise is much stronger than classical additive Gaussian noise* [37] *or classical multiplicative Gaussian noise* [36].

algorithms. There are adaptive filters, such as the Lee filter, the Kuan filter, or its improvement proposed by Wu and Maitre [42]: first and second order statistics computed in local windows are incorporated in the filtering process. Adaptive filters with some modeling of the scene, such as the Frost filter, have been proposed. The criterion is based on a MAP estimator, and Markov random fields can be used as in [40, 16]. Another class of filters are multitemporal, such as the Bruniquel filter [10]: by computing barycentric means, the standard deviation of the noise can be reduced (provided that several different images of the same scene are available). A last class of methods are variational methods as in [37, 36, 6], where the solution is computed with PDEs.

3. A variational multiplicative denoising model. The goal of this section is to propose a new variational model for denoising images corrupted by multiplicative noise and in particular for SAR images. We start from the following multiplicative model: $f = uv$, where f is the observed image, $u > 0$ the image to recover, and v the noise. We consider that f, u, and v are instances of some random variables F, U, and V. In the following, if X is a random variable, we denote by g_X its density function. We refer the interested reader to [25] for further details about random variables. In this section, we consider discretized images. We denote by $\mathcal{S}$ the set of the pixels of the image. Moreover, we assume that the samples of the noise on each pixel $s \in S$ are mutually independent and identically distributed (i.i.d.) with density function g_V.

3.1. Density laws with a multiplicative model. Our goal is to maximize $P(U|F)$, and thus thanks to the Bayes rule we need to know $P(F|U)$ and $g_{F|U}$.

PROPOSITION 3.1. *Assume that U and V are independent random variables, with continuous density functions g_U and g_V. Let us set $F = UV$. Then we have for $u > 0$*

$$g_V\left(\frac{f}{u}\right)\frac{1}{u} = g_{F|U}(f|u). \tag{3.1}$$

Proof. The proof is a standard result (see [25], for instance). We give the proof here for the sake of completeness.

Let $\mathcal{A}$ be an open subset in $\mathbb{R}$. We have

$$\int_{\mathbb{R}} g_{F|U}(f|u)\mathbf{1}_{\{f\in\mathcal{A}\}} = P(F\in\mathcal{A}|U) = \frac{P(F\in\mathcal{A},U)}{P(U)} = \frac{P\left(\left(V=\frac{F}{U}\right)\in\frac{\mathcal{A}}{U},U\right)}{P(U)}.$$

Using the fact that U and V are independent, we have

$$\begin{aligned}\frac{P\left(\left(V=\frac{F}{U}\right)\in\frac{\mathcal{A}}{U},U\right)}{P(U)} = P\left(\left(V=\frac{F}{U}\right)\in\frac{\mathcal{A}}{U}\right) &= \int_{\mathbb{R}} g_V(v)\mathbf{1}_{\{v\in\frac{\mathcal{A}}{u}\}}\,dv\\ &= \int_{\mathbb{R}} g_V(f/u)\mathbf{1}_{\{f\in\mathcal{A}\}}\,\frac{df}{u}.\end{aligned}$$ □

3.2. Our model via the MAP estimator. We assume the following multiplicative model: $f = uv$, where f is the observed image, u the image to recover, and v the noise. We assume that v follows a gamma law with mean 1 and with density function

$$g_V(v) = \frac{L^L}{\Gamma(L)}v^{L-1}e^{-Lv}\,\mathbf{1}_{\{v\geq 0\}}. \tag{3.2}$$

Using Proposition 3.1, we therefore get

$$g_{F|U}(f|u) = \frac{L^L}{u^L\Gamma(L)}f^{L-1}e^{-\frac{Lf}{u}}. \tag{3.3}$$

We also assume that U follows a Gibbs prior,

$$g_U(u) = \frac{1}{Z}\exp(-\gamma\phi(u)), \tag{3.4}$$

where Z is a normalizing constant and ϕ is a nonnegative given function. We aim to maximize $P(U|F)$. This will lead us to the classical MAP estimator. From the Bayes rule, we have $P(U|F) = \frac{P(F|U)\,P(U)}{P(F)}$. Maximizing $P(U|F)$ amounts to minimizing the log-likelihood:

$$-\log(P(U|F)) = -\log(P(F|U)) - \log(P(U)) + \log(P(F)). \tag{3.5}$$

We remind the reader that the image is discretized. We denote by $\mathcal{S}$ the set of the pixels of the image. Moreover, we assume that the samples of the noise on each pixel $s\in\mathcal{S}$ are mutually i.i.d. with density g_V. We therefore have $P(F|U) = \prod_{s\in\mathcal{S}} P(F(s)|U(s))$, where $F(s)$ (resp., $U(s)$) is the instance of the variable F (resp., U) at pixel s. Since $\log(P(F))$ is a constant, we just need to minimize:

$$-\log(P(F|U)) - \log(P(U)) = -\sum_{s\in\mathcal{S}}\left(\log(P(F(s)|U(s))) - \log(P(U(s)))\right). \tag{3.6}$$

Using (3.3), and since Z is a constant, we eventually see that minimizing $-\log(P(F|U))$ amounts to minimizing

$$\sum_{s\in\mathcal{S}}\left(L\left(\log U(s)+\frac{F(s)}{U(s)}\right)+\gamma\phi(U(s))\right). \tag{3.7}$$

The previous computation leads us to propose the following functional for restoring images corrupted with gamma noise:

$$\int\left(\log u+\frac{f}{u}\right)dx+\frac{\gamma}{L}\int\phi(u)\,dx. \tag{3.8}$$

Remarks. (1) It is easy to check that the function $u\rightarrow\log u+\frac{f}{u}$ reaches its minimum value $1+\log f$ over $\mathbb{R}_*^+$ for $u=f$.

(2) Multiplicative Gaussian noise: in the additive noise case, the most classical assumption is to assume that the noise is a white Gaussian noise. However, this can no longer be the case when dealing with multiplicative noise, except in the case of tiny noise. Indeed, if the model is $f=uv$, where v is a Gaussian noise with mean 1, then some instances of v are negative. Since the data f is assumed positive, this implies that the restored image u has some negative values, which is, of course, impossible. Nevertheless, numerically, if the standard deviation of the noise is smaller than 0.2 (i.e., in the case of tiny noise), then it is very unlikely that v takes some negative values. See also [32] where some limitations of the Bayesian estimator approach are investigated.

4. Mathematical study of the variational model.

In this section, we propose a nonconvex model for removing multiplicative noise, for which we prove the existence of a solution.

4.1. Preliminaries.

Throughout our study, we will use the following classical distributional spaces. $\Omega\subset\mathbb{R}^2$ will denote an open bounded set with Lipschitz boundary.

- $\mathcal{D}(\Omega)=C_0^\infty(\Omega)$ is the set of functions in $C^\infty(\Omega)$ with compact support in Ω. We denote by $\mathcal{D}'(\Omega)$ the dual space of $\mathcal{D}(\Omega)$, i.e., the space of distributions on Ω.
- $W^{m,p}(\Omega)$ denotes the space of functions in $L^p(\Omega)$ whose distributional derivatives $D^\alpha u$ are in $L^p(\Omega)$, $p\in[1,+\infty)$, $m\geq 1$, $m\in\mathbb{N}$, $|\alpha|\leq m$. For further details on these spaces, we refer the reader to [19, 20].
- $BV(\Omega)$ is the subspace of functions $u\in L^1(\Omega)$ such that the following quantity is finite:

$$J(u)=\sup\left\{\int_\Omega u(x)\operatorname{div}(\xi(x))\,dx\ /\ \xi\in C_0^\infty(\Omega,\mathbb{R}^2),\ \|\xi\|_{L^\infty(\Omega,\mathbb{R}^N)}\leq 1\right\}. \tag{4.1}$$

$BV(\Omega)$ endowed with the norm $\|u\|_{BV}=\|u\|_{L^1}+J(u)$ is a Banach space. If $u\in BV(\Omega)$, the distributional derivative Du is a bounded Radon measure, and (4.1) corresponds to the total variation, i.e., $J(u)=\int_\Omega|Du|$.

For $\Omega\subset\mathbb{R}^2$, if $1\leq p\leq 2$, we have $BV(\Omega)\subset L^p(\Omega)$. Moreover, for $1\leq p<2$, this embedding is compact. For further details on $BV(\Omega)$, we refer the reader to [1].

- Since $BV(\Omega)\subset L^2(\Omega)$, we can extend the functional J (which we still denote by J) over $L^2(\Omega)$:

$$J(u)=\begin{cases}\int_\Omega|Du| & \text{if } u\in BV(\Omega),\\ +\infty & \text{if } u\in L^2(\Omega)\backslash BV(\Omega).\end{cases} \tag{4.2}$$

We can then define the subdifferential ∂J of J [35]: $v \in \partial J(u)$ if and only if for all $w \in L^2(\Omega)$, we have $J(u+w) \geq J(u) + \langle v, w\rangle_{L^2(\Omega)}$, where $\langle ., .\rangle_{L^2(\Omega)}$ denotes the usual inner product in $L^2(\Omega)$.

- *Decomposability of $BV(\Omega)$.* If u in $BV(\Omega)$, then $Du = \nabla u\, dx + D_s u$, where $\nabla u \in L^1(\Omega)$ and $D_s u \perp dx$. ∇u is called the regular part of Du.
- *Weak * topology on $BV(\Omega)$.* If (u_n) is a bounded sequence in $BV(\Omega)$, then up to a subsequence, there exists $u \in BV(\Omega)$ such that $u_n \to u$ in $L^1(\Omega)$ strong, and $Du_n \to Du$ in the sense of measure; i.e., $\langle Du_n, \phi\rangle \to \langle Du, \phi\rangle$ for all ϕ in $(C_0^\infty(\Omega))^2$.
- *Approximation by smooth functions.* If u belongs to $BV(\Omega)$, then there exists a sequence u_n in $C^\infty(\Omega) \bigcap BV(\Omega)$ such that $u_n \to u$ in $L^1(\Omega)$ and $J(u_n) \to J(u)$ as $n \to +\infty$.
- In this paper, if a function f belongs to $L^\infty(\Omega)$, we denote by $\sup_\Omega f$ (resp., $\inf_\Omega f$) the sup ess of f (resp., the inf ess of f). We recall that $\operatorname{sup\,ess} f = \inf\{C \in \mathbb{R};\ f(x) \leq C \text{ a.e.}\}$ and $\operatorname{inf\,ess} f = \sup\{C \in \mathbb{R};\ f(x) \geq C \text{ a.e.}\}$.

4.2. The variational model. The application we have in mind is the denoising of nontextured SAR images. Inspired by the works of Rudin et al. [37, 36], we decide to choose $\phi(u) = J(u)$.

We thus propose the following restoration model (λ being a regularization parameter):

$$\inf_{u\in S(\Omega)} J(u) + \lambda \int_\Omega \left(\log u + \frac{f}{u}\right), \tag{4.3}$$

where $S(\Omega) = \{u \in BV(\Omega),\ u > 0\}$ and $f > 0$ in $L^\infty(\Omega)$ is the given data.

From now on, without loss of generality, we assume that $\lambda = 1$.

4.3. Existence of a minimizer. In this subsection, we show that problem (4.3) has at least one solution.

Theorem 4.1. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$; then problem* (4.3) *has at least one solution u in $BV(\Omega)$ satisfying*

$$0 < \inf_\Omega f \leq u \leq \sup_\Omega f \tag{4.4}$$

Proof. Let us denote $\inf f$ by α and $\sup f$ by β. Let us consider a minimizing sequence $(u_n) \in S(\Omega)$ for problem (4.3). Let us set

$$E(u) = J(u) + \int_\Omega \left(\log u + \frac{f}{u}\right). \tag{4.5}$$

We split the proof into two parts.

Part 1. We first show that we can assume without restriction that $\alpha \leq u_n \leq \beta$.

We remark that $x \mapsto \log x + \frac{f}{x}$ is decreasing if $x \in (0, f)$ and increasing if $x \in (f, +\infty)$. Therefore, if $M \geq f$, one always has

$$\left(\log(\min(x, M)) + \frac{f}{\min(x, M)}\right) \leq \log x + \frac{f}{x}. \tag{4.6}$$

Hence, if we let $M = \beta = \sup f$, we find that

$$\int_\Omega \left(\log \inf(u, \beta) + \frac{f}{\inf(u, \beta)}\right) \leq \int_\Omega \left(\log u + \frac{f}{u}\right). \tag{4.7}$$

Moreover, we have that (see Lemma 1 in section 4.3 of [27], for instance)

$$J(\inf(u,\beta)) \leq J(u).$$

We thus deduce that

$$E(\inf(u,\beta)) \leq E(u). \tag{4.8}$$

And in the same way we get that $E(\sup(u,\alpha)) \leq E(u)$, where $\alpha = \inf f$.

Part 2. From the first part of the proof, we know that we can assume that $\alpha \leq u_n \leq \beta$. This implies in particular that u_n is bounded in $L^1(\Omega)$.

By definition of (u_n), the sequence $E(u_n)$ is bounded; i.e., there exists a constant C such that $J(u_n) + \int_\Omega \left(\log u_n + \frac{f}{u_n}\right) \leq C$. Moreover, standard computations show that $\int_\Omega \left(\log u_n + \frac{f}{u_n}\right)$ reaches its minimum value $\int_\Omega (1 + \log f)$ when $u = f$, and thus we deduce that $J(u_n)$ is bounded.

Therefore we get that u_n is bounded in $BV(\Omega)$ and there exists u in $BV(\Omega)$ such that up to a subsequence, $u_n \to u$ in $BV(\Omega)$ weak * and $u_n \to u$ in $L^1(\Omega)$ strong. Necessarily, we have $0 \leq \alpha \leq u \leq \beta$, and thanks to the lower semicontinuity of the total variation and Fatou's lemma, we get that u is a solution of problem (4.3). □

4.4. Uniqueness and comparison principle. In this subsection, we address the problem of the uniqueness of a solution of problem (4.3). The question remains open in general, but we prove two results: we give a sufficient condition ensuring uniqueness and we show that a comparison principle holds.

PROPOSITION 4.2. *Let $f > 0$ be in $L^\infty(\Omega)$; then problem* (4.3) *has at most one solution $\hat{u}$ such that* $0 < \hat{u} < 2f$.

Proof. Let us set

$$h(u) = \log u + \frac{f}{u}. \tag{4.9}$$

We have $h'(u) = \frac{1}{u} - \frac{f}{u^2} = \frac{u-f}{u^2}$ and $h''(u) = -\frac{1}{u^2} + 2\frac{f}{u^3} = \frac{2f-u}{u^3}$. We deduce that if $0 < u < 2f$, then h is strictly convex, implying the uniqueness of a minimizer. □

We now state a comparison principle.

PROPOSITION 4.3. *Let f_1 and f_2 be in $L^\infty(\Omega)$ with $\inf_\Omega f_1 > 0$ and $\inf_\Omega f_2 > 0$. Let us assume that $f_1 < f_2$. We denote by u_1 (resp., u_2) a solution of* (4.3) *for $f = f_1$ (resp., $f = f_2$). Then we have $u_1 \leq u_2$.*

Proof. We use here the following classical notation: $u \vee v = \sup(u,v)$, and $u \wedge v = \inf(u,v)$.

From Theorem 4.1, we know that u_1 and u_2 do exist. We have, since u_i is a minimizer with data f_i,

$$J(u_1 \wedge u_2) + \int_\Omega \left(\log(u_1 \wedge u_2) + \frac{f_1}{u_1 \wedge u_2}\right) \geq J(u_1) + \int_\Omega \left(\log u_1 + \frac{f_1}{u_1}\right) \tag{4.10}$$

and

$$J(u_1 \vee u_2) + \int_\Omega \left(\log(u_1 \vee u_2) + \frac{f_2}{u_1 \vee u_2}\right) \geq J(u_2) + \int_\Omega \left(\log u_2 + \frac{f_2}{u_2}\right). \tag{4.11}$$

Adding these two inequalities and using the fact that $J(u_1 \wedge u_2) + J(u_1 \vee u_2) \leq J(u_1) + J(u_2)$ [12, 23], we get

$$\int_\Omega \left(\log(u_1 \wedge u_2) + \frac{f_1}{u_1 \wedge u_2} - \log u_1 - \frac{f_1}{u_1} + \log(u_1 \vee u_2) + \frac{f_2}{u_1 \vee u_2} - \log u_2 - \frac{f_2}{u_2}\right) \geq 0. \tag{4.12}$$

Writing $\Omega = \{u_1 > u_2\} \cup \{u_1 \leq u_2\}$, we easily deduce that

$$\int_{\{u_1>u_2\}} (f_1 - f_2)\frac{u_1 - u_2}{u_1 u_2} \geq 0. \tag{4.13}$$

Since $f_1 < f_2$, we thus deduce that $\{u_1 > u_2\}$ has a zero Lebesgue measure; i.e., $u_1 \leq u_2$ a.e. in Ω. □

4.5. Euler–Lagrange equation associated to problem (4.3). Let us now write an "Euler–Lagrange" equation for any solution of problem (4.3), the difficulty being that the ambient space is $BV(\Omega)$.

PROPOSITION 4.4. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$. If u in $BV(\Omega)$ is a solution of problem* (4.3)*, then we have*

$$-h'(u) \in \partial J(u). \tag{4.14}$$

Proof. This is a consequence of the maximum principle (4.4) of Theorem 4.1. Indeed, h can be replaced below the value $\inf_\Omega f$ by its C^1-quadratic extension, and this change does not alter the set of minimizers. The new functional has a Lipschitz derivative, and then standard results can be used to get (4.14). □

To give more insight into (4.14), we state the following result (see Proposition 1.10 in [2] for further details).

PROPOSITION 4.5. *Let (u, v) be in $L^2(\Omega)$ with u in $BV(\Omega)$. The following assertions are equivalent:*

(i) $v \in \partial J(u)$.

(ii) *Denoting by $X(\Omega)_2 = \{z \in L^\infty(\Omega, \mathbb{R}^2) : \mathrm{div}(z) \in L^2(\Omega)\}$, we have*

$$\int_\Omega vu\, dx = J(u) \tag{4.15}$$

and

$$\begin{array}{l}\exists z \in X(\Omega)_2,\ \|z\|_\infty \leq 1,\ z.N = 0 \textit{ on } \partial\Omega \\ \textit{such that } v = -\mathrm{div}(z) \textit{ in } \mathcal{D}'(\Omega).\end{array} \tag{4.16}$$

(iii) (4.16) *holds, and*

$$\int_\Omega (z, Du) = \int_\Omega |Du|. \tag{4.17}$$

From this proposition, we see that (4.14) means that $-h'(u) = \mathrm{div}\, z$, with z satisfying (4.16) and (4.17). This is a rigorous way to write $-\mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right) + h'(u) = 0$.

5. Numerical results.

We present in this section some numerical examples illustrating the capability of our model. We also compare it with some other existing models.

5.1. Algorithm. To numerically compute a solution to problem (4.3), we use the equation $-\mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right) + h'(u) = 0$ and, as it is classically done in image analysis, we embed it into a dynamical equation which we drive to a steady state:

$$\frac{\partial u}{\partial t} = \mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right) + \lambda\frac{f - u}{u^2} \tag{5.1}$$

with initial data $u(x,0) = \frac{1}{|\Omega|}\int_\Omega f$. We denote this model as the AA model. We use the following explicit scheme, with finite differences (we have checked numerically that for $\delta t > 0$ small enough, the sequence (u_n) satisfies a maximum principle):

$$\frac{u_{n+1} - u_n}{\delta t} = \left(\operatorname{div}\left(\frac{\nabla u_n}{\sqrt{|\nabla u_n|^2 + \beta^2}} \right) - \lambda h'(u_n) \right) \tag{5.2}$$

with β a small fixed parameter.

5.2. Other models. We have compared our results with some other classical variational denoising models.

*Additive model (*log*).* A natural way to turn a multiplicative model into an additive one is to use the logarithm transform (see [5, 22], for instance). Nevertheless, as can be seen on the numerical results, such a straightforward method does not produce satisfactory results. In the numerical results presented in this paper, we refer to this model as the log model. We first take the logarithm of the original image f. We then denoise $\log(f)$ by using the Rudin–Osher–Fatemi (ROF) model [37, 13], with the functional $\inf_y \left(J(y) + \frac{1}{2\lambda}\|x - z\|^2_{L^2} \right)$. We finally take the exponential to obtain the restored image. As can be seen in Figures 2 and 3, there is no maximum principle for this algorithm. In particular, the mean of the restored image is much smaller than that of the original image. In fact, in such an approach, the assumptions are not consistent with the modeling, as explained hereafter.

The original considered model is the following: $f = uv$, under the assumptions that u and v are independent, and $E(v) = 1$ (i.e., v is of mean 1). Hence $E(f) = E(u)$.

Now, if we take the logarithm, denoting by $x = \log(f)$, $y = \log(u)$, and $z = \log(v)$, we get the additive model $x = y + z$. To recover y from x, the classical assumption is $E(z) = 0$: this is the basic assumption in all the classical additive image restoration methods [11, 4] (total variation minimization, nonlinear diffusion, wavelet shrinkage, nonlocal means, heat equation, etc.).

But, from Jensen's inequality, we have $\exp(E(z)) \leq E(\exp(z))$, i.e., $1 \leq E(v)$. As soon as there is some noise, we are no longer in the case of equality in Jensen's inequality, which implies $E(v) > 1$. As a consequence, $E(u) < E(f)$ (in the numerical examples presented in Figures 2 and 3, we obtain $E(u) \approx E(f)/2$).

As a conclusion, if one wants to use the logarithm to get an additive model, then one cannot directly apply a standard additive noise removal algorithm. One needs to be more careful.

RLO model. The second model we use is a multiplicative version of the ROF model: it is a constrained minimization problem proposed by Rudin, Lions, and Osher in [36, 34], and we will call it the RLO model. In this approach, the model considered is $f = uv$, under the constraints that $\int_\Omega v = 1$ (mean 1) and $\int_\Omega (v-1)^2 = \sigma^2$ (given variance). The goal is then to minimize $\int_\Omega |Du|$ under the two previous constraints. The gradient projection method leads to

$$\frac{\partial u}{\partial t} = \operatorname{div}\left(\frac{\nabla u}{|\nabla u|} \right) - \lambda \frac{f^2}{u^3} - \mu \frac{f}{u^2}. \tag{5.3}$$

The two Lagrange multipliers λ and μ are dynamically updated to satisfy the constraints (as explained in [36]). With this algorithm, there is no regularization parameter to tune: the parameter to tune here is the number of iterations (since the considered flow is not associated to any functional). In practice, it appears that the Lagrange multipliers λ and μ are almost always of opposite signs.

Notice that the model proposed in this paper (AA) is specifically devoted to the denoising of images corrupted by gamma noise. The RLO model does not make such an assumption on the noise and therefore cannot be expected to perform as well as the AA model for speckle removal. Notice also that in the case of small Gaussian multplicative noise, both the RLO and AA models give very good results, as can be seen in Figure 5.

5.3. Deblurring. It is possible to modify our model to incorporate a linear blurring operator K. With u being the image to recover, we assume that the observed image f is obtained as $f = (Ku).v$. The functional to minimize in this case becomes

$$\inf_u \left(J(u) + \lambda \int_\Omega \left(\frac{f}{Ku} + \log(Ku) \right) \right). \tag{5.4}$$

The associated Euler–Lagrange equation is (denoting by K^T the transpose of K)

$$0 \in \partial J(u) + \lambda K^T \left(\frac{-f}{(Ku)^2} + \frac{1}{Ku} \right). \tag{5.5}$$

Numerically, we use a steepest gradient descent approach by solving

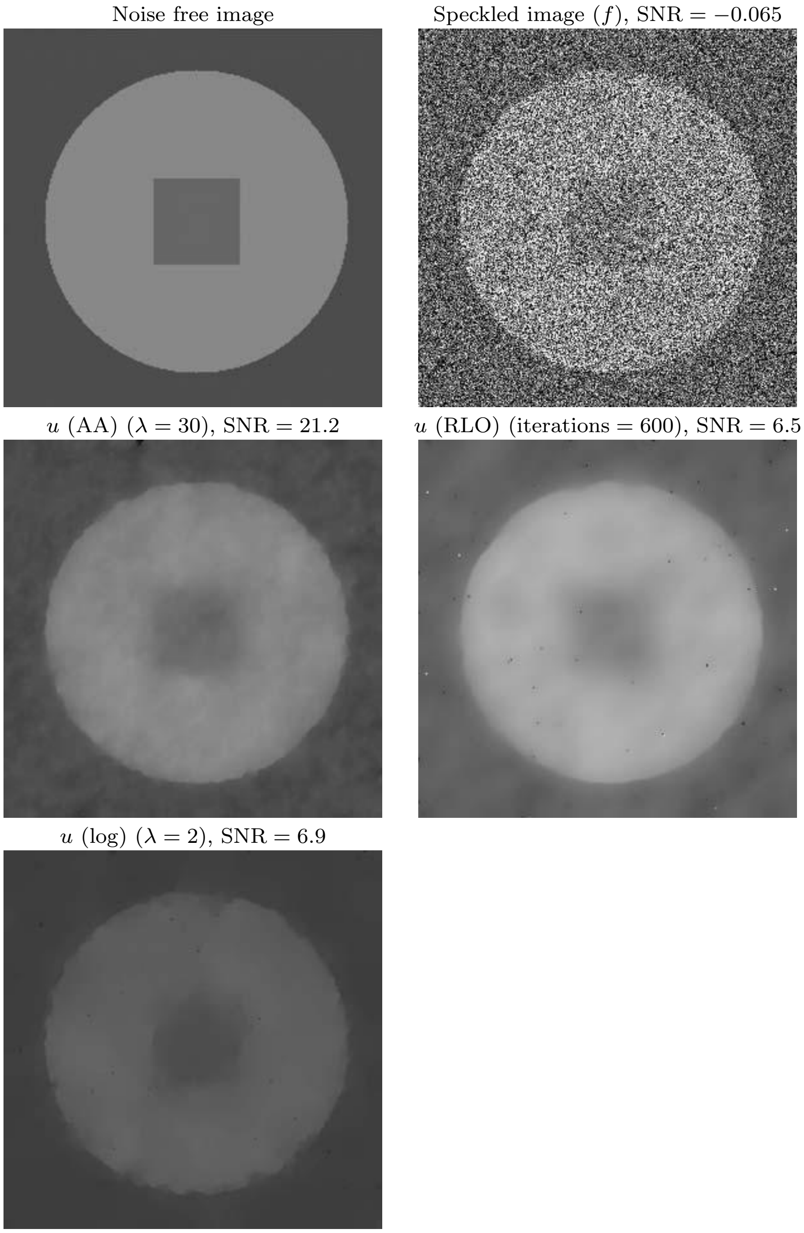
$$\frac{\partial u}{\partial t} = \operatorname{div}\left(\frac{\nabla u}{|\nabla u|} \right) + \lambda K^T \left(\frac{f - Ku}{(Ku)^2} \right). \tag{5.6}$$

5.4. Results. In Figure 2, we show a first example. The original synthetic image is corrupted by some multiplicative noise with gamma law of mean 1 (see (3.2)). We display the denoising results obtained by our approach (AA), as well as with the RLO model. Due to the very strong noise, the RLO model experiences some difficulties in bringing back in the range of the image some isolated points (white and black points on the denoised image) and at the same time keeping sharp edges: to remove these artifacts, one needs to regularize more, and therefore some part of the edges are lost. Moreover, the mean of the original image is not preserved (the mean of the restored image is quite larger than that of the original image): this is the reason why the signal-to-noise ratio (SNR) is not much improved, and also why the restored image with the RLO model looks lighter. We also display the results obtained with the log model: as explained before, this model gives bad results, due to the fact that the mean is not preserved (with the log model, the mean is much reduced). This is the reason why the restored image with the log model is much darker.

In Figure 3, we show how our model behaves with a complicated geometrical image. We also give the results with the RLO model and the log model (which have the same drawbacks as in Figure 2).

In Figure 4, we show the result we get on a SAR image provided by the CNES (French space agency). The reference image (also furnished by the CNES) has been obtained by amplitude summation.

In Figure 5, we show how our model behaves with multiplicative Gaussian noise. We have used the same parameters for the Gaussian noise as in [36], i.e., a standard deviation of 0.2 (and a mean equal to 1). The original image is displayed in Figure 3. In this case, we see that we get a very good restoration result. Notice that such a multiplicative Gaussian noise is much easier to handle than the speckle noise which was tackled in Figures 2–4. But, as far as we know, this is the type of multiplicative noise which was considered in all the variational approaches inspired by [36] (as used,



Noise free image

Speckled image (f), SNR $= -0.065$

u (AA) ($\lambda = 30$), SNR $= 21.2$

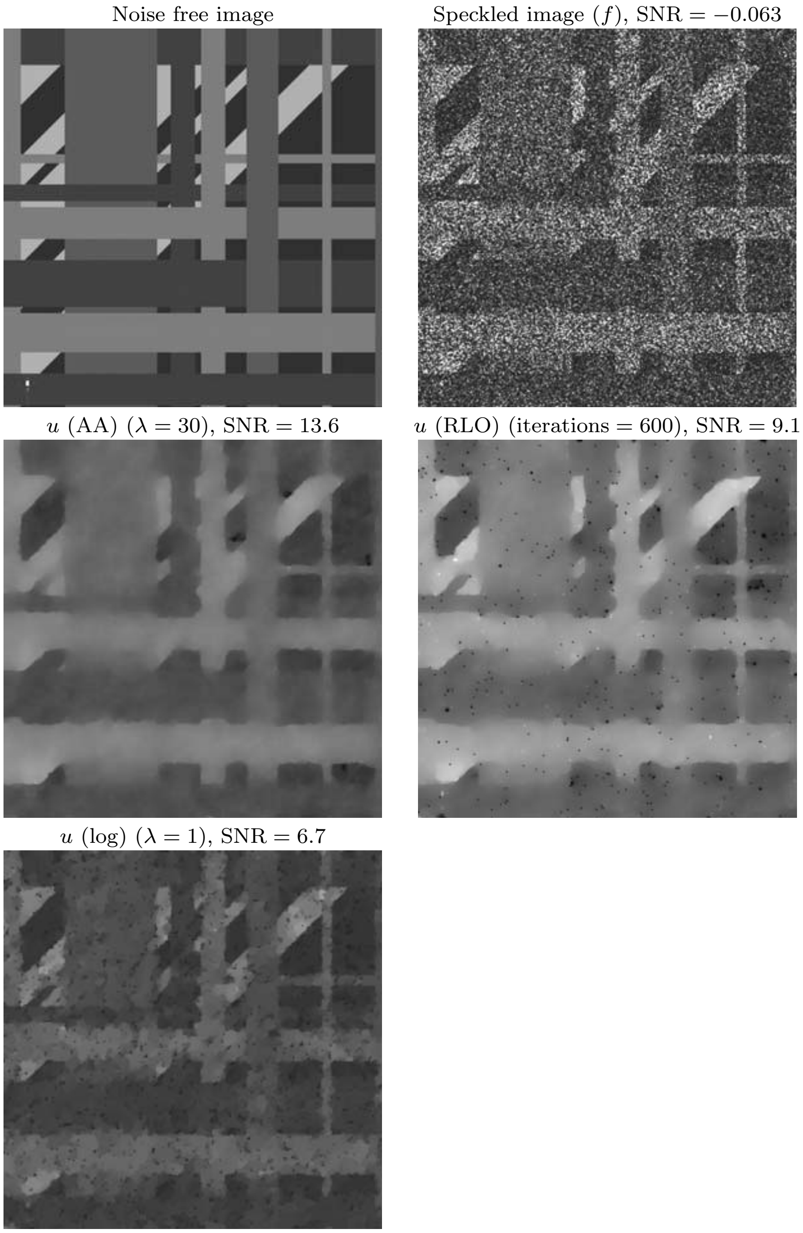
u (RLO) (iterations $= 600$), SNR $= 6.5$

u (log) ($\lambda = 2$), SNR $= 6.9$

FIG. 2. *Denoising of a synthetic image with gamma noise. f has been corrupted by some multiplicative noise with gamma law of mean* 1. *u is the denoised image.*

for instance, in [33, 28, 29, 38]). We also show the results with the RLO model and the log model. Notice that in this case all the models perform very well, even the log model: indeed, since the noise is small, Jensen's inequality is almost an equality.

In Figure 6, we finally show a deblurring example with our model (5.4). The original image (displayed in Figure 3) has been convolved with a Gaussian kernel of standard deviation $\sqrt{2}$ and then multiplied by a Gaussian noise of standard deviation 0.2 and mean 1 (we use the same parameters as in [36]). Even though the restored



Noise free image

Speckled image (f), SNR $= -0.063$

u (AA) ($\lambda = 30$), SNR $= 13.6$

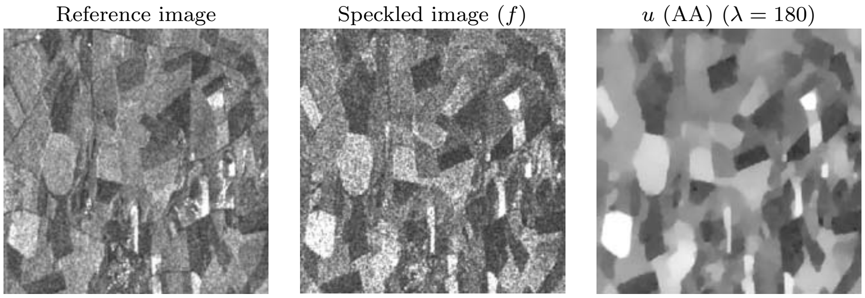
u (RLO) (iterations $= 600$), SNR $= 9.1$

u (log) ($\lambda = 1$), SNR $= 6.7$

FIG. 3. *Denoising of a synthetic image with gamma noise. f has been corrupted by some multiplicative noise with gamma law of mean* 1.

image is not as good as in the denoising case presented in Figure 5, we see that our model works well for deblurring.

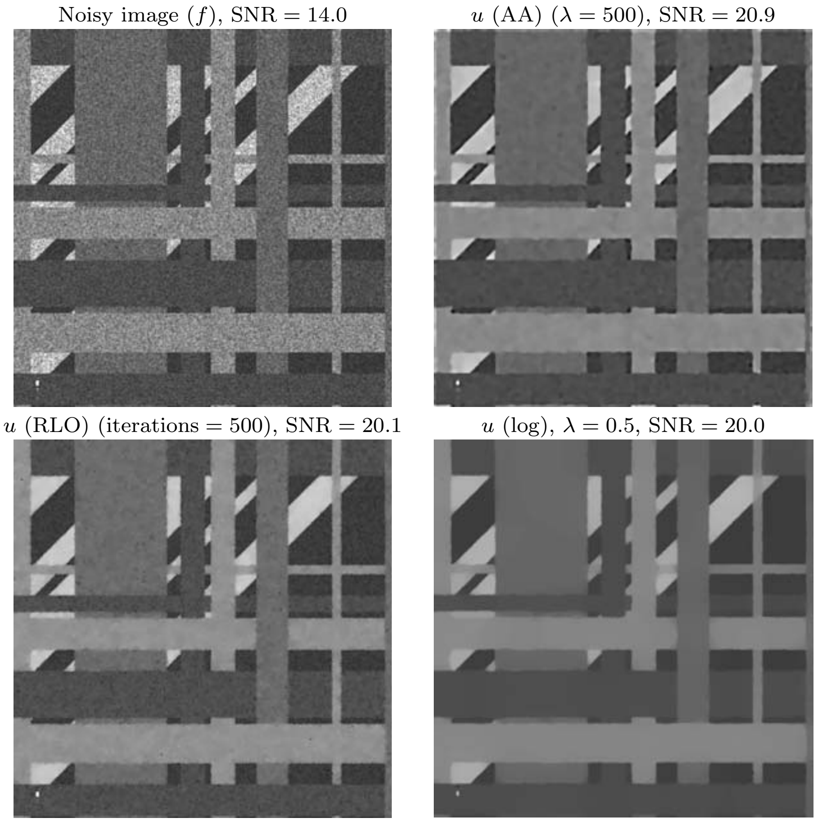
6. Evolution equation. In this section we study the evolution equation associated to (4.14). The motivation is that when searching for a numerical solution of (4.14) it is, in general, easier to compute a solution of the associated evolution equation (by using, for example, explicit or semi-implicit schemes) and then studying the asymptotic behavior of the process to get a solution of the stationary equation.


Reference image

Speckled image (f)

u (AA) ($\lambda = 180$)

FIG. 4. *Denoising of a SAR image provided by the CNES.*


Noisy image (f), SNR = 14.0

u (AA) ($\lambda = 500$), SNR = 20.9

u (RLO) (iterations = 500), SNR = 20.1

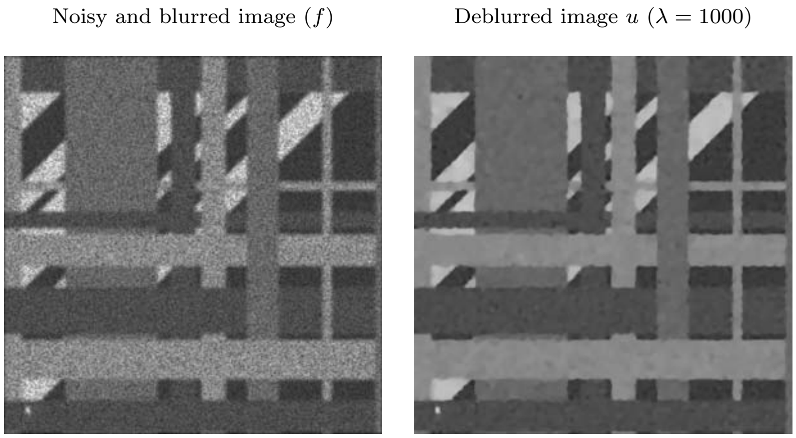
u (log), $\lambda = 0.5$, SNR = 20.0

FIG. 5. *Denoising of a synthetic image degraded by multiplicative Gaussian noise with* $\sigma = 0.2$. *The original noise free image is shown in Figure* 3.

We first consider a semidiscrete version of the problem: the space Ω is still included in R^2, but we discretize the time variable. We consider the case of a regular time discretization, (t_n), with t_0 given and $t_{n+1} - t_n = \delta t$ in $\mathbb{R}_+^*$ (in this section, δt is fixed). We define $u_n = u(.,t_n)$, and we consider the following implicit scheme:

$$0 \in \frac{u_{n+1} - u_n}{\delta t} + \partial J(u_{n+1}) + h'(u_{n+1}), \tag{6.1}$$

where J is the extended total variation as defined in (4.2). We first need to check that


Noisy and blurred image (f)

Deblurred image u ($\lambda = 1000$)

FIG. 6. *Deblurring of the synthetic image of Figure* 3 *(the original image, which is shown in Figure* 3*, was first convolved with a Gaussian kernel of standard deviation $\sigma = \sqrt{2}$ and then multiplied by some Gaussian noise of mean* 1 *and standard deviation $\sigma = 0.2$).*

(6.1) indeed defines a sequence (u_n). To this end, we intend to study the following functional:

$$\inf_{u \in BV(\Omega),\, u>0} F(u, u_n) \tag{6.2}$$

with

$$F(u, u_n) = \int_\Omega \frac{u^2}{2}\, dx - \int_\Omega u_n u \, dx + \delta t \left(J(u) + \int_\Omega h(u)\, dx \right). \tag{6.3}$$

We want to define u_{n+1} as

$$u_{n+1} = \underset{\{u \in BV(\Omega),\, u>0\}}{\operatorname{argmin}} F(u, u_n). \tag{6.4}$$

6.1. Existence and uniqueness of the sequence (u_n). We first need to check that the sequence (u_n) is indeed well defined.

PROPOSITION 6.1. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$. Let (u_n) be in $BV(\Omega)$ such that $\inf_\Omega f \leq u_n \leq \sup_\Omega f$. If $\delta t < 27(\inf_\Omega f)^2$, then there exists a unique u_{n+1} in $BV(\Omega)$ satisfying* (6.4)*. Moreover, we have*

$$\inf\left(\inf_\Omega f \,,\, \inf_\Omega u_0\right) \leq u_n \leq \sup\left(\sup_\Omega f \,,\, \sup_\Omega u_0\right). \tag{6.5}$$

Proof. We split the proof into two parts.

Part 1. We first show the existence and uniqueness of u^{n+1}. We consider: $g(u) = \delta t h(u) + u^2/2 - u_n u$. We have $g''(u) = 1 + \delta t \frac{f-u}{u^2} = \frac{u^3 - \delta t u + 2\delta t f}{u^3}$. A simple computation shows that if $\delta t < 27(\inf_\Omega f)^2$, then $g''(u) > 0$ for all $u > 0$, i.e., g strictly convex on $\mathbb{R}_+^*$. It is then standard to deduce the existence and uniqueness of u^{n+1}.

Part 2. As in the proof of Theorem 4.1, we have

$$\int_\Omega \left(\log \inf(u, \beta) + \frac{f}{\inf(u, \beta)} \right) \leq \int_\Omega \left(\log u + \frac{f}{u} \right) \tag{6.6}$$

and $J(\inf(u,\beta)) \leq J(u)$.

We remark that $x \mapsto x^2/2 - xu_n$ is decreasing if $x \in (0, u_n)$ and increasing if $x \in (u_n, +\infty)$. Therefore, proceeding as in the proof of Theorem 4.1, we get

$$\int_\Omega \frac{\left(\inf\left(u\,,\,\sup u_n\right)\right)^2}{2} - u_n \inf\left(u\,,\,\sup u_n\right) \leq \int_\Omega \frac{u^2}{2} - uu_n. \tag{6.7}$$

Thus the truncation procedure decreases the energy, and we deduce the right-hand side inequality in (6.5). We get the other one in the same way. □

We can thus derive the following theorem.

Theorem 6.2. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$, and u_0 in $L^\infty(\Omega) \bigcap BV(\Omega)$ with $\inf_\Omega u_0 > 0$ be given. If $\delta t < 27(\inf_\Omega f)^2$, then there exists a unique sequence (u_n) in $BV(\Omega)$ satisfying* (6.4). *Moreover, the following estimates hold:*

$$\inf\left(\inf_\Omega f\,,\, \inf_\Omega u_0\right) = \alpha \leq u_n \leq \beta = \sup\left(\sup_\Omega f\,,\, \sup_\Omega u_0\right) \tag{6.8}$$

and

$$J(u_n) \leq J(u_0) + \int_\Omega h(u_0)\,dx - \int_\Omega (1 + \log f). \tag{6.9}$$

Proof. This theorem is just a consequence (by induction) of Proposition 6.1, except for estimate (6.9) which we prove now.

From (6.4), we have $F(u_{n+1}, u_n) \leq F(u_n, u_n)$, which means

$$\begin{aligned} &\delta t \left(J(u_{n+1}) - J(u_n) + \int_\Omega h(u_{n+1}) - \int_\Omega h(u_n) \right) \\ &+ \frac{1}{2} \int_\Omega (u_{n+1} - u_n)^2 \leq 0. \end{aligned} \tag{6.10}$$

This implies

$$J(u_{n+1}) - J(u_n) + \int_\Omega h(u_{n+1}) - \int_\Omega h(u_n) \leq 0. \tag{6.11}$$

By summation, we obtain

$$J(u_{n+1}) \leq -\int_\Omega h(u_{n+1}) + \int_\Omega h(u_0) + J(u_0). \tag{6.12}$$

Standard computations show that $\int_\Omega h(u_{n+1}) \geq \int_\Omega (1 + \log f)\,dx$, from which we deduce (6.9). □

6.2. Euler–Lagrange equation. We have the following "Euler–Lagrange" equation.

Proposition 6.3. *The sequence (u_n) satisfying* (6.4) *is such that*

$$0 \in \frac{u_{n+1} - u_n}{\delta t} + \left(\partial J(u_{n+1}) + h'(u_{n+1}) \right). \tag{6.13}$$

Proof. The proof is similar to that of Proposition 4.4. □

6.3. Convergence of the sequence u_n. The following convergence result holds.

PROPOSITION 6.4. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$, and u_0 in $L^\infty(\Omega)\bigcap BV(\Omega)$ with $\inf_\Omega u_0 > 0$ be fixed. Let $\delta t < 27(\inf_\Omega f)^2$. The sequence (u_n) defined by (6.1) is such that there exists u in $BV(\Omega)$ with $u_n \rightharpoonup u$ (up to a subsequence) for the $BV(\Omega)$ weak * topology, and u is solution of*

$$0 \in \partial J(u) + h'(u) \tag{6.14}$$

in the distributional sense.

Proof. As in the proof of Theorem 6.2, we get the same kind of equation as (6.10):

$$\frac{1}{2}\int_\Omega (u_{n+1}-u_n)^2 \leq \delta t \left(J(u_n) - J(u_{n+1}) + \int_\Omega h(u_n) - \int_\Omega h(u_{n+1}) \right). \tag{6.15}$$

By summation, we obtain

$$\begin{aligned}\frac{1}{2}\sum_{n=0}^{N-1}\int_\Omega (u_{n+1}-u_n)^2 &\leq \delta t \left(J(u_0) - J(u_N) + \int_\Omega h(u_0) - \int_\Omega h(u_N) \right)\\ &\leq \delta t \left(J(u_0) + \int_\Omega h(u_0) - \int_\Omega h(f) \right) < +\infty\end{aligned}$$

(since $\int_\Omega h(u_N \geq \int_\Omega h(f))$. In particular, this implies that $u_{n+1} - u_n \to 0$ in $L^2(\Omega)$ strong.

From estimate (6.9), we know that there exists u in $BV(\Omega)$ such that up to a subsequence $u_n \rightharpoonup u$ for the $BV(\Omega)$ weak * topology. Moreover, $u_n \to u$ in $L^1(\Omega)$ strong. Let $v \in L^2(\Omega)$. From (6.13), we have

$$J(v) \geq J(u_{n+1}) + \left\langle v - u_{n+1}, -\frac{u_{n+1}-u_n}{\delta t} - h'(u_{n+1}) \right\rangle_{L^2(\Omega)}. \tag{6.16}$$

Using estimate (6.8) and the fact that $u_n \to u$ in $L^1(\Omega)$ strong, we deduce from Lebesgue's dominated convergence theorem that (up to a subsequence) $u_n \to u$ in $L^2(\Omega)$ strong. Moreover, since $u_{n+1} - u_n \to 0$ in $L^2(\Omega)$ strong, and thanks to the lower semicontinuity of the total variation, we get $J(v) \geq J(u) + \langle v-u, -h'(u)\rangle_{L^2(\Omega)}$. Hence (6.14) holds. □

6.4. Continuous setting. Let us consider the evolution equation

$$\frac{\partial u}{\partial t} \in -\partial J(u) - h'(u) \tag{6.17}$$

with the initial condition $u(0) = u_0$ and with $h(u) = \frac{f}{u} + \log u$, i.e., $h'(u) = \frac{u-f}{u^2}$. $J(u)$ still denotes the extended total variation of u with respect to the space variable x.

To show the existence and uniqueness of a solution for (6.17), we could apply the theory of maximal monotone operator [9, 8, 2]. This theory works provided that h' is Lipschitz. One need only replace h by its C^1-quadratic extension below $\inf_\Omega$. This would yield a solution in $L^2(\Omega)$. Here, we derive sharper bounds with the next result, whose proof is given in Appendix A.

THEOREM 6.5. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$, and u_0 in $L^\infty(\Omega) \bigcap BV(\Omega)$ with $\inf_\Omega u_0 > 0$. Then problem (6.17) has exactly one solution u in $L^\infty_w((0,T);BV(\Omega))$ $\bigcap W^{1,2}((0,T);L^2(\Omega))$.*

Remark. u belongs to $L^\infty_w((0,T);BV(\Omega))$ means that u belongs to $L^\infty((0,T)\times\Omega)$ and Du belongs to $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$. $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ is the space of equivalent classes of weak * measurable mappings μ that are essentially bounded; i.e., $\sup\operatorname{ess}_{x\in\Omega}\|\mu(x)\| < +\infty$ (we say that μ is weak * measurable if $\langle\mu(x),f\rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega;\mathbb{R}^2)}$ is measurable with respect to x for every f in $C_0(\Omega;\mathbb{R}^2)$; see Lemma A.5 and [7] for further details).

Appendix A. Evolution equation: Continuous setting.

To show that problem (6.17) has a solution, we start from the semidiscrete problem we have studied in the previous section. We therefore consider a sequence (u_n) satisfying (6.4). From Proposition 6.3, we know that (u_n) satisfies

$$0 \in \frac{u_{n+1}-u_n}{\delta t} + \Big(\partial J(u_{n+1}) + h'(u_{n+1})\Big) \tag{A.1}$$

and u_{n+1} satisfies Neumann boundary conditions $\frac{\partial u_{n+1}}{\partial N} = 0$ on the boundary of Ω. From Theorem 6.2, we know that the sequence (u_n) exists and is unique provided that $\delta t < 27(\inf_\Omega f)^2$.

A.1. Definitions of interpolate functions.

We classically introduce two functions defined on $\Omega\times\mathbb{R}^+$. We assume that $t_0 = 0$ and $t_n = n\delta t$. Then

$$\tilde{u}_{\delta t}(t,x) = u_{[t/\delta t]+1}(x) = u_{n+1}(x) \text{ if } t_n < t \leq t_{n+1}, \tag{A.2}$$

where $[t/\delta t]$ is the integer part of $t/\delta t$. $\tilde{u}_{\delta t}(.,x)$ is thus piecewise constant. We also introduce

$$\hat{u}_{\delta t}(t,x) = (t-t_n)\frac{u_{n+1}(x)-u_n(x)}{\delta t} + u_n(x) \tag{A.3}$$

with $n = [t/\delta t]$. $\hat{u}_{\delta t}(.,x)$ is piecewise affine and continuous, and we have

$$\frac{\partial \hat{u}_{\delta t}}{\partial t}(t,x) = \frac{u_{n+1}(x)-u_n(x)}{\delta t}, \quad t_n < t < t_{n+1}. \tag{A.4}$$

With this notation, we can rewrite (A.1) as

$$\frac{\tilde{u}_{\delta t}(t,x) - \tilde{u}_{\delta t}(t-\delta t,x)}{\delta t} \in -\partial J(\tilde{u}_{\delta t}(t,x)) - h'(\tilde{u}_{\delta t}(t,x)), \tag{A.5}$$

i.e.,

$$\frac{\partial \hat{u}_{\delta t}}{\partial t}(t,x) \in -\partial J(\tilde{u}_{\delta t}(t,x)) - h'(\tilde{u}_{\delta t}(t,x)). \tag{A.6}$$

A.2. A priori estimates.

We first need to show some a priori estimates.

Proposition A.1. *Let $T>0$ be fixed, f in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$, and u_0 in $L^\infty(\Omega)\bigcap BV(\Omega)$ with $\inf_\Omega u_0 > 0$. Then if $0\leq t\leq T$,*

$$\inf\Big(\inf_\Omega f\,,\,\inf_\Omega u_0\Big) = \alpha \leq \tilde{u}_{\delta t}, \hat{u}_{\delta t} \leq \beta = \sup\Big(\sup_\Omega f\,,\,\sup_\Omega u_0\Big) \tag{A.7}$$

and

$$\sup_{t\in(0,T)}\{J(\tilde{u}_{\delta t}), J(\hat{u}_{\delta t})\} \leq J(u_0) + \int_\Omega h(u_0) - \int_\Omega h(f). \tag{A.8}$$

Proof. (A.7) for $\tilde{u}_{\delta t}$ comes from (6.8) in Theorem 6.2, and (A.8) comes from (6.9). We then get the estimates for $\hat{u}_{\delta t}$ from (A.3). □

PROPOSITION A.2. *Let $T > 0$ be fixed. There exists a constant $C > 0$, which does not depend on δt, such that*

$$\int_0^T \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} \leq C. \tag{A.9}$$

Proof. Let us denote $[t/\delta t]$ by N. We have

$$\int_{t_n}^{t_{n+1}} \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} = \delta t \int_\Omega \left| \frac{u_{n+1}(x) - u_n(x)}{\delta t} \right|^2 dx. \tag{A.10}$$

By using (6.15), we get

$$\int_{t_n}^{t_{n+1}} \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} \leq 2 \left(J(u_n) - J(u_{n+1}) + \int_\Omega h(u_n) - \int_\Omega h(u_{n+1}) \right). \tag{A.11}$$

Hence,

$$\begin{aligned} \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} &\leq 2 \left(J(u_0) - J(u_N) + \int_\Omega h(u_0) - \int_\Omega h(u_N) \right) \\ &\leq 2 \left(J(u_0) + \int_\Omega h(u_0) - \int_\Omega h(f) \right). \end{aligned}$$

We thus deduce that

$$\int_0^T \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} dt \leq 2T \left(J(u_0) + \int_\Omega h(u_0) - \int_\Omega h(f) \right) + \int_{t_N}^T \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} dt. \tag{A.12}$$

But, by using (6.15), we have

$$\begin{aligned} \int_{t_N}^T \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} dt &\leq 2 \frac{T - t_n}{\delta t} \left(J(u_N) - J(u_{N+1}) + \int_\Omega h(u_N) - \int_\Omega h(u_{N+1}) \right) \\ &\leq 2 \left(J(u_0) - J(u_{N+1}) + \int_\Omega h(u_N) - \int_\Omega h(u_{N+1}) \right). \end{aligned}$$

We then get from (6.9) and (6.8) that there exists $B > 0$ which does not depend on N and δt such that $\int_{t_N}^T \left\| \frac{\partial \hat{u}_{\delta t}}{\partial t} \right\|^2_{L^2(\Omega)} dt \leq B$. We then conclude the proof thanks to (A.12). □

COROLLARY A.3. *Let $T > 0$ be fixed. Then*

$$\lim_{\delta t \to 0} \int_0^T \| \hat{u}_{\delta t} - \tilde{u}_{\delta t} \|^2_{L^2(\Omega)} dt = 0. \tag{A.13}$$

Proof. Let us denote $[t/\delta t]$ by N. We have

$$\int_0^T \| \hat{u}_{\delta t} - \tilde{u}_{\delta t} \|^2_{L^2(\Omega)} dt = \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \| \hat{u}_{\delta t} - \tilde{u}_{\delta t} \|^2_{L^2(\Omega)} dt + \int_{t_N}^T \| \hat{u}_{\delta t} - \tilde{u}_{\delta t} \|^2_{L^2(\Omega)} dt. \tag{A.14}$$

But

$$\text{(A.15)}\quad \sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}}\|\hat{u}_{\delta t}-\tilde{u}_{\delta t}\|^2_{L^2(\Omega)}\,dt=\sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}}\|(t-t_n-\delta t)(u_{n+1}-u_n)\|^2_{L^2(\Omega)}\,dt.$$

We then deduce from (A.4) that

$$\begin{aligned}\sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}}\|\hat{u}_{\delta t}-\tilde{u}_{\delta t}\|^2_{L^2(\Omega)}\,dt &\le \sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}}\left\|\delta t\frac{\partial\hat{u}_{\delta t}}{\partial t}\right\|_{L^2(\Omega)}dt\\ &\le \underbrace{(\delta t)^2\int_0^T\left\|\frac{\partial\hat{u}_{\delta t}}{\partial t}\right\|^2_{L^2(\Omega)}dt\,.}_{\to 0\text{ as }\delta t\to 0}\end{aligned}$$

And

$$\text{(A.16)}\qquad \int_{t_N}^{T}\|\hat{u}_{\delta t}-\tilde{u}_{\delta t}\|^2_{L^2(\Omega)}\,dt\le\underbrace{(\delta t)^3\left\|\frac{u_{N+1}-u_N}{\delta t}\right\|^2_{L^2(\Omega)}dt\,.}_{\to 0\text{ as }\delta t\to 0}\qquad\square$$

We summarize the a priori estimates we have proved in the following corollary.

COROLLARY A.4. *Let $T>0$ be fixed. There exists a constant $C>0$ such that*

$$\text{(A.17)}\qquad \sup\left\{\sup_{t\in(0,T)}\|\tilde{u}_{\delta t}\|_{L^\infty(\Omega)}\,,\ \sup_{t\in(0,T)}\|\hat{u}_{\delta t}\|_{L^\infty(\Omega)}\right\}\le C,$$

$$\text{(A.18)}\qquad \sup\left\{\sup_{t\in(0,T)}J(\tilde{u}_{\delta t})\,,\ \sup_{t\in(0,T)}J(\hat{u}_{\delta t})\right\}\le C,$$

$$\text{(A.19)}\qquad \int_0^T\left\|\frac{\partial\hat{u}_{\delta t}}{\partial t}\right\|^2_{L^2(\Omega)}\le C,$$

$$\text{(A.20)}\qquad \lim_{\delta t\to 0}\int_0^T\|\hat{u}_{\delta t}-\tilde{u}_{\delta t}\|^2_{L^2(\Omega)}\,dt=0.$$

A.3. Existence of a solution. We can now prove Theorem 6.5.

Proof. The uniqueness of u will come from Proposition A.6. Here we just show the existence of u.

We first remark that, from inequalities (A.17) and (A.19), $\hat{u}_{\delta t}$ is uniformly bounded in $W^{1,2}((0,T);L^2(\Omega))$. Thus, up to a subsequence, there exists u in $W^{1,2}((0,T);L^2(\Omega))$ such that $\hat{u}_{\delta t}\rightharpoonup u$ in $W^{1,2}((0,T);L^2(\Omega))$ weak. Since $W^{1,2}((0,T);L^2(\Omega))$ is compactly embedded in $L^2((0,T);L^2(\Omega))$ (see [39, Theorem 2.1, Chapter 3]), $\hat{u}_{\delta t}\to u$ strongly in $L^2((0,T);L^2(\Omega))$.

Since (A.17) and (A.18) hold, we can apply Lemma A.5 (stated below) with $(\tilde{u}_{\delta t})$. Thus, up to a subsequence, there exists $\tilde{u}$ in $L^\infty_w((0,T);BV(\Omega))$ such that $\tilde{u}_{\delta t}\rightharpoonup\tilde{u}$ in $L^\infty(\Omega\times(0,T))$ weak * and $D_x\tilde{u}_{\delta t}\rightharpoonup D_x\tilde{u}$ in $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ weak *. From (A.20), we have that $\tilde{u}_{\delta t}\to u$ strongly in $L^2((0,T);L^2(\Omega))$, and we thus deduce that $\tilde{u}=u$.

The semidiscrete implicit scheme writes for a.e. $t \in (0,T)$

$$-\frac{\partial \hat{u}_{\delta t}}{\partial t} - h'(\tilde{u}_{\delta t}) \in \partial J(\tilde{u}_{\delta t}); \tag{A.21}$$

i.e., for all v in $BV(\Omega)$, $v > 0$, and a.e. $t \in (0,T)$,

$$J(v) \geq J(\tilde{u}_{\delta t}) + \left\langle v - \tilde{u}_{\delta t}, -\frac{\partial \hat{u}_{\delta t}}{\partial t} - h'(\tilde{u}_{\delta t}) \right\rangle_{L^2(\Omega)\times L^2(\Omega)}. \tag{A.22}$$

Let ϕ in $C_c^0(0,T)$ be a test function, $\phi \geq 0$. We multiply (A.22) by ϕ and integrate on $(0,T)$:

$$\int_0^T J(v)\phi(t)\,dt \geq \int_0^T J(\tilde{u}_{\delta t})\phi(t)\,dt + \int_0^T \int_\Omega (v - \tilde{u}_{\delta t}) \left(-\frac{\partial \hat{u}_{\delta t}}{\partial t} - h'(\tilde{u}_{\delta t}) \right) \phi(t)\,dt\,dx. \tag{A.23}$$

We want to let $\delta t \to 0$ in (A.23). By convexity, we have

$$\liminf \int_0^T J(\tilde{u}_{\delta t})\phi(t)\,dt \geq \int_0^T J(u)\phi(t)\,dt. \tag{A.24}$$

Now, since $\tilde{u}_{\delta t} \to u$ strongly in $L^2((0,T);L^2(\Omega))$, $\frac{\partial \hat{u}_{\delta t}}{\partial t} \rightharpoonup \frac{\partial u}{\partial t}$ in $L^2((0,T);L^2(\Omega))$ weak, and h' is bounded on the interval $[\alpha,\beta]$, the second term on the right-hand side of (A.23) tends to

$$\int_0^T \int_\Omega (v-u) \left(-\frac{\partial u}{\partial t} - h'(u) \right) \phi(t)\,dt\,dx.$$

We thus get

$$\int_0^T J(v)\phi(t)\,dt \geq \int_0^T J(u)\phi(t)\,dt + \int_0^T \int_\Omega (v-u) \left(-\frac{\partial u}{\partial t} - h'(u) \right) \phi(t)\,dt\,dx. \tag{A.25}$$

This inequality holds for all $\phi \geq 0$, and we deduce that for a.e. t in $(0,T)$

$$J(v) \geq J(u) + \int_\Omega (v-u) \left(-\frac{\partial u}{\partial t} - h'(u) \right) dx; \tag{A.26}$$

i.e., $-\frac{\partial u}{\partial t} \in \partial J(u) + h'(u)$. Hence we deduce that u is a solution of (6.17) in the distributional sense. □

In the above proof, we have used the following lemma.

LEMMA A.5. *Let (u_n) be a bounded sequence in $L^\infty_w(\Omega\times(0,T))$, such that $(D_x u_n)$ is a bounded sequence in $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$. Then, up to a subsequence, there exists u in $L^\infty_w((0,T);BV(\Omega))$ such that $u_n \rightharpoonup u$ in $L^\infty(\Omega\times(0,T))$ weak * and $D_x u_n \rightharpoonup D_x u$ in $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ weak *; i.e., for all ψ in $L^1((0,T);C_0(\Omega))$,*

$$\int_0^T \langle Du_n, \psi\rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega;\mathbb{R}^2)}\,dt \to \int_0^T \langle Du, \psi\rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega;\mathbb{R}^2)}\,dt, \tag{A.27}$$

where $\langle .,. \rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega)}$ denotes the duality product between bounded measures on Ω and $C_0(\Omega;\mathbb{R}^2)$ denotes the space of functions continuous on Ω and vanishing in $\partial\Omega$.

Proof. From the Riesz representation theorem [1, 20], there is an isometric isomorphism between $\mathcal{M}_b(\Omega)$ and the dual space of $C_0(\Omega)$. Moreover, since $C_0(\Omega)$ is separable, there is an isometric isomorphism between $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ and the dual space of $L^1((0,T);C_0(\Omega))$ (see [7] or [18, page 588]). Up to a subsequence, there exist u in $L^\infty(\Omega\times(0,T))$ and v in $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ such that $u_n \rightharpoonup u$ in $L^\infty(\Omega\times(0,T))$ weak *, and $D_x u_n \rightharpoonup v$ in $L^\infty_w((0,T);\mathcal{M}_b(\Omega))$ weak *. We therefore have for all ψ in $L^1((0,T);C_0(\Omega))$

$$\int_0^T \langle Du_n,\psi\rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega;\mathbb{R}^2)}\,dt \to \int_0^T \langle v,\psi\rangle_{\mathcal{M}_b(\Omega)\times C_0(\Omega;\mathbb{R}^2)}\,dt. \tag{A.28}$$

Moreover, we have $D_x u_n \to D_x u$ in $\mathcal{D}'(\Omega\times(0,T))$ and $D_x u_n \to v$ in $\mathcal{D}'(\Omega\times(0,T))$: this implies that $D_x u = v$. □

A.4. Uniqueness of the solution. A uniqueness result holds.

PROPOSITION A.6. *Let f be in $L^\infty(\Omega)$ with $\inf_\Omega f > 0$, and let u_0 be in $L^\infty(\Omega)\bigcap BV(\Omega)$ with $\inf_\Omega u_0 > 0$. Then problem* (6.17) *has at most one solution u such that $0<\alpha\le u\le\beta$.*

Proof. This is a standard result. It is based on the convexity of J, the fact that h' is Lipschitz on $[\inf_\Omega f, +\infty)$, and the Gronwall inequality. □

Acknowledgments. We would like to thank one of the anonymous referees for his very interesting comments about this paper (in particular for shortening the proof of Theorem 4.1 and suggesting the result of Proposition 4.3). We also would like to thank Professor Antonin Chambolle and Professor Vicent Caselles for fruitful discussions.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.

[2] F. Andreu-Vaillo, V. Caselles, and J. M. Mazon, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Progr. Math. 223, Birkhäuser Verlag, Basel, Switzerland, 2004.

[3] G. Anzellotti, *The Euler equation for functionals with linear growth*, Trans. Amer. Math. Soc., 290 (1985), pp. 483–501.

[4] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Appl. Math. Sci. 147, Springer, New York, 2002.

[5] J.-F. Aujol, *Contribution à l'analyse de textures en traitement d'images par méthodes variationnelles et équations aux dérivées partielles*, Ph.D. thesis, Université de Nice-Sophia Antipolis, Nice, France, 2004.

[6] J. F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle, *Image decomposition into a bounded variation component and an oscillating component*, J. Math. Imaging Vision, 22 (2005), pp. 71–88.

[7] J.-M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions (Nice, 1988), Lecture Notes in Phys. 344, M. Slemrod, M. Rascle, and D. Serre, eds., Springer, Berlin, 1989, pp. 207–215.

[8] P. Benilan, *Equation d'evolution dans un espace de Banach quelconque*, Ph.D. thesis, Université Paris-Sud 11, Orsay, France, 1972.

[9] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[10] J. Bruniquel and A. Lopes, *Analysis and enhancement of multitemporal SAR data*, in Proc. SPIE 2315, SPIE, Bellingham, WA, 1994, pp. 342–353.

[11] A. Buades, B. Coll, and J. M. Morel, *A review of image denoising algorithms, with a new one*, Multiscale Model. Simul., 4 (2005), pp. 490–530.

[12] A. Chambolle, *An algorithm for mean curvature motion*, Interfaces Free Bound., 6 (2004), pp. 1–24.

[13] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97.

[14] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, SIAM, Philadelphia, 2005.
[15] T. R. Crimmins, *Geometric filter for reducing speckle*, Opt. Eng., 25 (1986), pp. 651–654.
[16] J. Darbon, M. Sigelle, and F. Tupin, *A Note on Nice-Levelable MRFs for SAR Image Denoising with Contrast Preservation*, preprint, 2006, http://www.enst.fr/_data/files/docs/id_619_1159280203_271.pdf.
[17] D. L. Donoho and M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.
[18] R. E. Edwards, *Functional Analysis: Theory and Applications*, Holt, Rinehart and Winston, New York, Toronto, London, 1965.
[19] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1991.
[20] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[21] D. Geman and S. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., 6 (1984), pp. 721–741.
[22] J. Gilles, *Décomposition et détection de structures géométriques en imagerie*, Ph.D. thesis, ENS Cachan, Cachan, France, 2006.
[23] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, Switzerland, 1994.
[24] J. W. Goodman, *Statistical properties of laser speckle patterns*, in Laser Speckle and Related Phenomena, 2nd ed., Topics Appl. Phys. 11, Springer, Berlin, 1984.
[25] G. Grimmett and D. Welsh, *Probability: An Introduction*, Oxford Sci. Publ., The Clarendon Press, Oxford University Press, New York, 1986.
[26] F. M. Henderson and A. J. Lewis, *Principles and Applications of Imaging Radar: Manual of Remote Sensing*, Vol. 2, 3rd ed., Wiley and Sons, New York, 1998.
[27] P. Kornprobst, R. Deriche, and G. Aubert, *Image sequence analysis via partial differential equations*, J. Math. Imaging Vision, 11 (1999), pp. 5–26.
[28] K. Krissian, K. Vosburgh, R. Kikinis, and C.-F. Westin, *Anisotropic Diffusion of Ultrasound Constrained by Speckle Noise Model*, Technical report, 2004.
[29] T. Le and L. Vese, *Additive and Multiplicative Piecewise-Smooth Segmentation Models in a Variational Level Set Approach*, UCLA CAM Report 03-52, University of California at Los Angeles, Los Angeles, CA, 2003.
[30] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures*, Univ. Lecture Ser. 22, AMS, Providence, RI, 2001.
[31] D. C. Munson, Jr., and R. L. Visentin, *A signal processing view of strip-mapping synthetic aperture radar*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2131–2147.
[32] M. Nikolova, *Counter-examples for Bayesian and MAP restoration*, IEEE Trans. Image Process., to appear.
[33] A. Ogier and P. Hellier, *A modified total variation denoising method in the context of* 3*D ultrasound images*, in MICCAI'04, Lecture Notes in Comput. Sci. 3216, 2004, pp. 70–77.
[34] S. Osher and N. Paragios, eds., *Geometric Level Set Methods in Imaging, Vision, and Graphics*, Springer, New York, 2003.
[35] T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
[36] L. Rudin, P.-L. Lions, and S. Osher, *Multiplicative denoising and deblurring: Theory and algorithms*, in Geometric Level Set Methods in Imaging, Vision, and Graphics, S. Osher and N. Paragios, eds., Springer, New York, 2003, pp. 103–119.
[37] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[38] E. Tadmor, S. Nezzar, and L. Vese, *A multiscale image representation using hierarchical* (BV, L^2) *decompositions*, Multiscale Model. Simul., 2 (2004), pp. 554–579.
[39] R. Temam, *Navier Stokes Equations*, North–Holland, Amsterdam, 1984.
[40] F. Tupin, M. Sigelle, A. Chkeif, and J.-P. Veran, *Restoration of SAR images using recovery of discontinuities and non-linear optimization*, in EMMCVPR'97, Lecture Notes in Comput. Sci. 1223, J. Van Leuween, G. Goos, and J. Hartemis, eds., Springer, London, 1997.
[41] M. Tur, C. Chin, and J. W. Goodman, *When is speckle noise multiplicative?*, Appl. Optics, 21 (1982), pp. 1157–1159.
[42] Y. Wu and H. Maitre, *Smoothing speckled synthetic aperture radar images by using maximum homogeneous region filters*, Opt. Eng., 31 (1992), pp. 1785–1792.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 947–980

HILL'S EQUATION WITH RANDOM FORCING TERMS*

FRED C. ADAMS† AND ANTHONY M. BLOCH‡

Abstract. Motivated by a class of orbit problems in astrophysics, this paper considers solutions to Hill's equation with forcing strength parameters that vary from cycle to cycle. The results are generalized to include period variations from cycle to cycle. The development of the solutions to the differential equation is governed by a discrete map. For the general case of Hill's equation in the unstable limit, we consider separately the cases of purely positive matrix elements and those with mixed signs; we then find exact expressions, bounds, and estimates for the growth rates. We also find exact expressions, estimates, and bounds for the infinite products of several 2×2 matrices with random variables in the matrix elements. In the limit of sharply spiked forcing terms (the delta function limit), we find analytic solutions for each cycle and for the discrete map that matches solutions from cycle to cycle; for this case we find the growth rates and the condition for instability in the limit of large forcing strength, as well as the widths of the stable/unstable zones.

Key words. Hill's equation, random matrices, orbit problems

AMS subject classifications. 34C11, 15A52

DOI. 10.1137/070689322

1. Introduction.

This paper presents new results concerning Hill's equation of the form

$$\frac{d^2y}{dt^2} + [\lambda_k + q_k \hat{Q}(t)]y = 0, \tag{1}$$

where the function $\hat{Q}(t)$ is periodic, so that $\hat{Q}(t+\pi) = \hat{Q}(t)$, and normalized, so that $\int_0^\pi \hat{Q} dt = 1$. The parameter q_k is denoted here as the forcing strength, which we consider to be a random variable that takes on a new value every cycle (the index k determines the cycle). The parameter λ_k, which determines the oscillation frequency in the absence of forcing, also varies from cycle to cycle. In principal, the duration of the cycle could also vary; our first result (see Theorem 1) shows that this generalized case can be reduced to the problem of (1).

Hill's equations [HI] arise in a wide variety of contexts [MW], and hence the consideration of random variations in the parameters (q_k, λ_k) is a natural generalization of previous work. This particular form of Hill's equation was motivated by a class of orbit problems in astrophysics [AB]. In many astrophysical systems, orbits take place in extended mass distributions with triaxial forms. Examples include dark matter halos that envelop galaxies and galaxy clusters, stellar bulges found at the centers of spiral galaxies, elliptical galaxies, and young embedded star clusters. These systems thus occur over an enormous range of scales, spanning factors of millions in size and

*Received by the editors April 24, 2007; accepted for publication (in revised form) September 28, 2007; published electronically January 30, 2008. This work was supported at the University of Michigan by the Michigan Center for Theoretical Physics; by NASA through the Spitzer Space Telescope Theoretical Research Program; and by the NSF through grants CMS-0408542 and DMS-604307. Some of this work was completed at the Kavli Institute for Theoretical Physics and at U. C. Santa Barbara and was supported in part by the National Science Foundation under grant PHY05-51164.

http://www.siam.org/journals/siap/68-4/68932.html

†Michigan Center for Theoretical Physics, Physics Department, and Astronomy Department, University of Michigan, Ann Arbor, MI 48109 (fca@umich.edu).

‡Michigan Center for Theoretical Physics, Physics Department, and Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (abloch@umich.edu).

factors of trillions in mass. Nonetheless, the basic form of the potential is similar [NF, BE, AB] for all of these systems, and the corresponding orbit problem represents a sizable fraction of the orbital motion that takes place in our universe. In this context, when a test particle (e.g., a star or a dark matter particle) orbits within the triaxial potential, motion that is initially confined to a particular orbital plane can be unstable to motion in the perpendicular direction [AB]. The equation that describes the development of this instability takes the form of (1). Further, the motion in the original orbital plane often displays chaotic behavior, which becomes more extreme as the axis ratios of the potential increase [BT]. In this application, the motion in the original orbit plane—in particular, the distance to the center of the coordinate system—determines the magnitude of the forcing strength q_k that appears in Hill's equation. The crossing time, which varies from orbit to orbit, determines the value of the oscillation parameter λ_k. As a result, the chaotic behavior in the original orbital plane leads to random forcing effects in the differential equation that determines instability of motion out of the plane (see Appendix A for further discussion).

Given that Hill's equations arise in a wide variety of physical problems [MW], we expect that applications with random forcing terms will be common. Although the literature on stochastic differential equations is vast (e.g., see the review of [BL]), specific results regarding Hill's equations with random forcing terms are relatively rare.

In this application, Hill's equation is periodic or nearly periodic (we generalize to the case of varying periods for the basic cycles), and the forcing strength q_k varies from cycle to cycle. Since the forcing strength is fixed over a given cycle, one can solve the Hill's equation for each cycle using previously developed methods [MW], and then match the solutions from cycle to cycle using a discrete map. As shown below, the long-time solution can be developed by repeated multiplication of 2×2 matrices that contain a random component in their matrix elements.

The subject of random matrices, including the long term behavior of their products, is also the subject of a great deal of previous work [BL, DE, BD, FK, FU, LR, ME, VI]. In this application, however, Hill's equation determines the form of the random matrices, and the repeated multiplication of this type of matrix represents a new and specific application. Given that instances where analytic results can be obtained are relatively rare, this set of solutions adds new examples to the list of known cases.

This paper is organized as follows. In section 2, we present the basic formulation of the problem, define relevant quantities, and show that aperiodic generalizations of the problem can be reduced to random Hill's equations. The following section (section 3) presents the main results of the paper: We find specific results regarding the growth rates of instability for random Hill's equations in the limit of large forcing strengths (i.e., in the limit where the equations are robustly unstable). These results are presented for purely positive and for mixed signs in the 2×2 matrix map. We also find limiting forms and constraints on the growth rates. Finally, we find bounds and estimates for the errors incurred by working in the limit of large forcing strengths. This work is related to the general existence results of [FU] but provides much more detailed information in our setting. In the next section (section 4) we consider the limit where the forcing terms are Dirac delta functions; this case allows for analytic solutions to the original differential equation. We note that the growth rates calculated here (section 3) depend on the distribution of the ratios of the principal solutions to (1), rather than (directly) on the distributions of the parameters (λ_k, q_k). Using the analytic solutions for the delta function limit (section 4), we thus gain insight into the transformation between the distributions of the input parameters (λ_k, q_k) and the parameters that specify the growth rates. Finally, we conclude, in section 5, with a summary and discussion of our results.

2. Formulation.

Definition. *A* random Hill's equation *is defined here to be of the form given by* (1)*, where the forcing strength q_k and oscillation parameter λ_k vary from cycle to cycle. Specifically, the parameters q_k and λ_k are stochastic variables that take on new values every cycle $0 \leq [t] \leq \pi$, and the values are sampled from known probability distributions $P_q(q)$ and $P_\lambda(\lambda)$.*

2.1. Hill's equation with fixed parameters.

Over a single given cycle, a random Hill's equation is equivalent to an ordinary Hill's equation and can be solved using known methods [MW].

Definition. *The* growth factor f_c *per cycle (the Floquet multiplier) is given by the solution to the characteristic equation and can be written as*

$$f_c = \frac{\Delta + (\Delta^2 - 4)^{1/2}}{2}, \tag{2}$$

where the discriminant $\Delta = \Delta(q, \lambda)$ is defined by

$$\Delta \equiv y_1(\pi) + \frac{dy_2}{dt}(\pi), \tag{3}$$

and where y_1 and y_2 are the principal solutions [MW].

It follows from Floquet's theorem that $|\Delta| > 2$ is a sufficient condition for instability [MW, AS]. In addition, periodic solutions exist when $|\Delta| = 2$.

2.2. Random variations in forcing strength.

We now generalize to the case where the forcing strength q_k and oscillation parameter λ_k vary from cycle to cycle. In other words, we consider each period from $t = 0$ to $t = \pi$ as a cycle and consider the effects of successive cycles with varying values of (q_k, λ_k).

During any given cycle, the solution can be written as a linear combination of the two principal solutions y_1 and y_2. Consider two successive cycles. The first cycle has parameters (q_a, λ_a) and solution

$$f_a(t) = \alpha_a y_{1a}(t) + \beta_a y_{2a}(t), \tag{4}$$

where the solutions $y_{1a}(t)$ and $y_{2a}(t)$ correspond to those for an ordinary Hill's equation when evaluated using the values (q_a, λ_a). Similarly, for the second cycle with parameters (q_b, λ_b) the solution has the form

$$f_b(t) = \alpha_b y_{1b}(t) + \beta_b y_{2b}(t). \tag{5}$$

Next we note that the new coefficients α_b and β_b are related to those of the previous cycle through the relations

$$\alpha_b = \alpha_a y_{1a}(\pi) + \beta_a y_{2a}(\pi) \qquad \text{and} \qquad \beta_b = \alpha_a \frac{dy_{1a}}{dt}(\pi) + \beta_a \frac{dy_{2a}}{dt}(\pi). \tag{6}$$

The new coefficients can thus be considered as a two dimensional vector, and the transformation between the coefficients in one cycle and the next cycle is a 2 × 2 matrix. Here we consider the case in which the equation is symmetric with respect to the midpoint $t = \pi/2$. This case arises in the original orbit problem that motivated this study—the forcing function is determined by the orbit as it passes near the center of the potential and this passage is symmetric (or very nearly so). It also makes sense to consider the symmetric case, which is easier, first. Since the Wronskian of the original differential equation is unity, the number of independent matrix coefficients is reduced further, from four to two. We thus have the following result.

PROPOSITION 1. *The transformation between the coefficients α_a, β_a of one cycle and the coefficients α_b, β_b of the next may be written in the form*

$$\begin{bmatrix}\alpha_b\\ \beta_b\end{bmatrix} = \begin{bmatrix} h & (h^2-1)/g \\ g & h\end{bmatrix}\begin{bmatrix}\alpha_a\\ \beta_a\end{bmatrix} \equiv \mathbf{M}(q_a)\begin{bmatrix}\alpha_a\\ \beta_a\end{bmatrix}, \tag{7}$$

where the matrix $\mathbf{M}$ *(defined in the second equality) depends on the values* (q_a, λ_a) *and* $h = y_1(\pi)$ *and* $g = \dot{y}_1(\pi)$ *for a given cycle.*

Proof. This result can be verified by standard matrix multiplication, which yields (6) above. □

After N cycles with varying values of (q_k, λ_k), the solution retains the general form given above, where the coefficients are determined by the product of matrices according to

$$\begin{bmatrix}\alpha_N\\ \beta_N\end{bmatrix} = \mathbf{M}^{(N)}\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}, \qquad \text{where} \qquad \mathbf{M}^{(N)} \equiv \prod_{k=1}^{N}\mathbf{M}_k(q_k,\lambda_k). \tag{8}$$

This formulation thus transforms the original differential equation (with a random element) into a discrete map. The properties of the product matrix $\mathbf{M}^{(N)}$ determine whether the solution is unstable and the corresponding growth rate.

DEFINITION. *The* asymptotic growth rate γ_∞ *is that experienced by the system when each cycle amplifies the growing solution by the growth factor appropriate for the given value of the forcing strength for that cycle, i.e.,*

$$\gamma_\infty \equiv \lim_{N\to\infty}\frac{1}{\pi N}\log\left[\prod_{k=1}^{N}\frac{1}{2}\left\{\Delta_k + \sqrt{\Delta_k^2-4}\right\}\right], \tag{9}$$

where $\Delta_k = \Delta(q_k, \lambda_k)$ *is defined by* (3), *and where this expression is evaluated in the limit* $N \to \infty$. *In this definition, it is understood that if* $|\Delta_k| < 2$ *for a particular cycle, then the growth factor is unity for that cycle, resulting in no net contribution to the product (for that cycle).*

Notice that the factor of π appears in this definition of the growth rate because the original Hill's equation is taken to be π-periodic [MW, AS]. As we show below, the growth rates of the differential equation are determined by the growth rates resulting from matrix multiplication. In many cases, however, the growth rates for matrix multiplication are given without the factor of π [BL, FK], so there is a mismatch of convention (by a factor of π) between growth rates of Hill's equations and growth rates of matrix multiplication.

Notice that this expression for the asymptotic growth rate takes the form

$$\gamma_\infty = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\gamma(q_k,\lambda_k) \to \langle\gamma\rangle, \tag{10}$$

where $\gamma(q_k, \lambda_k)$ is the growth rate for a given cycle. The asymptotic growth rate is thus given by the expectation value of the growth rate per cycle for a given probability distribution for the parameters q_k and λ_k.

We note that a given system does not necessarily experience growth at the rate γ_∞ because the solutions must remain continuous across the boundaries between subsequent cycles. This requirement implies that the solutions during every cycle will contain an admixture of both the growing solution and the decaying solution for that cycle, thereby leading to the possibility of slower growth. In some cases, however, the growth rate is larger than γ_∞; i.e., the stochastic component of the problem aids and abets the instability. One could also call γ_∞ the direct growth rate.

2.3. Generalization to aperiodic variations.

THEOREM 1. *Consider a generalization of Hill's equation so that the cycles are no longer exactly π-periodic. Instead, each cycle has period $\mu_k\pi$, where μ_k is a random variable that averages to unity. Then variations in period are equivalent to variations in (q,λ); i.e., the problem with three stochastic variables (q_k,λ_k,μ_k) reduces to a π-periodic problem with only two stochastic variables (q_k,λ_k).*

Proof. With this generalization, the equation of motion takes the form

$$\frac{d^2y}{dt^2}+\left[\lambda_k+q_k\hat{Q}(\mu_k t)\right]y=0, \tag{11}$$

where we have normalized the forcing frequency to have unit amplitude ($\hat{Q}=Q/q_k$). Since $\hat{Q}$ (and Q) are π-periodic, the jth cycle is defined over the time interval $0\le\mu_k t\le\pi$, or $0\le t\le\pi/\mu_k$. We can rescale both the time variable and the "constants" according to

$$t\to\mu_k t\,,\qquad \lambda_k\to\lambda_k/\mu_k^2=\widetilde{\lambda}_j\,,\qquad\text{and}\qquad q_k\to q_k/\mu_k^2=\widetilde{q}_j, \tag{12}$$

so the equation of motion reduces to the familiar form

$$\frac{d^2y}{dt^2}+\left[\widetilde{\lambda}_j+\widetilde{q}_j\hat{Q}(t)\right]y=0. \tag{13}$$

Thus, the effects of varying period can be incorporated into variations in the forcing strength q_k and oscillation parameter λ_k. □

3. Hill's equation in the unstable limit. In this section we consider Hill's equation in the general form (for the delta function limit see section 4) but restrict our analysis to the case of symmetric potentials so that $y_1(\pi)=h=\dot{y}_2(\pi)$. We also consider the *highly unstable limit*, where we define this limit to correspond to large $h\gg 1$. Since the 2×2 matrix of the discrete map must have its determinant equal to unity, the matrix of the map has the form given by (7), where the values of h and g depend on the form of the forcing potential.

The discrete map can be rewritten in the general form

$$\mathbf{M}=h\begin{bmatrix}1 & x\\ 1/x & 1\end{bmatrix}+\begin{bmatrix}0 & -1/g\\ 0 & 0\end{bmatrix}. \tag{14}$$

In the highly unstable limit $h\to\infty$, the matrix simplifies to the approximate form

$$\mathbf{M}\approx h\begin{bmatrix}1 & x\\ 1/x & 1\end{bmatrix}\equiv h\mathbf{C}, \tag{15}$$

where we have defined $x\equiv h/g$, and where the second equality defines the matrix $\mathbf{C}$.

In this problem we are concerned with both the long-time limit $N\to\infty$ and the "unstable" limit $h\to\infty$. In the first instance considered here we take the unstable limit first, but below we analyze precisely the difference between taking the long time limit first and then the unstable limit.

3.1. Fixed matrix of the discrete map. The simplest case occurs when the stochastic component can be separated from the matrix, i.e., when the matrix $\mathbf{C}$ does not vary from cycle to cycle. This case arises when the Hill's equation does not contain a random component; it also arises when the random component can be factored out so that x does not vary from cycle to cycle, although the leading factors h_k can vary. In either case, the matrix $\mathbf{C}$ is fixed. Repeated multiplications of the matrix $\mathbf{C}$ are then given by

$$\mathbf{C}^N = 2^{N-1}\mathbf{C}. \tag{16}$$

With this result, after N cycles the Floquet multiplier (eigenvalue) of the product matrix and the corresponding growth rate take the form

$$\Lambda = \prod_{k=1}^{N} (2h_k) \qquad \text{and} \qquad \gamma = \lim_{N\to\infty} \frac{1}{\pi N} \sum_{k=1}^{N} \log(2h_k). \tag{17}$$

Note that this result applies to the particular case of Hill's equation in the delta function limit (section 4), where the forcing strength q_k varies from cycle to cycle but the frequency parameter λ_k is constant. The growth rate in (17) is equal to the asymptotic growth rate γ_∞ (see (9)) for this case.

3.2. General results in the unstable limit. We now generalize to the case where the parameters of the differential equation vary from cycle to cycle. For a given cycle, the discrete map is specified by a matrix of the form specified by (15), where $x = x_k = h_k/g_k$, with varying values from cycle to cycle. The values of x_k depend on the parameters (q_k, λ_k) through the original differential equation. After N cycles, the product matrix $\mathbf{M}^{(N)}$ takes the form

$$\mathbf{M}^{(N)} = \prod_{k=1}^{N} h_k \, \prod_{k=1}^{N} \mathbf{C}_k, \tag{18}$$

where we have separated the two parts of the problem. One can show (by induction) that the product of N matrices $\mathbf{C}_k$ has the form

$$\mathbf{C}^{(N)} = \prod_{k=1}^{N} \mathbf{C}_k = \begin{bmatrix} \Sigma_{T(N)} & x_1 \Sigma_{T(N)} \\ \Sigma_{B(N)}/x_1 & \Sigma_{B(N)} \end{bmatrix}, \tag{19}$$

where x_1 is the value of the variable for the first cycle and where the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ are given by

$$\Sigma_{T(N)} = \sum_{j=1}^{2^{N-1}} r_j \qquad \text{and} \qquad \Sigma_{B(N)} = \sum_{j=1}^{2^{N-1}} \frac{1}{r_j}, \tag{20}$$

where the variables r_j are ratios of the form

$$r_j = \frac{x_{a_1} x_{a_2} \dots x_{a_n}}{x_{b_1} x_{b_2} \dots x_{b_n}}. \tag{21}$$

The ratios r_j arise from repeated multiplication of the matrices $\mathbf{C}_k$, and hence the indices lie in the range $1 \le a_i, b_i \le N$. The r_j always have the same number of factors in the numerator and the denominator, but the number of factors (n) varies from 0 (where $r_j = 1$) up to $N/2$. This upper limit arises because each composite ratio r_j has $2n$ values of x_j, which must all be different, and because the total number of possible values is N.

Next we define a composite variable

$$\widetilde{S} \equiv \frac{1}{2^N}\left[\Sigma_{T(N)} + \Sigma_{B(N)}\right] = \frac{1}{2^N}\sum_{j=1}^{2^{N-1}}\left(r_j + \frac{1}{r_j}\right). \tag{22}$$

With this definition, the (growing) eigenvalue Λ of the product matrix $\mathbf{M}^{(N)}$ takes the form

$$\Lambda = \widetilde{S}\prod_{k=1}^{N}(2h_k) \tag{23}$$

and the corresponding growth rate of the instability has the form

$$\gamma = \lim_{N\to\infty}\left[\frac{1}{\pi N}\sum_{k=1}^{N}\log(2h_k) + \frac{1}{N\pi}\log\widetilde{S}\right]. \tag{24}$$

The first term is the asymptotic growth rate γ_∞ defined by (9) and is thus an average of the growth rates for the individual cycles. All of the additional information regarding the stochastic nature of the differential equation is encapsulated in the second term through the variable $\widetilde{S}$. For example, if the composite variable $\widetilde{S}$ is finite in the limit $N \to \infty$, then the second term would vanish. As shown below, however, the stochastic component can provide a significant contribution to the growth rate and can provide either a stabilizing or destabilizing influence. In the limit $N \to \infty$, we can thus write the growth rate in the from

$$\gamma = \gamma_\infty + \Delta\gamma, \tag{25}$$

where we have defined the correction term $\Delta\gamma$,

$$\Delta\gamma \equiv \lim_{N\to\infty}\frac{1}{N\pi}\log\widetilde{S}. \tag{26}$$

Since the asymptotic growth rate γ_∞ is straightforward to evaluate, the remainder of this section focuses on evaluating $\Delta\gamma$ as well as finding corresponding estimates and constraints. This correction term $\Delta\gamma$ is determined by the discrete map $\mathbf{C}$, whose matrix elements are given by the ratios $x = h/g$, where h and g are determined by the solutions to Hill's equation over one cycle. One should keep in mind that the parameters in the original differential equation are (λ_k, q_k). The distribution of these parameters determines the distributions of the principal solutions (the distributions of h_k and g_k), whereas the distribution of the ratios x_k of these latter quantities determines the correction $\Delta\gamma$ to the growth rate. The problem thus separates into two parts: (1) the transformation between the distributions of the parameters (λ_k, q_k) and the resulting distribution of the ratios x_k that define the discrete map, and (2) the growth rate of the discrete map for a given distribution of x_k. The following analysis focuses on the latter issue (whereas section 4 provides an example of the former issue).

3.3. Growth rates for positive matrix elements. This subsection addresses the cases where the ratios x_k that define the discrete map $\mathbf{C}$ all have the same sign. For this case, the analysis is simplified, and a number of useful results can be obtained.

THEOREM 2. *Consider the general form of Hill's equation in the unstable limit so that $h = y_1(\pi) = \dot{y}_2(\pi) \gg 1$. For the case of positive matrix elements, $r_j > 0$, the growth rate is given by* (25), *where the correction term $\Delta\gamma$ is given by*

$$\Delta\gamma = \lim_{N\to\infty} \frac{1}{\pi N} \sum_{j=1}^{N} \log(1 + x_{j1}/x_{j2}) - \frac{\log 2}{\pi}, \tag{27}$$

where x_{j1} and x_{j2} represent two different (independent) samples of the x_j variable.[1]

Proof. Using the same induction argument that led to (19), one finds that from one cycle to the next the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ vary according to

$$\Sigma_{T(N+1)} = \Sigma_{T(N)} + \frac{x}{x_0}\Sigma_{B(N)} \tag{28}$$

and

$$\Sigma_{B(N+1)} = \Sigma_{B(N)} + \frac{x_0}{x}\Sigma_{T(N)}. \tag{29}$$

In this notation, the variable x (no subscript) represents the value of the x variable at the current cycle, whereas x_0 represents the value at the initial cycle. The growing eigenvalue of the product matrix of (19) is given by $\Lambda = \Sigma_{T(N)} + \Sigma_{B(N)}$. As a result, the eigenvalue (growth factor) varies from cycle to cycle according to

$$\Lambda^{(N+1)} = \Lambda^{(N)} + \frac{x}{x_0}\Sigma_{B(N)} + \frac{x_0}{x}\Sigma_{T(N)} = \Lambda^{(N)}\left[1 + \frac{(x/x_0)\Sigma_{B(N)} + (x_0/x)\Sigma_{T(N)}}{\Sigma_{B(N)} + \Sigma_{T(N)}}\right]. \tag{30}$$

The overall growth factor is then determined by the product

$$\Lambda^{(N)} = \prod_{j=1}^{N}\left[1 + \frac{(x/x_0)\Sigma_{B(N)} + (x_0/x)\Sigma_{T(N)}}{\Sigma_{B(N)} + \Sigma_{T(N)}}\right]. \tag{31}$$

The growth rate of matrix multiplication is determined by setting the above product equal to $\exp[N\pi\gamma]$. The growth rate $\Delta\gamma$ also includes the factor of 2 per cycle that is included in the definition of the asymptotic growth rate γ_∞. We thus find that

$$\Delta\gamma \approx \frac{1}{N\pi}\sum_{j=1}^{N}\log\left[1 + \frac{(x_{j1}/x_{j2})\Sigma_{B(N)} + (x_{j2}/x_{j1})\Sigma_{T(N)}}{\Sigma_{B(N)} + \Sigma_{T(N)}}\right] - \frac{\log 2}{\pi}. \tag{32}$$

Note that this expression provides the correction $\Delta\gamma$ to the growth rate. The full growth rate is given by $\gamma = \gamma_\infty + \Delta\gamma$ (where γ_∞ is specified by (9) and $\Delta\gamma$ is specified by (27)). In the limit of large N, the ratio of the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ approaches unity, almost surely, so that

$$\frac{\Sigma_{T(N)}}{\Sigma_{B(N)}} \to 1 \qquad \text{as} \qquad N \to \infty. \tag{33}$$

[1] Specifically, the index j labels the cycle number, and the indices $j1$ and $j2$ label two successive samples of the x variable; since the stochastic parameters of the differential equations are assumed to be independent from cycle to cycle, however, the variables x_{j1} and x_{j2} can be any independent samples.

This result follows from the definition of $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$: The terms in each of these two sums are the product of ratios x_a/x_b, and the terms r_j in the first sum $\Sigma_{T(N)}$ are the inverse of those $(1/r_j)$ in the second sum $\Sigma_{B(N)}$. Since the fundamental variables x_k that make up these ratios, and the products of these ratios, are drawn from the same distribution, the above condition (33) must hold. As a consequence, the expression for the growth rate given by (32) approaches that of (27). □

COROLLARY 2.1. *Let σ_0 be the variance of the composite variable* $\log(x_{j1}/x_{j2})$ *(see Theorem* 2*). The correction to the growth rate is positive semidefinite; specifically, $\Delta\gamma \geq 0$ and $\Delta\gamma \to 0$ in the limit $\sigma_0 \to 0$. Further, in the limit of small variance, the growth rate approaches the asymptotic form $\Delta\gamma \to \sigma_0^2/(8\pi)$.*

Proof. In the limit of small σ_0, we can write $x_j = 1 + \delta_j$, where $|\delta_j| \ll 1$. In this limit, (27) for the growth rate becomes

$$\Delta\gamma = \lim_{N\to\infty} \frac{1}{\pi N} \sum_{j=1}^{N} \log\big[2 + \delta_{j1} - \delta_{j2} + \delta_{j2}^2 - \delta_{j1}\delta_{j2} + \mathcal{O}(\delta^3)\big] - \frac{\log 2}{\pi}. \tag{34}$$

In the limit $|\delta_j| \ll 1$, we can expand the logarithm, and the above expression simplifies to the form

$$\Delta\gamma = \lim_{N\to\infty} \frac{1}{2\pi N} \sum_{j=1}^{N} \big[\delta_{j1} - \delta_{j2} + \delta_{j2}^2 - \delta_{j1}\delta_{j2} - (\delta_{j1} - \delta_{j2})^2/4 + \mathcal{O}(\delta^3)\big]. \tag{35}$$

Evaluation of the above expression shows that

$$\Delta\gamma = \frac{1}{2\pi}\left[\langle\delta_{j2}^2\rangle - \frac{1}{4}\langle(\delta_{j1} - \delta_{j2})^2\rangle + \mathcal{O}(\delta^3)\right] \to \frac{\sigma_0^2}{8\pi}. \tag{36}$$

As a result, $\Delta\gamma \geq 0$. In the limit $\sigma_0 \to 0$, all of the x_j approach unity and $\delta_j \to 0$; therefore, $\Delta\gamma \to 0$ as $\sigma_0 \to 0$. □

Although (27) is exact, the computation of the expectation value can be difficult in practice. As a result, it is useful to have simple constraints on the growth rate in terms of the variance of the probability distribution for the variables x_k. In particular, a simple bound can be derived.

THEOREM 3. *Consider the general form of Hill's equation in the unstable limit so that $h = y_1(\pi) = \dot{y}_2(\pi) \gg 1$. Take the variables $r_j > 0$. Then the growth rate is given by* (25) *and the correction term $\Delta\gamma$ obeys the constraint*

$$\Delta\gamma \leq \frac{\sigma_0^2}{4\pi}, \tag{37}$$

where σ_0^2 is the variance of the distribution of the variable $\xi = \log(x_{j1}/x_{j2})$, and where x_j are independent samplings of the ratios $x_j = h_j/g_j$.

Proof. First we define the variable $\xi_j = \log r_j$, where r_j is given by (21) above with a fixed value of n. In the limit of large n, the variable ξ_j has zero mean and will be normally distributed. If the variables x_j are independent, the variance of the composite variable ξ_j will be given by

$$\sigma_\xi^2 = n\sigma_0^2. \tag{38}$$

As shown below, in order to obtain 2^N terms in the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$, almost all of the variables r_j will fall in the large n limit; in addition, $n \to \infty$ in the limit $N \to \infty$.

As a result, we can consider the large n limit to be valid for purposes of evaluating the correction term $\Delta\gamma$. In practice, the variables will not be completely independent, so the actual variance will be smaller than that given by (38); nonetheless, this form can be used to find the desired upper limit.

Given the large n limit and a log-normal distribution of r_j, the expectation values $\langle r_j \rangle$ and $\langle 1/r_j \rangle$ are given by

$$\langle r_j \rangle = \exp\big[n\sigma_0^2/2\big] = \langle 1/r_j \rangle. \tag{39}$$

Note that the variable ξ_j is normally distributed, and we are taking the expectation value of $r_j = \exp \xi_j$; since the mean of the exponential is not necessarily equal to the exponential of the mean, the above expression contains the (perhaps counterintuitive) factor of 2. As expected, larger values of n allow for a wider possible distribution and result in larger expectation values. The maximum expectation values thus occur for the largest values of n. Since $n < N/2$, these results, in conjunction with (22), imply that $\widetilde{S}$ obeys the constraint

$$\widetilde{S} < \exp\big[N\sigma_0^2/4\big]. \tag{40}$$

The constraint claimed in (37) then follows immediately.

Combinatorics. To complete the argument, we must show that most of the variables r_j have a large number n of factors (in the limit of large N). The number of terms in the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ is large, namely, 2^{N-1}. Further, the ratios r_j must contain $2n$ different values of the variables x_k. The number $P(n|N)$ of different ways to choose the $2n$ variables for N cycles (and hence N possible values of x_k) is given by the expression

$$P(n|N) = \frac{N!}{(N-2n)!(n!)^2}. \tag{41}$$

Notice that this expression differs from the more familiar binomial coefficient because the values of r_j depend on whether or not the x_k factors are in the numerator or denominator of the ratio r_j. Next we note that if $n \ll N$, then the following chain of inequalities holds for large N:

$$P(n|N) < \frac{N^{2n}}{(n!)^2} \ll 2^{N-1}. \tag{42}$$

For large N and $n \ll N$, the central expression increases like a power of N, whereas the right-hand expression increases exponentially with N. As a result, for $n \ll N$, there are not enough different ways to choose the x_k values to make the required number of composite ratios r_j. In order to allow for enough different r_j, the number n of factors must be large (namely, large enough so that $n \ll N$ does not hold) for most of the r_j. This conclusion thus justifies our use of the large n limit in the proof of Theorem 3 (where we used a log-normal form for the composite distribution to evaluate the expectation values $\langle r_j \rangle$ and $\langle 1/r_j \rangle$). □

Estimate. Theorem 3 provides an upper bound on the contribution of the correction term $\Delta\gamma$ to the overall growth rate. This bound depends on the value of n, which determines the magnitude of the expectation value $\langle r_j \rangle$. It is useful to have an estimate of the "typical" size of n. In rough terms, the value of n must be large enough so that the number of possible combinations is large enough to account for the 2^{N-1} terms in the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$. For each n, we have $P(n|N)$ combinations.

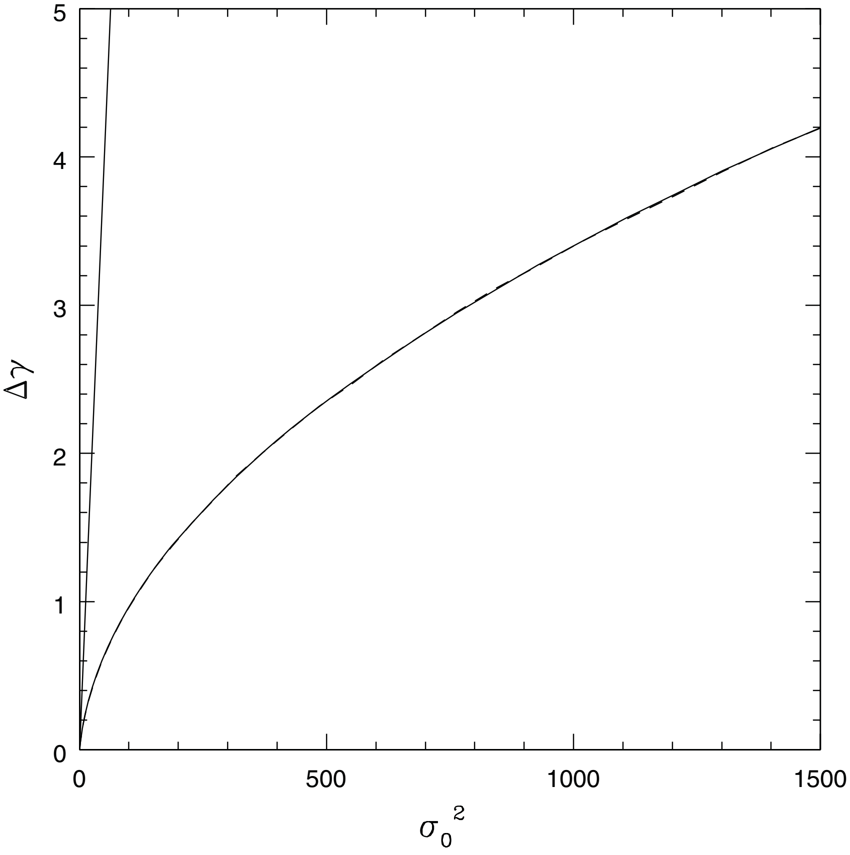


FIG. 1. *Comparison of the bound of Theorem* 3 *and the prediction of Theorem* 2 *with results from numerical experiments. All cases use matrices* $\mathbf{C}_k$ *of the form given by* (15)*, where the variables* x_k *are chosen according to distributions with variance* σ_0^2*. For each distribution, the growth rate* $\Delta\gamma$ *due to matrix multiplication is plotted versus the variance of the distribution of the composite variable* $\xi = \log(x_j/x_k)$*, where* $x_k = y_{1k}(\pi)/\dot{y}_{1k}(\pi)$ *and, similarly,* $x_j = y_{1j}(\pi)/\dot{y}_{1j}(\pi)$*. The solid curve shows the results obtained by averaging* 1000 *realizations of the numerical experiments; the overlying dashed curve shows the prediction of Theorem* 2*. The straight solid line shows the upper bound of Theorem* 3*, i.e.,* $\Delta\gamma \le \sigma_0^2/(4\pi)$*.*

As a rough approximation, $nP(n|N)$ accounts for all of the combinations of size less than n. If we set $nP(n|N) = 2^N$, we can solve for the ratio n/N required to have enough terms and find $n/N \approx 0.11354\cdots \approx 1/9$. As a result, we expect the ratio n/N to lie in the range

$$\frac{1}{9} < \frac{n}{N} < \frac{1}{2}. \tag{43}$$

If we use this range of n/N to evaluate the expectation value using (39) and estimate the growth rate, the upper end of this range provides a rigorous upper bound (Theorem 3). The lower end of the range represents only a rough guideline, however, since the variables are not fully independent. Nonetheless, it can be used to estimate the expectation values $\langle r_j \rangle$.

Notice that the upper bound is conservative. Figure 1 shows a comparison of the actual growth rate (from Theorem 2) and the bound (Theorem 3). At large variance, the actual growth rate is much less than our bound. In fact, as shown in the following section, in the limit of large variance, the growth rate $\Delta\gamma \propto \sigma_0$ (rather than $\propto \sigma_0^2$).

For this numerical experiment, we used a particular form for the x_k variables, namely, $x_k = 0.01 + (10a\xi_k)^a$, where ξ_k is a random variable in the range $0 \le \xi_k \le 1$ and a is a parameter that is chosen to attain varying values of σ_0^2. The exact form of the curve $\Delta\gamma(\sigma_0^2)$ depends on the distribution of the x_k. However, all of the distributions studied result in the general form shown in Figure 1, and all of the cases show the same agreement between numerical experiments and the predictions of Theorem 2.

3.4. Error bounds and estimates. The analysis presented thus far is valid in the highly unstable limit, as defined at the beginning of this section. In other words, we have found an exact solution to the reduced problem, as encapsulated in (15). In this problem we are taking two limits—the long-time limit $N \to \infty$ and the "unstable" limit $h \to \infty$. In the reduced problem, as analyzed above, we take the limit $h \to \infty$ first and then consider the long-time limit $N \to \infty$. In this subsection, we consider the accuracy of this approach by finding bounds (and estimates) for the errors in the growth rates incurred from working in the highly unstable limit. In other words, we find bounds on the difference between the results for the full problem (with large but finite h_k) and the reduced problem.

To assess the error budget, we write the general matrix (for the full problem) in the form

$$\mathbf{M} = h\mathbf{B}, \qquad \text{where} \qquad \mathbf{B} \equiv \begin{bmatrix} 1 & x\phi \\ 1/x & 1 \end{bmatrix}. \tag{44}$$

This form is the same as the matrix of the reduced problem (in the unstable limit) except for the correction factor ϕ in the (1,2) matrix element, where $\phi \equiv (1 - 1/h^2)$.

Let $(\Delta\gamma)_B$ denote the growth rate for the matrix $\mathbf{B}$ for the full problem defined in (44). Similarly, let $(\Delta\gamma)_C$ denote the growth rate found previously for the reduced problem using the matrix $\mathbf{C}$ defined in (15). Through repeated matrix multiplications, the product of matrices $\mathbf{B}_k$ will be almost the same as for the product of matrices $\mathbf{C}_k$, where the difference is due to the continued accumulation of factors ϕ_k. Note that the index k, as introduced here, denotes the cycle number, and that all of these quantities vary from cycle to cycle.

Proposition 2. *The error $\varepsilon_{BC} = (\Delta\gamma)_C - (\Delta\gamma)_B$ introduced by using the reduced form of the problem (the matrices $\mathbf{C}_k$) instead of the full problem (the matrices $\mathbf{B}_k$) is bounded by*

$$0 < \varepsilon_{BC} < -\frac{1}{\pi}\langle \log \phi_k \rangle. \tag{45}$$

Proof. Since $\phi_k < 1$, by definition, we see immediately that the growth rate for the full problem is bounded from above by that of the reduced problem, i.e.,

$$(\Delta\gamma)_B < (\Delta\gamma)_C. \tag{46}$$

Next we construct a new matrix of the form

$$\mathbf{A} \equiv \phi \begin{bmatrix} 1 & x \\ 1/x & 1 \end{bmatrix} = \phi\mathbf{C}. \tag{47}$$

The products of the matrices $\mathbf{A}_k$ will be almost the same as those for the matrices $\mathbf{B}_k$, where the difference is again due to the inclusion of additional factors of ϕ_k. Since the $\phi_k < 1$, we find that the growth rate for this benchmark problem is less than (or equal to) that of the full problem, i.e., $(\Delta\gamma)_A < (\Delta\gamma)_B$. Further, the growth rate $(\Delta\gamma)_A$ for this new matrix can be found explicitly and is given by

$$(48)\qquad (\Delta\gamma)_A = (\Delta\gamma)_C + \lim_{N\to\infty} \frac{1}{\pi N} \log \left[\prod_{k=1}^{N} \phi_k\right] = (\Delta\gamma)_C + \lim_{N\to\infty} \frac{1}{\pi N} \sum_{k=1}^{N} \log \phi_k .$$

Combining (46) and (48) shows that the growth rate for the full problem $(\Delta\gamma)_B$ is bounded on both sides and obeys the constraint

$$(49)\qquad (\Delta\gamma)_C + \frac{1}{\pi}\langle \log \phi_k \rangle < (\Delta\gamma)_B < (\Delta\gamma)_C .$$

Notice that the expectation value $\langle \log \phi_k \rangle < 0$ since $\phi_k < 1$. The error ε_{BC} introduced by using the reduced form of the problem (the matrices $\mathbf{C}_k$) instead of the full problem (the matrices $\mathbf{B}_k$) is thus bounded by

$$(50)\qquad 0 < \varepsilon_{BC} < -\frac{1}{\pi}\langle \log \phi_k \rangle .$$

This bound can be made tighter by a factor of 2. Note that the product of two matrices of the full problem has the form

$$(51)\qquad \mathbf{B}_2\mathbf{B}_1 = \begin{bmatrix} 1 + (x_2/x_1)\phi_2 & x_1\phi_1 + x_2\phi_2 \\ 1/x_1 + 1/x_2 & 1 + (x_1/x_2)\phi_1 \end{bmatrix} .$$

Thus, the product of two matrices contains only linear factors of ϕ_k. As a result, we can define a new reference matrix $\widetilde{\mathbf{A}} = \phi^{1/2}\mathbf{C}$ that accumulates factors of ϕ_k only half as quickly as the original matrix $\mathbf{A}$ in the above argument, so that

$$(52)\qquad \widetilde{\mathbf{A}}_2\widetilde{\mathbf{A}}_1 = \phi_1^{1/2}\phi_2^{1/2} \begin{bmatrix} 1 + x_2/x_1 & x_1 + x_2 \\ 1/x_1 + 1/x_2 & 1 + x_1/x_2 \end{bmatrix} = \phi_1^{1/2}\phi_2^{1/2}\mathbf{C}_2\mathbf{C}_1 .$$

The new reference matrix still grows more slowly than the matrix $\mathbf{B}$ of the full problem, but the product of N such matrices accumulates only N extra factors of $\phi_k^{1/2}$. Using this reference matrix in the above argument results in the tighter bound

$$(53)\qquad 0 < \varepsilon_{BC} < -\frac{1}{2\pi}\langle \log \phi_k \rangle .$$

In the limit where all of the $h_k \gg 1$, $\log \phi_k \approx -1/h_k^2$, and the above bound approaches the approximate form

$$(54)\qquad 0 < \varepsilon_{BC} < \frac{1}{2\pi}\langle h_k^{-2} \rangle .$$

This expression shows that the errors are well controlled. For large but finite h_k, the departure of the growth rates from those obtained in the highly unstable limit (Theorem 2) are $\mathcal{O}(h_k^{-2})$. □

Given the above considerations, we can write the growth rate $(\Delta\gamma)_B$ for the full problem in the form

$$(55)\qquad (\Delta\gamma)_B = (\Delta\gamma)_C - \frac{K_\varepsilon}{\pi}\langle h_k^{-2} \rangle ,$$

where $(\Delta\gamma)_C$ is the growth rate for the reduced problem and where K_ε is a constant of order unity. In the limit of large h_k (specifically for $\log\phi_k \approx 1/h_k^2$), the constant is bounded and lies in the range $0 < K_\varepsilon < 1/2$. Our numerical exploration of parameter space suggests that $K_\varepsilon \approx 1/4$ provides a good estimate for the correction term. In any case, however, the correction term depends on h_k through the quantity $\langle h_k^{-2}\rangle$ and decreases with the size of this expectation value.

3.5. Matrix elements with varying signs. We now consider the case in which the signs of the variables r_j can be either positive or negative. Suppose that the system has equal probability of attaining positive and negative factors. In the limit $N \to \infty$, one expects the sums $\Sigma_{T(N)}, \Sigma_{B(N)} \to 0$, which would seem to imply no growth. However, two effects counteract this tendency. First, the other factor that arises in the repeated matrix multiplication diverges in the same limit, i.e.,

$$\prod_k^N (2h_k) \to \infty \qquad \text{as} \qquad N \to \infty. \tag{56}$$

Second, the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ can random walk away from zero with an increasing number N of cycles, where the effective step length is determined by the variance σ_0 defined previously. If the random walk is fast enough, the system can be unstable even without considering the diverging product of (56). In order to determine the stability (or instability) of the Hill's equation in this case, we must thus determine how the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ behave with increasing N.

Theorem 4. *Consider the case of Hill's equation in the unstable limit with both positive and negative signs for the matrix elements. Let positive signs occur with probability p and negative signs occur with probability $1-p$. Then the general form of the growth rate is given by*

$$\Delta\gamma = \lim_{N\to\infty} \frac{1}{\pi N} \left\{ \left[p^2 + (1-p)^2\right] \sum_{j=1}^N \log\Big(1 + \Big|\frac{x_{j1}}{x_{j2}}\Big|\Big) + 2p(1-p) \sum_{k=1}^N \log\Big|1 - \Big|\frac{x_{k1}}{x_{k2}}\Big|\Big| \right\} - \frac{\log 2}{\pi}. \tag{57}$$

Proof. The same arguments leading to (32) in the proof of Theorem 2 can be used, where the signs of the ratios x_{j1}/x_{j2} must be taken into account. If p is the probability of the x_j variables being positive, the probability of the ratio of two variables being positive will be given by $p^2 + (1-p)^2$, i.e., the probability of getting either two positive signs or two negative signs. The probability of the ratio being negative is then $2p(1-p)$. With this consideration of signs, the intermediate form of (32) is modified to take the form

$$\Delta\gamma + \frac{\log 2}{\pi} \approx \frac{1}{N\pi} \sum_{j=1}^{N_P} \log\left[1 + \frac{|x_{j1}/x_{j2}|\Sigma_{B(N)} + |x_{j2}/x_{j1}|\Sigma_{T(N)}}{\Sigma_{B(N)} + \Sigma_{T(N)}}\right] + \frac{1}{N\pi} \sum_{j=1}^{N_Q} \log\left[1 - \frac{|x_{j1}/x_{j2}|\Sigma_{B(N)} + |x_{j2}/x_{j1}|\Sigma_{T(N)}}{\Sigma_{B(N)} + \Sigma_{T(N)}}\right], \tag{58}$$

where N_P is the number of terms where the ratios have positive signs and N_Q is the number of terms where the ratios have negative signs. In the limit $N \to \infty$, we argue (as before) that the sums $\Sigma_{B(N)}$ and $\Sigma_{T(N)}$ approach the same value. Notice also that the two sums can be either positive or negative, but they will both have the same sign (by construction). As a result, we can divide the sums out of the expression as before. In the limit $N \to \infty$, the fraction $N_P/N \to p^2 + (1-p)^2$ and the fraction $N_Q/N \to 2p(1-p)$. After some rearrangement, we obtain the form of (57). □

COROLLARY 4.1. *Let $P(\xi)$ denote the probability distribution of the composite variable $\xi = x_k/x_j$, and assume that the integral $\int d\xi (dP/d\xi) \log|\xi|$ exists. Then for Hill's equation in the unstable limit, and for the case of the variables x_k having mixed signs, in the limit of small variance the correction to the growth rate $\Delta\gamma$ approaches the following limiting form:*

$$\lim_{\sigma_0 \to 0} \Delta\gamma = \frac{2p(1-p)}{\pi} \left[\log \sigma_0 + C_0 - \log 2 \right], \tag{59}$$

where C_0 is a constant that depends on the probability distribution of the variables x_k.

Proof. In the limit of small σ_0, the variables x_k can be written in the form $x_k = 1 + \delta_k$, where $|\delta_k| \ll 1$. To leading order, the expression of (57) for the growth rate becomes

$$\begin{aligned} \Delta\gamma + \frac{\log 2}{\pi} = \lim_{N \to \infty} \frac{1}{\pi N} \Bigg\{ & [p^2 + (p-1)^2] \sum_{j=1}^{N} \log(2 + \delta_{j1} - \delta_{j2}) \\ & + 2p(1-p) \sum_{k=1}^{N} \log \left| \delta_{k1} - \delta_{k2} \right| \Bigg\}. \end{aligned} \tag{60}$$

In the limit of small variance $\sigma_0 \to 0$, the variables $\delta_k \to 0$, and the above expression reduces to the form

$$\Delta\gamma = \frac{2p(1-p)}{\pi} \left[\langle \log|\delta_{k1} - \delta_{k2}| \rangle - \log 2 \right]. \tag{61}$$

We thus need to evaluate the expectation value given by

$$\langle \log|\delta_k - \delta_j| \rangle = \int d\xi \log|\xi| \frac{dP}{d\xi}, \tag{62}$$

where we have defined the composite variable $\xi = \delta_k - \delta_j$. Notice that in the limit $|\delta| \ll 1$, the variance of ξ is σ_0^2. Next we define a dimensionless variable $z \equiv \xi/\sigma_0$ so that the integral becomes

$$I = \int dz \frac{dP}{dz} \log(\sigma_0 z) = \log \sigma_0 \int dz \frac{dP}{dz} + \int dz \frac{dP}{dz} \log z = \log \sigma_0 + \int dz \frac{dP}{dz} \log z. \tag{63}$$

As long as the differential probability distribution dP/dz allows the integral in the final expression to converge, then $I = \log \sigma_0 + C_0$, where C_0 is some fixed number that depends only on the shape of the probability distribution. This convergence requirement is given by the statement of the corollary, so that Corollary 4.1 holds. Notice also that in the limit of small σ_0, the $\log \sigma_0$ term dominates for any fixed C_0 so that $\Delta\gamma \sim 2p(1-p)(\log \sigma_0)/\pi$. □

Figure 2 shows the growth rates as a function of the variance σ_0 for the case of mixed signs. For the case of positive signs only, $p = 1$, the correction $\Delta\gamma$ to the growth rate goes to zero as $\sigma_0 \to 0$. For the case of mixed signs, the correction to the growth rate has the form $\Delta\gamma \propto \log \sigma_0$ as implied by Corollary 4.1.

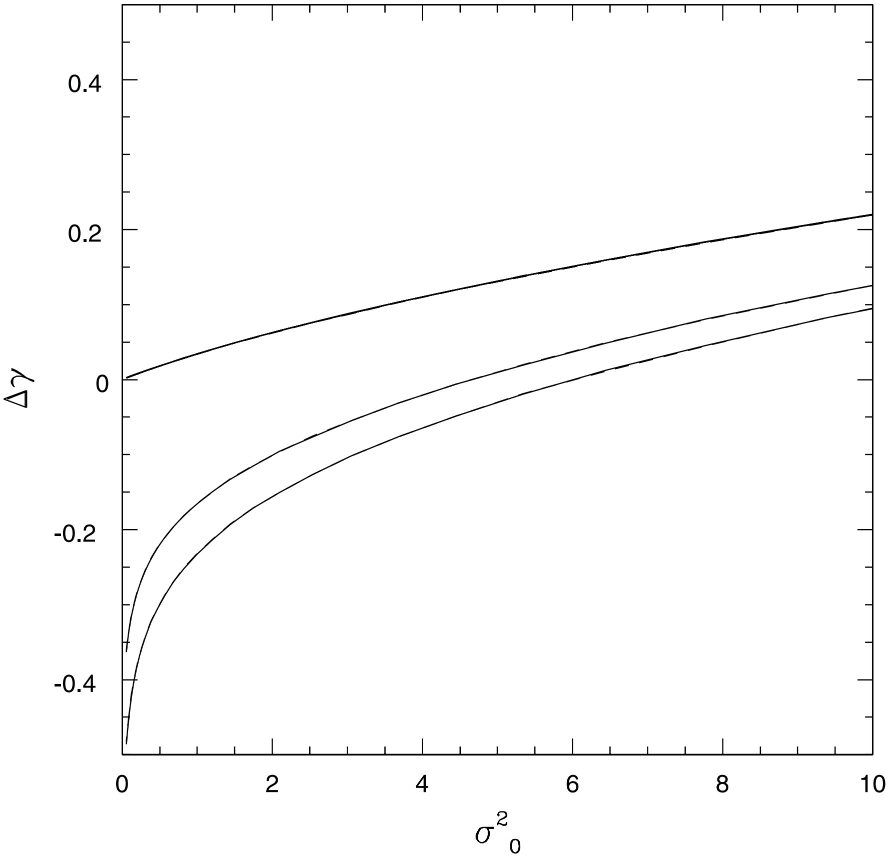


FIG. 2. *Correction $\Delta\gamma$ to the growth rate for the case in which the signs of the random variables x_k are both positive and negative. The three curves show the results for a* 50/50 *distribution (bottom),* 75/25 *(center), and the case of all positive signs (top). For all three cases, the solid curves show the results of numerical matrix multiplication, where* 1000 *realizations of each product are averaged. The overlying dashed curves, which are virtually indistinguishable, show the exact results from Theorem* 4.

Sometimes it is useful to explicitly denote when the growth rates under consideration are the result of purely positive signs or mixed signs for the variables x_k. Here, we use the notation $\Delta\gamma_p$ to specify the growth rate when all the signs are positive. Similarly, $\Delta\gamma_q$ denotes growth rates for the case of mixed signs.

COROLLARY 4.2. *In the limit of large variance, $\sigma_0 \to \infty$, the growth rates for the case of positive signs only and for the case of mixed signs converge, i.e.,*

$$\lim_{\sigma_0\to\infty} \Delta\gamma_q = \Delta\gamma_p, \tag{64}$$

where $\Delta\gamma_p$ denotes the case of all positive signs and $\Delta\gamma_q$ denotes the case of mixed signs.

Proof. The difference in the growth rates for two cases is given by

$$\Delta\gamma_p - \Delta\gamma_q = \frac{2p(1-p)}{\pi} \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \Big[\log(1 + |x_{j1}/x_{j2}|) - \log \Big| 1 - |x_{j1}/x_{j2}| \Big| \Big], \tag{65}$$

where p is the probability for the sign of x_k being positive. In the limit of large variance $\sigma_0^2 \to \infty$, the ratios $|x_j/x_k|$ are almost always far from unity. Only the cases with $|x_j/x_k| \gg 1$ have a significant contribution to the sums. For those cases, however, both of the logarithms in the sums reduce to the same form, $\log|x_j/x_k|$, and hence (65) becomes

$$\lim_{\sigma_0\to\infty} \Delta\gamma_p - \Delta\gamma_q = \frac{2p(1-p)}{\pi}\left[\langle\log|x_j/x_k|\rangle - \langle\log|x_j/x_k|\rangle\right] \to 0. \tag{66}$$

As a result, (64) is valid. □

COROLLARY 4.3. *In the limit of large variance $\sigma_0 \to \infty$, the growth rate $\Delta\gamma$ approaches the form given by*

$$\lim_{\sigma_0\to\infty} \Delta\gamma = \frac{\sigma_0}{\pi} C_\infty, \tag{67}$$

where C_∞ is a constant that depends on the form of the probability distribution for the variables x_k. In general, $C_\infty \le 1/2$.

Proof. Let the composite variable $\xi = \log(x_k/x_j)$ have a probability distribution $dP/d\xi$. Since the growth rate for the case of mixed signs converges to that for all positive signs in the limit of interest (from Corollary 4.2), we need only to consider the latter case (from Theorem 2). The growth rate is then given by the expectation value

$$\Delta\gamma = \frac{1}{\pi}\int_{-\infty}^{\infty} d\xi \frac{dP}{d\xi}\log(1+\mathrm{e}^{\xi}). \tag{68}$$

The integral can be separated into the domains $\xi < 0$ and $\xi > 0$. For the positive integral, we expand the integrand into two terms; for the negative domain, we change the variables of integration so that $\xi \to -\xi$. We thus obtain the three terms

$$\Delta\gamma = \frac{1}{\pi}\int_0^\infty d\xi\frac{dP}{d\xi}\xi + \frac{1}{\pi}\int_0^\infty d\xi\frac{dP}{d\xi}\log(1+\mathrm{e}^{-\xi}) + \frac{1}{\pi}\int_0^\infty d\xi\frac{d\widetilde{P}}{d\xi}\log(1+\mathrm{e}^{-\xi}). \tag{69}$$

In the third integral, the probability distribution $(d\widetilde{P}/d\xi)(\xi) = (dP/d\xi)(-\xi)$; the second and third terms will thus be the same since the distribution is symmetric (by construction, the composite variable ξ is the difference between two variables $\log x_k$ drawn from the same distribution). The sum of the second two integrals is bounded from above by $\log 2$ and can be neglected in the limit of interest. In the first integral, we change variables according to $z = \xi/\sigma$, so that

$$\Delta\gamma \to \frac{\sigma_0}{\pi}\langle z\rangle_{(\xi\ge0)}, \qquad \text{where} \qquad \langle z\rangle_{(\xi\ge0)} \equiv \int_0^\infty dz\frac{dP}{dz}z. \tag{70}$$

Since $\langle 1\rangle = 1$ and $\langle z^2\rangle = 1$, by definition, we expect the quantity $\langle z\rangle_{(\xi\ge0)} = C_\infty$ to be of order unity. Further, one can show that C_∞ as defined here is bounded from above by $1/2$. As a result, in this limit, we obtain a bound of the form $\pi(\Delta\gamma) \le \sigma_0/2 + \log 2$. We note that the constant C_∞ cannot be bounded from below (in the absence of further constraints placed on the probability distribution $dP/d\xi$). □

COROLLARY 4.4. *In the limit of large variance $\sigma_0^2 \gg 1$, the difference $\Delta(\Delta\gamma)$ between the growth rate for strictly positive signs and that for mixed signs takes the form*

$$\lim_{\sigma_0\to\infty} \Delta(\Delta\gamma) = \frac{8p(1-p)}{\pi\sigma_0} C_\Delta, \tag{71}$$

where C_Δ is a constant that depends on the form of probability distribution, and where p is the probability of positive matrix elements for the case of mixed signs.

Proof. Using the results from Theorems 2 and 4 to specify the growth rates for the cases of positive signs and mixed signs, respectively, the difference can be written in the form

$$\Delta(\Delta\gamma) = \frac{2p(1-p)}{\pi}\int_{-\infty}^{\infty} d\xi \, \frac{dP}{d\xi}\left[\log(1+\mathrm{e}^{\xi}) - \log\left|1-\mathrm{e}^{\xi}\right|\right]. \tag{72}$$

Next we separate the integrals into positive and negative domains and change the integration variable for the negative domain ($\xi \to -\xi$). The integral (I) then becomes

$$I = \int_0^{\infty} d\xi \frac{dP}{d\xi}\log\left(\frac{1+\mathrm{e}^{-\xi}}{1-\mathrm{e}^{-\xi}}\right) + \int_0^{\infty} d\xi \frac{d\widetilde{P}}{d\xi}\left(\frac{1+\mathrm{e}^{-\xi}}{1-\mathrm{e}^{-\xi}}\right), \tag{73}$$

where $\widetilde{P}(\xi) = P(-\xi)$. Since we are working in the large σ_0 limit, the variable ξ will be large over most of the domain where the integrals have support, so we can expand, using $\mathrm{e}^{-\xi}$ as a small parameter. In this case, the integral I becomes

$$I = 2\int_{-\infty}^{\infty} d\xi \, \frac{dP}{d\xi}\mathrm{e}^{-|\xi|} \; = \; 2\int_{-\infty}^{\infty} dz \frac{dP}{dz}\mathrm{e}^{-\sigma_0|z|}, \tag{74}$$

where we have made the substitution $z = \xi/\sigma$. For large σ_0, the decaying exponential dominates the behavior of the integrand. In the limit $\sigma \to \infty$, the exponential term decays to zero before the probability dP/dz changes so that $dP/dz \to C_\Delta$, where C_Δ is a constant. The integral thus becomes $I = 4C_\Delta/\sigma_0$, and the difference between the growth rates becomes

$$\Delta(\Delta\gamma) = \frac{8p(1-p)}{\pi\sigma_0}C_\Delta, \tag{75}$$

as claimed by Corollary 4.4. □

Figure 3 illustrates the behavior implied by the last three corollaries. In the limit of large variance, the growth rates for mixed signs and positive signs converge only (Corollary 4.2). Further, growth rates for both cases approach the form $\Delta\gamma \propto \sigma_0$ (as in Corollary 4.3). Finally, the difference between the growth rates for the two cases has the characteristic form $\Delta(\Delta\gamma) \propto 1/\sigma_0$ (from Corollary 4.4).

Corollary 4.5. *For the case of mixed signs, the crossover point between growing solutions and decaying solutions is given by the condition*

$$[p^2 + (1-p)^2]\langle\log\left|1 + |x_j/x_k|\right|\rangle + 2p(1-p)\langle\log\left|1 - |x_j/x_k|\right|\rangle \; = \log 2. \tag{76}$$

Proof. This result follows from Theorem 4 by inspection. □

Estimate for the crossover condition. Equation (76) is difficult to evaluate in practice. In order to obtain a rough estimate of the threshold for instability, we can consider the r_j to be independent variables and use elementary methods to estimate the conditions necessary for systems with mixed signs to be unstable. We first note that the sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ add up the composite variables r_j, which are made up of the variables x_j (which in turn are set by the form of the original differential equation). If the signs of the variables x_j are symmetrically distributed, then the signs of the composite variables r_j are also symmetrically distributed. We can thus focus on the variables r_j.

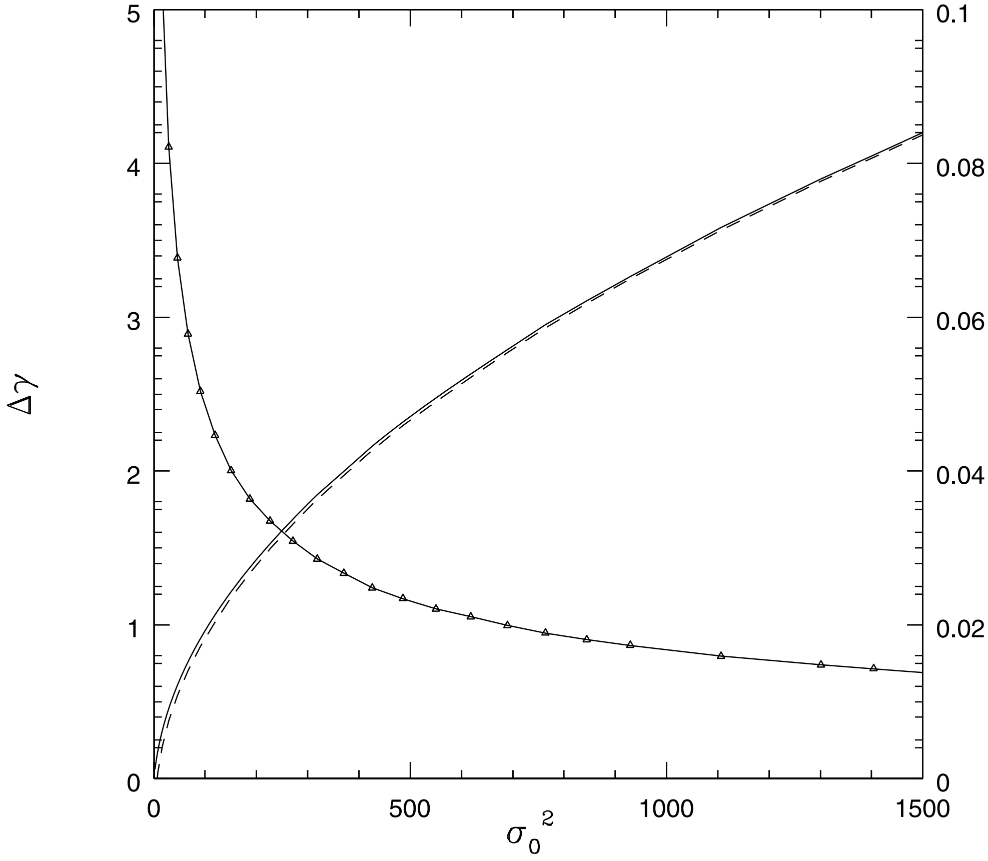


FIG. 3. *Convergence of growth rates in the limit of large variance. The increasing solid curve shows the growth rate as a function of variance for the case of all positive signs. The dashed curve shows the growth rate for the cased of mixed signs with a* 50/50 *sign distribution, i.e.,* $p = 1/2$. *The decreasing curve marked by triangles shows the difference between the two curves (where the axis on the right applies).*

Since the signs can be either positive or negative, the probability of a net excess of positive (or negative) terms is governed by the binomial distribution (which has a Gaussian form in the limit of large N). The probability P of having a net excess of m signs is given by the distribution

$$P(m) = (\pi N_S/2)^{-1/2} \exp\big[-m^2/2N_S\big], \tag{77}$$

where N_S is the number of steps in the random walk. The sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ have $N_S = 2^N$ steps, where N is the number of cycles of the Hill's equation.

If the net excess of signs of one type is m, the sums are reduced (from those obtained with purely positive variables) so that

$$\widetilde{S} = \widetilde{S}_0 \frac{m}{N_S}, \tag{78}$$

where $\widetilde{S}_0$ is the value of the composite sum obtained when the variables x_j have only one sign.

The probability of a growing solution is given by

$$P_G = \int_{m_*}^{\infty} P(m) dm, \tag{79}$$

where m_* is the minimum number of steps needed for instability. We can write m_* in the form

$$m_* = N_S \mathrm{e}^{-N\pi\Delta\gamma_0} = \exp\bigl[N(\log(2) - \pi\Delta\gamma_0)\bigr], \tag{80}$$

where $\Delta\gamma_0$ is the correction to the growth rate for the case of positive signs only.

The integral can be written in terms of the variable $\xi = m/(2N_S)^{1/2}$ so that

$$P_G = \frac{2}{\sqrt{\pi}} \int_{z_*}^{\infty} \mathrm{e}^{-z^2} dz, \tag{81}$$

where

$$z_* = \exp\left[N\left(\frac{1}{2}\log 2 - \pi\Delta\gamma_0\right)\right]. \tag{82}$$

Thus, the crossover for growth occurs under the condition

$$\Delta\gamma_0 \approx \log 2/(2\pi)\ . \tag{83}$$

Keep in mind that this result was derived under the assumption that the variables in the random walk are completely independent. We can derive the above approximate result from a simpler argument: The sums $\Sigma_{T(N)}$ and $\Sigma_{B(N)}$ random walk away from zero according to $\ell\sqrt{N_S} = \langle r_j^2\rangle^{1/2} 2^{N/2} = \exp[n\sigma_0^2 + (N/2)\log 2]$. As a result, $\widetilde{S} \approx \exp[n\sigma_0^2 - (N/2)\log 2]$ and hence $\Delta\gamma \approx (n/N)(\sigma_0^2/\pi) - (\log 2)/2\pi$.

3.6. Specific results for a normal distribution. In this section we consider the particular case where the composite variable $\xi = \log(x_k/x_j)$ has a normal distribution. Specifically, we let the differential probability distribution take the form

$$\frac{dP}{d\xi} = \frac{1}{\sqrt{2\pi}\sigma_0} \mathrm{e}^{-\xi^2/2\sigma_0^2}, \tag{84}$$

so that σ_0^2 is the variance of the distribution. In order to determine the growth rates, we must evaluate the integrals

$$J_{\pm} = \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} d\xi\, \mathrm{e}^{-\xi^2/2\sigma_0^2} \log\left|1 \pm \mathrm{e}^{\xi}\right|. \tag{85}$$

In the limit $\sigma_0 \to 0$, the correction part of the growth rate ($\Delta\gamma$) can be evaluated and has the form

$$\lim_{\sigma_0 \to 0} \Delta\gamma = \frac{1}{\pi}\left\{\left[p^2 + (1-p)^2\right]\frac{\sigma_0^2}{8} + 2p(1-p)\left[\log\sigma_0 - \frac{\gamma_{\mathrm{em}}}{2}\right] - 3p(1-p)\log 2\right\}, \tag{86}$$

where $\gamma_{\mathrm{em}} = 0.577215665\ldots$ is the Euler–Mascheroni constant. Note that for the case of positive signs only ($p = 1$), this expression reduces to the form $\Delta\gamma = \sigma_0^2/(8\pi)$ as in Corollary 2.1. For the case of mixed signs, this expression reduces to the form $\Delta\gamma \propto \log\sigma_0$ from Corollary 4.1.

We can also evaluate the growth rate in the limit of large σ_0, and find the asymptotic form

$$\lim_{\sigma_0 \to \infty} \Delta\gamma = \frac{\sigma_0}{\sqrt{2}\pi^{3/2}}. \tag{87}$$

As a result, the constant C_∞ from Corollary 4.3 is given by $C_\infty = 1/\sqrt{2\pi}$. Note that in this limit, the growth rate is independent of the probabilities p and $(1-p)$ for the variables x_k to have positive and negative signs, consistent with Corollary 4.2. In this limit, we can also evaluate the difference between the cases of positive signs and mixed signs, i.e.,

$$\Delta\gamma_p - \Delta\gamma_q = \frac{8p(1-p)}{\sqrt{2}\pi^{3/2}\sigma_0}. \tag{88}$$

Thus, the constant C_Δ from Corollary 4.4 is given by $C_\Delta = 1/\sqrt{2\pi}$ for the case of a normal distribution. Note that although $C_\Delta = C_\infty$ for this particular example, these constants will not be the same in general.

Finally, for the case of purely positive signs, we can connect the limiting forms for small variance and large variance to construct a rough approximation for the whole range of σ_0, i.e.,

$$\Delta\gamma \approx \frac{\sigma_0^2/\pi}{8 + \sqrt{2\pi}\sigma_0}. \tag{89}$$

This simple expression, which is exact in the limits $\sigma_0 \to 0$ and $\sigma_0 \to \infty$, has a maximum error of about 18% over the entire range of σ_0.

3.7. Matrix decomposition for small variance. For completeness, and as a consistency check, we can study the growth rates by breaking the transformation matrix into separate parts. In this section we consider the case of small variance (see Appendix B for an alternate, more general, separation). In the limit of small variance, $\sigma_0^2 \ll 1$, the variables x_k have only small departures from unity and can be written in the form

$$x_k = 1 + \delta_k, \tag{90}$$

where $|\delta_k| \ll 1$. The matrices of the discrete map can then be decomposed into two parts so that

$$\mathbf{C}_k = \mathbf{A}_k + s_k\delta_k\mathbf{B}_k, \tag{91}$$

where $s_k = \pm 1$ is the sign of the kth term, and where

$$\mathbf{A}_k = \begin{bmatrix} 1 & s_k \\ s_k & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{B}_k = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{92}$$

The matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ have simple multiplicative properties. In particular,

$$\mathbf{A}_j\mathbf{A}_k = 2\mathbf{A}_j \quad \text{if } s_j = s_k\,, \qquad \text{but} \qquad \mathbf{A}_j\mathbf{A}_k = 0 \quad \text{if } s_j \neq s_k, \tag{93}$$

and

$$\mathbf{B}_k^2 = -\mathbf{I}\,, \qquad \mathbf{B}_k^3 = -\mathbf{B}_k\,, \qquad \text{and} \qquad \mathbf{B}_k^4 = \mathbf{I}. \tag{94}$$

The product matrix $\prod \mathbf{C}_k$ will contain long strings of matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ multiplied by each other. If any two matrices $\mathbf{A}_k$ have opposite signs in such a multiplication string, then the product of the two matrices will be zero and the entire string will vanish. As a result, after a large number N of cycles, the only matrices that are guaranteed to survive in the product are those with only $\mathbf{B}_k$ factors and those with only one $\mathbf{A}_k$ factor. Although it is possible for strings with larger numbers of $\mathbf{A}_k$ to survive, it becomes increasingly unlikely (exponentially) as the number of factors increases. To a good approximation, the eigenvalue of the resulting product matrix will be given by the product

$$\Lambda^{(N)} \approx \prod_{k=1}^{N} \delta_k . \tag{95}$$

We could correct for the possibility of longer surviving strings of $\mathbf{A}_k$ by multiplying by a factor of order unity; however, such a factor would have a vanishing contribution to the growth rate. The corresponding growth rate thus takes the form

$$\Delta\gamma = \lim_{N\to\infty} \frac{2p(1-p)}{N\pi} \sum_{k=1}^{N} \log|\delta_k| , \tag{96}$$

where the factor $2p(1-p)$ arises because the matrices with all positive signs lead to a zero growth rate in the limit $\sigma_0 \to 0$, so only the fraction of the cases with mixed signs contribute. Next we note that the sum converges to an expectation value

$$\langle|\delta_k|\rangle = \int d\delta \frac{dP}{d\delta} \log|\delta| . \tag{97}$$

Next we make the substitution $z = \delta/\sigma_0$ and rewrite the integral in the form

$$\langle|\delta_k|\rangle = \sigma_0 \int dz \frac{dP}{dz} + \int dz \frac{dP}{dz} \log z . \tag{98}$$

In the limit of interest, $\sigma_0 \to 0$, the first term dominates and the growth rate (to leading order) approaches the form

$$\Delta\gamma = \frac{2p(1-p)}{\pi} \log\sigma_0 . \tag{99}$$

This form agrees with the leading order expression found earlier in Corollary 4.1 (see also Figure 2, which shows the growth rate as a function of the variance).

4. Hill's equation in the delta function limit. In many physical applications, including the astrophysical orbit problem that motivated this analysis, we can consider the forcing potential to be sufficiently sharp so that $\hat{Q}(t)$ can be considered as a Dirac delta function. For this limit, we specify the main equation considered in this section.

Definition. Hill's equation in the delta function limit *is defined to have the form*

$$\frac{d^2y}{dt^2} + [\lambda + q\delta([t] - \pi/2)]y = 0 , \tag{100}$$

where q measures the strength of the forcing potential and where $\delta(t)$ is the Dirac delta function. In this form, the time variable is scaled so that the period of one cycle is

π. The argument of the delta function is written in terms of $[t]$, which corresponds to the time variable mod-π, so that the forcing potential is π-periodic.

This form of Hill's equation allows for analytic solutions, as outlined below, which can be used to further elucidate the instability for random Hill's equations. In particular, in this case, we can solve for the transformation between the variables (λ_k, q_k) that appear in Hill's equation and the derived composite variables x_k that determine the growth rates.

4.1. Principal solutions. To start the analysis, we first construct the principal solutions to (100) for a particular cycle with given values of forcing strength q and oscillation parameter λ. The equation has two linearly independent solutions $y_1(t)$ and $y_2(t)$, which are defined through their initial conditions

$$y_1(0) = 1, \quad \frac{dy_1}{dt}(0) = 0, \qquad \text{and} \qquad y_2(0) = 0, \quad \frac{dy_2}{dt}(0) = 1. \tag{101}$$

The first solution y_1 has the generic form

$$y_1(t) = \cos\sqrt{\lambda}t \qquad \text{for} \quad 0 \le t < \pi/2, \tag{102}$$

and

$$y_1(t) = A\cos\sqrt{\lambda}t + B\sin\sqrt{\lambda}t \qquad \text{for} \quad \pi/2 < t \le \pi, \tag{103}$$

where A and B are constants that are determined by matching the solutions across the delta function at $t = \pi/2$. We define $\theta \equiv \sqrt{\lambda}\pi/2$ and find

$$A = 1 + (q/\sqrt{\lambda})\sin\theta\cos\theta \qquad \text{and} \qquad B = -(q/\sqrt{\lambda})\cos^2\theta. \tag{104}$$

Similarly, the second solution y_2 has the form

$$y_2(t) = \sin\sqrt{\lambda}t \qquad \text{for} \quad 0 < t < \pi/2, \tag{105}$$

and

$$y_2(t) = C\cos\sqrt{\lambda}t + D\sin\sqrt{\lambda}t \qquad \text{for} \quad \pi/2 < t \le \pi, \tag{106}$$

where

$$C = (q/\lambda)\sin^2\theta \qquad \text{and} \qquad D = \frac{1}{\sqrt{\lambda}} - (q/\lambda)\sin\theta\cos\theta. \tag{107}$$

For the case of constant parameters (q, λ), we can find the criterion for instability and the growth rate for unstable solutions. Since the forcing potential is symmetric, $y_1(\pi) = dy_2/dt(\pi)$, from Theorem 1.1 of [MW]. The resulting criterion for instability reduces to the form

$$H \equiv \left| \frac{q}{2\sqrt{\lambda}}\sin(\sqrt{\lambda}\pi) - \cos(\sqrt{\lambda}\pi) \right| > 1, \tag{108}$$

and the growth rate γ is given by

$$\gamma = \frac{1}{\pi}\log[H + \sqrt{H^2 - 1}]. \tag{109}$$

In the delta function limit, the solution to Hill's equation is thus specified by two parameters—the frequency parameter λ and the forcing strength q. Figure 4 shows the plane of possible parameter space for Hill's equation in this limit, with the unstable regions shaded. Note that a large fraction of the plane is unstable.

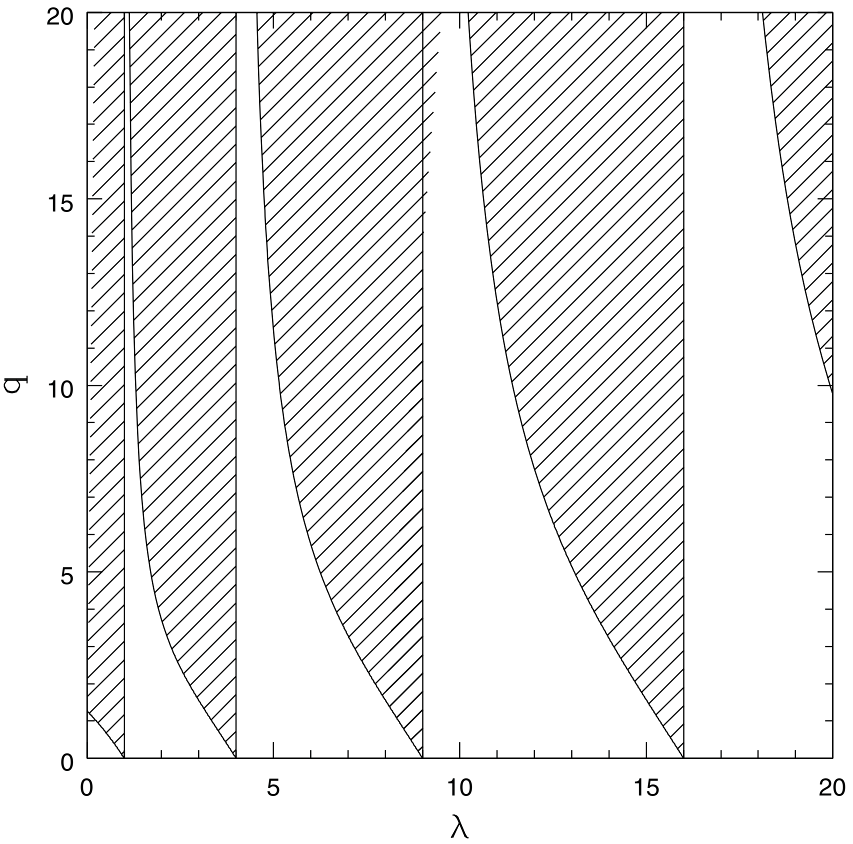


FIG. 4. *Regions of instability for Hill's equation in the delta function limit. The shaded regions show the values of (λ, q) that correspond to exponentially growing (unstable) solutions, which represent unstable growth of the perpendicular coordinate for orbits in our triaxial potential that are initially confined to one of the principal planes.*

4.2. Random variations in the forcing strength. We now generalize to the case where the forcing strength q varies from cycle to cycle, but the oscillation parameter λ is fixed. This version of the problem describes orbits in triaxial, extended mass distributions [AB] and is thus of interest in astrophysics. As outlined in section 2.2, the solutions from cycle to cycle are connected by the transformation matrix given by (7). Here, the matrix elements are given by

$$h = \cos(\sqrt{\lambda}\pi) - \frac{q}{2\sqrt{\lambda}}\sin(\sqrt{\lambda}\pi) \qquad \text{and} \qquad g = -\sqrt{\lambda}\sin(\sqrt{\lambda}\pi) - q\cos^2(\sqrt{\lambda}\pi/2). \tag{110}$$

THEOREM 5. *Consider a random Hill's equation in the delta function limit. For the case of fixed λ, the growth rate of instability approaches the asymptotic growth rate γ_∞ in the highly unstable limit $q/\sqrt{\lambda} \gg 1$, where the correction term has the following order:*

$$\gamma \to \gamma_\infty \left\{ 1 + \mathcal{O}\Big(\lambda/q^2\Big) \right\}. \tag{111}$$

Corollary 5.1. *In the delta function limit, the random Hill's equation with fixed λ is unstable when the asymptotic growth rate $\gamma_\infty > 0$.*

Remark 5.2. *Note that $\gamma_\infty > 0$ requires only that a nonvanishing fraction of the cycles be unstable.*

Proof. For this version of the problem, the matrix $\mathbf{M}$ represents the transition from one cycle to the next, where the solutions are written as linear combinations of y_1 and y_2 for the given cycle. In other words, this transformation operates in the (y_1, y_2) basis of solutions. However, one can also consider the purely growing and decaying solutions, which we denote here as f_+ and f_-.

For a given cycle, the eigenvectors $V_\pm$ of the matrix $\mathbf{M}$ take the form

$$V_\pm = \begin{bmatrix} 1 \\ \pm g/k \end{bmatrix}, \tag{112}$$

where the + (−) sign refers to the growing (decaying) solution. The eigenvalues have the form $\Lambda_\pm = h \pm k$, where $k \equiv (h^2 - 1)^{1/2}$. Keep in mind that $h = y_1(\pi)$ and $g = \dot{y}_2(\pi)$, and that $\Lambda_- = 1/\Lambda_+$. We can write any general solution in the form

$$f = AV_+ + BV_-, \tag{113}$$

where the coefficients (A, B) are related to the coefficients (α, β) in the first basis through the transformation

$$\begin{bmatrix} A \\ B \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & k/g \\ 1 & -k/g \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{114}$$

In the basis of eigenvectors, the action of the differential equation over any cycle is to amplify the growing solution (eigenvector) and attenuate the decaying solution, and this action can be written as the matrix transformation

$$\begin{bmatrix} A' \\ B' \end{bmatrix} = \begin{bmatrix} \Lambda_+ & 0 \\ 0 & \Lambda_- \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}. \tag{115}$$

At the end of the cycle, we can transform back to the original basis through the inverse of the transformation (114). As a result, the original matrix $\mathbf{M}$ can be decomposed into three components so that

$$\mathbf{M}(q, \lambda) = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ g/k & -g/k \end{bmatrix} \begin{bmatrix} \Lambda_+ & 0 \\ 0 & \Lambda_- \end{bmatrix} \begin{bmatrix} 1 & k/g \\ 1 & -k/g \end{bmatrix}. \tag{116}$$

For each cycle, the values of (q, λ) can vary. The next cycle will have a new matrix of the same general form, with the matrix elements specified by (q', λ').

We now shift our view to the basis of eigenvectors, so that each cycle amplifies the growing solution. Between the applications of the amplification factors, the action of successive cycles "rotates" the solution according to a transition matrix of the form

$$\mathbf{T}(q, \lambda; q', \lambda') = \frac{1}{2} \begin{bmatrix} 1 & k'/g' \\ 1 & -k'/g' \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ g/k & -g/k \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 + \mathcal{R} & 1 - \mathcal{R} \\ 1 - \mathcal{R} & 1 + \mathcal{R} \end{bmatrix}, \tag{117}$$

where the primes denote the second cycle and where we have defined $\mathcal{R} \equiv k'g/(kg')$. For the case in which successive cycles have the same values of the original parameters (q, λ), the transition matrix $\mathbf{T}$ becomes the identity matrix (as expected).

For simplicity, we now specialize to the case where λ is held constant from cycle to cycle, but the forcing strength q varies. We can evaluate the transition matrix for the case in which Hill's equation lies in the delta function limit and where we also take the limit $q/\sqrt{\lambda} \gg 1$. In this regime,

$$\mathcal{R} = 1 + \frac{q-q'}{q'}\,\frac{2\sqrt{\lambda}}{q}\,\frac{1-2\cos(\sqrt{\lambda}\pi)}{\sin(\sqrt{\lambda}\pi)} + \mathcal{O}\Big(\frac{\lambda}{q^2}\Big) \equiv 1+2\delta. \tag{118}$$

Note that $\mathcal{R} = 1+2\delta$ to leading order, where δ (defined through the above relation) is small compared to unity and the sign of δ can be both positive and negative. Thus, not only is the parameter δ small, but it can average to zero. Repeated iterations of the mapping lead to the (1,1) matrix element growing according to the product

$$M_{(1,1)} = \prod_{k=1}^{N}\left[\Lambda_k(1+\delta_k)\right] \approx \left[\prod_{k=1}^{N}\Lambda_k\right]\left[1+\sum_{k=1}^{N}\delta_k + \sum_{k=1}^{N}\mathcal{O}(\delta_k^2)\right]. \tag{119}$$

The other matrix elements are of lower order (in powers of $1/q$) so that to leading order the growing eigenvalue of the product matrix is equal to the (1,1) matrix element. Further, for sufficiently well-behaved distributions of the parameter q, the sum of δ_k averages to zero as $N \to \infty$. The growth rate is thus given by

$$\gamma = \frac{1}{\pi N}\sum_{k=1}^{N}\log(\Lambda_k) + \frac{1}{\pi N}\sum_{k=1}^{N}\log(1+\delta_k) = \gamma_\infty + \mathcal{O}\Big(\frac{\lambda}{q^2}\Big). \tag{120}$$

The condition required for the δ_k to average to zero can be expressed in the form

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\frac{q'-q}{qq'} = \Big\langle\frac{1}{q}\Big\rangle - \Big\langle\frac{1}{q'}\Big\rangle = 0, \tag{121}$$

which will hold provided that the expectation value $\langle 1/q\rangle$ exists. This constraint is nontrivial in that a uniform probability distribution $P(q) =$ *constant* that extends to $q = 0$ will produce a divergent expectation value for $\langle 1/q\rangle$. Fortunately, in the physical application that motivated this analysis, the value of q is determined by the distance to the center of an orbit (appropriately weighted) so that the minimum value of q corresponds to the maximum value of the distance. Since physical orbits have a maximum outer turning point (due to conservation of energy), physical orbit problems will satisfy the required constraint on the probability distribution. □

4.3. Second matrix decomposition. Another way to decompose the transformation matrix is to separate it into two separate rotations, one part that is independent of the forcing strength q, and another that is proportional to q. We can thus write the matrix in the form

$$\mathbf{M}(q,\lambda) = \mathbf{A} - \frac{q}{2\sqrt{\lambda}}\mathbf{B} \equiv \begin{bmatrix}\cos 2\theta & (\sin 2\theta)/\sqrt{\lambda}\\ -\sqrt{\lambda}\sin 2\theta & \cos 2\theta\end{bmatrix} - \frac{q}{2\sqrt{\lambda}}\begin{bmatrix}\sin 2\theta & (2\sin^2\theta)/\sqrt{\lambda}\\ 2\sqrt{\lambda}\cos^2\theta & \sin 2\theta\end{bmatrix}, \tag{122}$$

where the second equality defines the matrices $\mathbf{A}$ and $\mathbf{B}$. With these definitions, one finds that

$$\mathbf{A}^N(\theta) = \mathbf{A}(N\theta) \qquad \text{and} \qquad \mathbf{B}^N(\theta) = (2\sin 2\theta)^{N-1}\mathbf{B}(\theta), \tag{123}$$

where we again take λ to be constant from cycle to cycle. As a result, after N cycles, the effective transformation matrix can be written in the form

$$\mathbf{M}^{(N)} = \prod_{k=1}^{N} \Big(\mathbf{A} - \frac{q_k}{2\sqrt{\lambda}} \mathbf{B} \Big). \tag{124}$$

In the asymptotic limit $q/\sqrt{\lambda} \to \infty$, the matrix approaches the form

$$\mathbf{M}^{(N)} = (-1)^N \left[\prod_{k=1}^{N} \frac{q_k}{2\sqrt{\lambda}} \right] (2 \sin 2\theta)^{N-1} \mathbf{B}(\theta). \tag{125}$$

The condition for stability takes the form $|\mathrm{Tr}\mathbf{M}^{(N)}| \geq 2$, i.e.,

$$\left[\prod_{k=1}^{N} q_k \right] \left[\frac{\sin 2\theta}{\sqrt{\lambda}} \right]^N \geq 1. \tag{126}$$

When the system is unstable, the factor on the left-hand side of this equation represents the growth factor over the entire set of N cycles. The growth rate γ is thus given by

$$\gamma = \lim_{N\to\infty} \frac{1}{\pi N} \log \left[\prod_{k=1}^{N} \Big(q_k \frac{\sin 2\theta}{\sqrt{\lambda}} \Big) \right] = \lim_{N\to\infty} \frac{1}{\pi N} \sum_{k=1}^{N} \log \Big(q_k \frac{\sin 2\theta}{\sqrt{\lambda}} \Big). \tag{127}$$

Since $H_k = q_k(\sin 2\theta)/\sqrt{\lambda}$ in this asymptotic limit, the above expression for the growth rate can be rewritten in the form

$$\gamma = \lim_{N\to\infty} \frac{1}{\pi N} \sum_{k=1}^{N} \log(2H_k) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \gamma_k = \gamma_\infty, \tag{128}$$

in agreement with Theorem 5.

4.4. Width of stable and unstable zones. In the plane of parameters (e.g., Figure 4), the width of the stable and unstable zones can be found for the delta function limit. In this case, the leading edge of the zone of stability is given by the condition

$$\theta = \sqrt{\lambda}\pi = n\pi, \tag{129}$$

where n is an integer that can be used to label the zone in question. The beginning of the next unstable zone is given by the condition $|h| = 1$. In the limit of large $q \gg 1$, the width of the stable regime is narrow, and the boundary will fall at $\theta = n\pi + \varphi$, where φ is small. In particular, φ will be smaller than $\pi/2$, so that the angle θ will lie in either the first or the third quadrant, which in turn implies that $\sin\theta$ and $\cos\theta$ have the same sign. As a result, the condition at the boundary takes the form

$$\frac{q}{2\sqrt{\lambda}} = \frac{1+\cos\varphi}{\sin\varphi} \approx \frac{2}{\varphi}. \tag{130}$$

If we solve this expression for φ and use the definition $\varphi = \theta - n\pi$, we can solve for the value of λ at the boundary of the zone, i.e.,

$$\lambda \approx \frac{n^2}{(1-4/q\pi)^2} \approx n^2 \left[1 + \frac{8}{q\pi} + \mathcal{O}(q^{-2}) \right]. \tag{131}$$

The width of the stable zone can then be expressed in the form

$$\Delta\lambda = \frac{8n^2}{\pi q}. \tag{132}$$

For any finite q, there exists a zone number n such that $n^2 > q$ and the width of the zone becomes wide. In the limit $q \to \infty$, the zones are narrow for all finite n.

Note that when the forcing strength q_k varies from cycle to cycle, we can define the expectation value of the zone widths:

$$\langle \Delta\lambda \rangle = \frac{8n^2}{\pi} \left\langle \frac{1}{q_k} \right\rangle. \tag{133}$$

This expectation value exists under the same conditions required for Theorem 5 to be valid.

4.5. Variations in $(\boldsymbol{\lambda_k, q_k})$ and connection to the general case. As outlined earlier, the growth rates $\Delta\gamma$ depend on the ratios of the principal solutions, rather than on the input parameters (λ_k, q_k) that appear in the original differential equation (1). Since we have analytic expressions for the principal solutions in the delta function limit, we can study the relationship between the distributions of the fundamental parameters (λ_k, q_k) and the distribution of the composite variable $\xi = \log(x_k/x_j)$ that appears in the theorems of this paper.

As a starting point, we first consider the limiting case where $q_k \to \infty$ and the parameter λ_k is allowed to vary. We also focus the discussion on the correction $\Delta\gamma$ to the growth rate, which depends on the ratios x_k. In this limit, using (110), we see the variables x_k reduce to the simple form

$$x_k = \frac{\pi}{\theta_k} \frac{\sin\theta_k}{1+\cos\theta_k}, \tag{134}$$

where $\theta_k \equiv \sqrt{\lambda}\pi$. In this case the distribution of $\xi = \log(x_k/x_j)$ depends only on the distribution of the angles θ_k, which is equivalent to the distribution of λ_k. Since the x_j and x_k are drawn independently from the same distribution (of θ_k), the variance of the composite variable $\sigma_0^2 = 2\sigma_x^2$, where σ_x^2 is the variance of $\log x_k$.

As a benchmark case, we consider the distribution of θ to be uniformly distributed over the interval $[0, 2\pi]$. For this example,

$$\sigma_x^2 = \int_0^{2\pi} \frac{d\theta}{2\pi} \left[\log\left(\frac{\pi}{\theta} \frac{\sin\theta}{1+\cos\theta} \right) \right]^2 - \left[\int_0^{2\pi} \frac{d\theta}{2\pi} \log\left(\frac{\pi}{\theta} \frac{\sin\theta}{1+\cos\theta} \right) \right]^2. \tag{135}$$

Numerical evaluation indicates that $\sigma_0 \approx 2.159$. Further, the correction to the growth rate is bounded by $\Delta\gamma \le \sigma_0^2/(4\pi) \approx 0.371$ and is expected to be given approximately by $\Delta\gamma \sim 0.13$. In this limit we expect the asymptotic growth rate to dominate. For example, if $q_k \sim 1000$, a typical value for one class of astrophysical orbits [AB], then $\gamma_\infty \approx 2$, which is an order of magnitude greater than $\Delta\gamma$. Note that in the limit of large (but finite) q_k, the corrections to (134) are of order $\mathcal{O}(1/q_k)$, which will be small, so that the variance σ_0^2 of the composite variable ξ will be nearly independent of the distribution of q_k in this limit.

As another way to illustrate the transformation between the (λ_k, q_k) and the matrix elements x_k, we consider the case of fixed λ_k and large but finite (and varying) values of q_k. We are thus confining the parameter space in Figure 4 to a particular vertical line, which is chosen to be in an unstable band. We thus define $\theta = \sqrt{\lambda}\pi$, and the x_k take the form

$$x_k = \frac{q_k(\pi/\theta)\sin\theta - 2\cos\theta}{q_k(1+\cos\theta)/2 + (\theta/\pi)\sin\theta}. \tag{136}$$

For purposes of illustration, we can make a further simplification by taking θ to have a particular value; for example, if $\theta = \pi/2$, the x_k are given by

$$x_k = \frac{2q_k}{q_k+1}. \tag{137}$$

For this case, the relevant composite variable ξ is given by

$$\xi = \log\left[\frac{q_k}{q_j}\frac{q_j+1}{q_k+1}\right], \tag{138}$$

where q_j and q_k are the values for two successive cycles. In the limit of large $q_j, q_k \gg 1$, the composite variable takes the approximate form $\xi \approx (q_k - q_j)/(q_k q_j)$, which illustrates the relationship between the original variables (only the q_k in this example) and the x_k, or the composite variable ξ, that appear in the growth rates.

Before leaving this section, we note that the more general case of Hill's equation with a square barrier of finite width can also be solved analytically (e.g., let $\hat{Q}(t) = 1/w$ for a finite time interval of width $\Delta t = w$, with $\hat{Q}(t) = 0$ otherwise). For this case, in the limit of large q_k, the solution for h_k takes the form

$$|h_k| \propto \sin(wq_k)^{1/2}\left(\frac{q_k}{w\lambda_k}\right)^{1/2}. \tag{139}$$

In the limit of large but finite q_k and vanishing width $w \to 0$, we recover the result from the delta function limit; i.e., the dependence on the width w drops out and $|h_k| \propto q_k$. In the limit of finite w and large q_k (specifically, when $(wq_k) \ll 1$ does *not* hold); then $|h_k| \propto \sqrt{q_k}$. This example vindicates our expectation that large q_k should lead to large h_k, but the dependence depends on the shape of the barrier $\hat{Q}(t)$. An interesting problem for further study is to place constraints on the behavior of the matrix elements h_k (and g_k) as a function of the forcing strengths q_k for general $\hat{Q}(t)$.

5. Discussion and conclusion.

This paper has considered Hill's equation with forcing strengths and oscillation parameters that vary from cycle to cycle. We denote such cases as random Hill's equations. Our first result is that Hill-like equations where the period is not constant, but rather varies from cycle to cycle, can be reduced to a random Hill's equation (Theorem 1). The rest of the paper thus focuses on random Hill's equations, specifically, general equations in the unstable limit (section 3) and the particular cases of the delta function limit (section 4), where the solutions can be determined in terms of elementary functions.

For a general Hill's equation in the limit of a large forcing parameter, we have found general results governing instability. In all cases, the growth rates depend on the distribution of values for the elements of the transition matrix that maps the solution for one cycle onto the next. The relevant composite variable ξ is determined by the principal solutions via the relation $\xi = \log[y_{1k}(\pi)\dot{y}_{1j}(\pi)/\dot{y}_{1k}(\pi)y_{1j}(\pi)]$, where k and j denote two successive cycles; our results are then presented in terms of the variance σ_0 of the distribution of ξ. The growth rate can be separated into two parts—the asymptotic growth rate γ_∞ that would result if each cycle grew at the rate appropriate for an ordinary Hill's equation, and the correction term $\Delta\gamma$ that results from matching the solutions across cycles. The asymptotic growth rate γ_∞ is determined by the appropriate average of the growth rates for individual cycles (see (9) and (10)). In contrast, the correction term $\Delta\gamma$ results from a type of random walk behavior and depends on the variance of the distribution of the composite variable ξ defined above.

For the case of purely positive matrix elements, the correction term $\Delta\gamma$ has a simple form (Theorem 2) and is positive semidefinite and bounded from above and below. In the limit of small variance, the correction term $\Delta\gamma \propto \sigma_0^2$, whereas in the limit of large variance, $\Delta\gamma \propto \sigma_0$. For all σ_0, the correction term to the growth rate is bounded by $\Delta\gamma \le \sigma_0^2/4\pi$ (Theorem 3). A sharper bound could be obtained in the future.

For the case of matrix elements with varying signs, we have found the growth rate of instability (Theorem 4), where the results depend on the probability p of the matrix elements having a positive sign. In the limit of small variance, the correction term $\Delta\gamma$ is always negative and approaches the form $\Delta\gamma \propto \log\sigma_0$ (unless $p = 1$, where $\Delta\gamma \to 0$ in this limit). As a result, the total growth rate $\gamma = \gamma_\infty + \Delta\gamma$ will always be negative—and hence the system will be stable—for sufficiently small variance σ_0 and any admixture of mixed signs. In the opposite limit of large variance, the growth rate for mixed signs and that for purely positive signs converge, with both cases approaching the form $\Delta\gamma \propto \sigma$; the difference between the growth rates for the two cases decreases as $\Delta(\Delta\gamma) \propto 1/\sigma_0$.

For the delta function limit, we can find the solution explicitly for each cycle and thus analytically define the matrix elements of the discrete map that develops the solution (see (7) and (110)). For the case in which only the forcing strength varies, the growth rate of the general solution approaches the asymptotic growth rate (see (9)), which represents the growth the solution would have if every cycle grows at the rate appropriate for a standard (nonstochastic) Hill's equation. We have calculated the widths of the stable and unstable zones for Hill's equation in the limit of delta function forcing and large growth rates, which represents a specific case of the results presented in [WK], where this specific case includes random forcing terms. Finally, we have used the analytic solutions for the delta function limit to illustrate the transformation between the original variables (λ_k, q_k) that appear in Hill's equation and the variables x_k that determine the growth rates (section 4.5).

Although this paper takes a step forward in our understanding of Hill's equation (in particular, generalizing it to include random forcing terms) and the multiplication of random matrices (of the particular form motivated by Hill's equation), additional work along these lines can be carried out. The analysis presented herein works primarily in the limit of large q_k, where the solutions are highly unstable, although we have bounded the errors incurred by working in this limit. Nonetheless, the case in which some cycles have stable solutions, while others have unstable solutions, should be considered in greater detail. This paper presents bounds on the correction term $\Delta\gamma$ to the growth rate, but a sharper bound could be found. In the treatment of this paper, we considered the probability distribution of the composite variable $\xi = \log(x_k/x_j)$ to be symmetric, which implies that x_k and x_j are independently drawn from their distribution. In future work, correlations between successive cycles can be considered and would lead to asymmetric probability distributions. Most of the results of this paper are presented in terms of the distributions of the composite variables x_k, rather than the original parameters that appear in Hill's equation; the transformation between the distributions of the (λ_k, q_k) and the x_k thus represents another interesting problem for future study. Another case of interest we intend to consider is the case where $\hat{Q}(t)$ takes the form of a finite Fourier series. Finally, the relationship between solutions to random Hill's equations and the multiplication of random matrices should be explored in greater generality.

Random Hill's equations, and the properties of their solutions, have a wide variety of applications. The original motivation for this work was a class of orbit problems in astrophysics. In that context, many astrophysical systems—young embedded star clusters, galactic bulges, and dark matter halos—are essentially triaxial extended mass distributions. Orbits within these mass distributions are often chaotic; further, when motion is initially confined to a plane, the equation of motion for the perpendicular direction is described by a random Hill's equation. The instability explored here thus determines how quickly an orbiting body will explore the perpendicular direction. For example, this class of behavior occurs in young embedded star clusters, which begin in highly flattened configurations but quickly become rounder, in part due to the instability described here. Dark matter halos are found (numerically) to display nearly universal forms for their density distributions [NF, BE], but an a priori explanation for this form remains lacking. Since the orbits of dark matter particles will be subject to the instability studied herein, random Hill's equations must play a role in the explanation. As yet another example, galactic bulges often harbor supermassive black holes at their centers; the resulting stellar orbits, including the instability considered here, play a role in feeding stars into the central black hole. Finally, we note that in addition to astrophysical applications, random Hill's equations are likely to arise in a number of other settings.

Appendix A: Astrophysical motivation. This appendix outlines the original astrophysical problem that motivated this study of Hill's equation with random forcing. In the initial setting, the goal was to understand orbits in potentials resulting from a density profile of the form

$$\rho = \rho_0 \frac{f(m)}{m}, \tag{A1}$$

where ρ_0 is a density scale. This form arises in many different astrophysical contexts, including dark matter halos, galactic bulges, and young embedded star clusters. The density field is constant on ellipsoids and the variable m has a triaxial form

$$m^2 = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}, \tag{A2}$$

where, without loss of generality, $a > b > c > 0$. The radial coordinate ξ is given by $\xi^2 = x^2 + y^2 + z^2$. The function $f(m)$ is assumed to approach unity as $m \to 0$ so that the density profile approaches the form $\rho \sim 1/m$. For this inner limit, one can find an analytic form for both the potential and the force terms [AB]. For purposes of illustration, we write the force terms for the three spatial directions in the form

$$\mathcal{F}_x = -\frac{2x}{F(a)} \ln \left| \frac{2F(a)\sqrt{\Gamma} + 2\Gamma - \Lambda a^2}{a^2\left[2F(a)\xi + \Lambda - 2a^2\xi^2\right]} \right|, \tag{A3}$$

$$\mathcal{F}_y = -\frac{2y}{|F(b)|} \left[\sin^{-1}\left(\frac{\Lambda - 2b^2\xi^2}{\sqrt{\Lambda^2 - 4\xi^2\Gamma}} \right) - \sin^{-1}\left(\frac{2\Gamma/b^2 - \Lambda}{\sqrt{\Lambda^2 - 4\xi^2\Gamma}} \right) \right], \tag{A4}$$

(A5)
$$\mathcal{F}_z = -\frac{2z}{F(c)} \ln \left| \frac{2F(c)\sqrt{\Gamma} + 2\Gamma - \Lambda c^2}{c^2 \left[2F(c)\xi + \Lambda - 2c^2\xi^2 \right]} \right| .$$

The coefficients in the numerators are given by the following quadratic functions of the coordinates:

(A6)
$$\Lambda \equiv (b^2 + c^2)x^2 + (a^2 + c^2)y^2 + (a^2 + b^2)z^2 \qquad \text{and} \qquad \Gamma \equiv b^2c^2x^2 + a^2c^2y^2 + a^2b^2z^2 .$$

The remaining function F is defined by

(A7)
$$F(\alpha) \equiv \left[\xi^2\alpha^4 - \Lambda\alpha^2 + \Gamma \right]^{1/2} .$$

Equations (A3)–(A7) define the force terms that determine the orbital motion of a test particle moving in the potential under consideration (i.e., that resulting from a triaxial density distribution of the form (A1)). The work of [AB] shows that when the orbit begins in any of the three principal planes, the motion is (usually) highly unstable to perturbations in the perpendicular direction. For example, for an orbit initially confined to the $x - z$ plane, the amplitude of the y coordinate will (usually) grow exponentially with time. In the limit of small y, the equation of motion for the perpendicular coordinate simplifies to the form

(A8)
$$\frac{d^2y}{dt^2} + \omega_y^2 y = 0, \qquad \text{where} \qquad \omega_y^2 = \frac{4/b}{\sqrt{c^2x^2 + a^2z^2} + b\sqrt{x^2 + z^2}} .$$

Here, the time evolution of the coordinates (x, z) is determined by the orbit in the original $x - z$ plane. Since the orbital motion is nearly periodic, the (x, z) dependence of ω_y^2 represents a periodic forcing term. The forcing strengths, and hence the parameters q_k appearing in Hill's equation (1), are thus determined by the inner turning points of the orbit (with appropriate weighting from the axis parameters $[a, b, c]$). Further, since the orbit in the initial plane often exhibits chaotic behavior, the distance of closest approach of the orbit, and hence the strength q_k of the forcing, varies from cycle to cycle. The orbit also has outer turning points, which provide a minimum value of ω_y^2, which defines the unforced oscillation frequency λ_k appearing in Hill's equation (1). As a result, the equation of motion (A8) for the perpendicular coordinate has the form of Hill's equation, where the period, the forcing strength, and the oscillation frequency generally vary from cycle to cycle.

Appendix B: Growth rate for an ancillary matrix. In this appendix, we separate the transformation matrix for the general case (not in the limit of small variance) and find the growth rate for one of the matrices. We include this result because examples where one can explicitly find the growth rates (Lyapunov exponents) for random matrices are rare. Specifically, the transition matrix can be written in the form given by (91), where the second term in the sum has the form

(B1)
$$s_k(x_k - 1)\mathbf{B}_k, \qquad \text{where} \qquad \mathbf{B}_k = \begin{bmatrix} 0 & 1 \\ -1/x_k & 0 \end{bmatrix} .$$

Note that any pair of matrices $\mathbf{A}_k$ with opposite signs will vanish, and so will all subsequent products.

The products of the second term (the matrices $\mathbf{B}_k$ along with the leading factor) have a well-defined growth rate.

Proposition 3. *The growth rate of matrix multiplication for the matrix* $\mathbf{M}_k = (x_k - 1)\mathbf{B}_k$ *is given by*

$$\gamma_B = \lim_{N\to\infty} \frac{1}{2\pi N} \left\{ \sum_{k=1}^{N} \log|x_k - 1| + \sum_{j=1}^{N} \log|1/x_j - 1| \right\}. \tag{B2}$$

Proof. The products of the matrices $\mathbf{B}_k$ follow cycles as shown by the first three nontrivial cases:

$$\mathbf{B}_2\mathbf{B}_1 = \begin{bmatrix} -1/x_1 & 0 \\ 0 & -1/x_2 \end{bmatrix}, \qquad \mathbf{B}_3\mathbf{B}_2\mathbf{B}_1 = \begin{bmatrix} 0 & -1/x_2 \\ 1/(x_1x_3) & 0 \end{bmatrix}, \tag{B3}$$

and

$$\mathbf{B}_4\mathbf{B}_3\mathbf{B}_2\mathbf{B}_1 = \begin{bmatrix} 1/(x_1x_3) & 0 \\ 0 & 1/(x_2x_4) \end{bmatrix}. \tag{B4}$$

Thus, the even products of the matrices are diagonal matrices, whereas the odd products produce matrices with only off-diagonal elements. As a result, the product matrix will approach the form

$$\begin{aligned} \mathbf{M}^{(N)} &\sim \left(\prod_{k=1}^{N}(x_k - 1)\right) \begin{bmatrix} P_{\text{odd}} & 0 \\ 0 & P_{\text{even}} \end{bmatrix} \qquad \text{or} \\ \mathbf{M}^{(N)} &\sim \left(\prod_{k=1}^{N}(x_k - 1)\right) \begin{bmatrix} 0 & -P_{\text{even}} \\ P_{\text{odd}} & 0 \end{bmatrix}, \end{aligned} \tag{B5}$$

where we have defined

$$P_{\text{odd}} \equiv \prod_{k=1,odd}^{N} \frac{1}{x_k} \qquad \text{and} \qquad P_{\text{even}} \equiv \prod_{k=2,even}^{N} \frac{1}{x_k}. \tag{B6}$$

For N even (odd), the eigenvalues are $\Lambda = P_{\text{even}}, P_{\text{odd}}$ ($\Lambda = \pm i\sqrt{P_{\text{even}}P_{\text{odd}}}$). Since $|P_{\text{even}}| = |P_{\text{odd}}|$ in the limit $N \to \infty$, the eigenvalues (and hence the growth rates) have the same magnitudes in either case. To compute the growth rate γ_B, we need to account for the fact that only half of the factors (either the even or the odd terms) appear in the products P_{odd} and P_{even}. After some rearrangement, we obtain equation (B2). ◻

Acknowledgments. We would like to thank Charlie Doering, Gus Evrard, Divakar Vishwanath, and Michael Weinstein for many useful conversations, and research students Michael Busha, Suzanne Butler, Jeff Druce, Jake Ketchum, and Eva Proszkow for performing numerical calculations that guided the initial formulation of this project. We also thank an anonymous referee for many useful comments and criticisms that improved the paper.

REFERENCES

[AS] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[AB] F. C. Adams, A. M. Bloch, S. C. Butler, J. M. Druce, and J. A. Ketchum, *Orbits and instabilities in a triaxial cusp potential*, Astrophys. J., 670 (2007), pp. 1027–1047.

[BD] J. Baik, P. Deift, and E. Strahov, *Products and ratios of polynomials of random Hermitian matrices*, J. Math. Phys., 44 (2003), pp. 3657–3670.

[BE] M. T. Busha, A. E. Evrard, F. C. Adams, and R. H. Wechsler, *The ultimate halo mass in a ΛCDM universe*, Monthly Notices R. Astron. Soc., 363 (2005), pp. L11–L15.

[BT] J. Binney and S. Tremaine, *Galactic Dynamics*, Princeton University Press, Princeton, NJ, 1987.

[BL] P. Bougerol and J. Lacroix, *Products of Random Matrices with Applications to Schrödinger Operators*, Birkhäuser Boston, Boston, 1985.

[DE] P. Deift, *Orthogonal Polynomials and Random Matrices, A Riemann-Hilbert Approach*, CIMS Lecture Notes, New York University, New York, NY, 1999.

[FU] H. Furstenberg, *Non-commuting random products*, Trans. Amer. Math. Soc., 108 (1963), pp. 377–428.

[FK] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Statist., 31 (1960), p. 457.

[HI] G. W. Hill, *On the part of the motion of the lunar perigee which is a function of the mean motions of the Sun and Moon*, Acta. Math., 8 (1886), pp. 1–36.

[LR] R. Lima and M. Rahibe, *Exact Lyapunov exponent for infinite products of random matrices*, J. Phys. A, 27 (1994), pp. 3427–3438.

[MW] W. Magnus and S. Winkler, *Hill's Equation*, Wiley, New York, 1966.

[ME] M. Mehta, *Random Matrices*, 2nd ed., Academic Press, Boston, 1991.

[NF] J. F. Navarro, C. S. Frenck, and S. D. M. White, *A universal density profile from hierarchical clustering*, Astrophys. J., 490 (1997), pp. 493–508.

[VI] D. Viswanath, *Random Fibonacci sequences and the number* 1.13198824..., Math. Comp., 69 (2000), pp. 1131–1155.

[WK] M. I. Weinstein and J. B. Keller, *Asymptotic behavior of stability regions for Hill's equation*, SIAM J. Appl. Math., 47 (1987), pp. 941–958.

SIAM J. Appl. Math.
Vol. 68, No. 4, pp. 981–1003

A FLUID DYNAMIC MODEL FOR TELECOMMUNICATION NETWORKS WITH SOURCES AND DESTINATIONS*

CIRO D'APICE[†], ROSANNA MANZO[†], AND BENEDETTO PICCOLI[‡]

Abstract. This paper proposes a macroscopic fluid dynamic model dealing with the flows of information on a telecommunication network with sources and destinations. The model consists of a conservation law for the packet density and a semilinear equation for traffic distribution functions, i.e., functions describing packet paths. We describe methods to solve Riemann problems at junctions assigning different traffic distribution functions and two "routing algorithms." Moreover, we prove the existence of solutions to Cauchy problems for small perturbations of network equilibria.

Key words. data flows on telecommunication networks, sources and destinations, conservation laws, fluid dynamic models

AMS subject classifications. 35L65, 35L67, 90B20

DOI. 10.1137/060674132

1. Introduction.

This paper is concerned with the description and analysis of a macroscopic fluid dynamic model dealing with flows of information on a telecommunication network with sources and destinations. The latter are, respectively, areas from which packets start their travels on the network and areas where they end.

There are various approaches to telecommunication and data networks (see, for example, [1], [3], [4], [15], [21], [22]). A first model for telecommunication networks, similar to that introduced recently for car traffic, has been proposed in [10], where two algorithms for dynamics at nodes were considered and the existence of solutions to Cauchy problems was proved. The idea is to follow the approach used in [12] for road networks (see also [7], [9], [11], [14], [16], [17], [18]), introducing sources and destinations in the telecommunication model described in [10] and thus taking care of the paths of the packets inside the network.

A telecommunication network consists of a finite collection of transmission lines, modelled by closed intervals of $\mathbb{R}$ connected together by nodes (routers, hubs, switches, etc.). We assume that each node receives and sends information encoded in packets, which can be seen as particles travelling on the network. Taking the Internet network as a model, we assume the following:

(1) Each packet travels on the network with a fixed speed and with an assigned final destination.
(2) Nodes receive, process, and then forward packets. Packets may be lost with a probability increasing with the number of packets to be processed. Each lost packet is sent again.

Since each lost packet is sent again until it reaches the next node, looking at the macroscopic level, it is assumed that the number of packets is conserved. This leads

*Received by the editors November 6, 2006; accepted for publication (in revised form) November 5, 2007; published electronically February 6, 2008.
http://www.siam.org/journals/siap/68-4/67413.html

[†]Department of Information Engineering and Applied Mathematics, University of Salerno, Fisciano (SA), Salerno 84084, Italy (dapice@diima.unisa.it, manzo@diima.unisa.it).
[‡]Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Roma 00179, Italy (b.piccoli@iac.cnr.it).

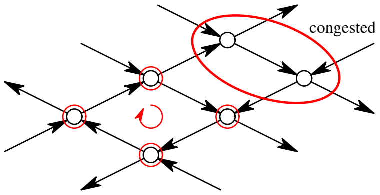


Fig. 1.1. *A possible cycling effect of* (RA2).

to a conservation law for the packet density ρ on each line:

$$\rho_t + f(\rho)_x = 0. \tag{1.1}$$

The flux $f(\rho)$ is given by $v \cdot \rho$, where v is the average speed of packets among nodes, derived considering the amount of packets that may be lost.

Recently, a conservation law model was obtained in [2] for supply chains, which have a dynamics somehow related to our case.

On each transmission line we also consider a vector π describing the traffic types, i.e., the percentage of packets going from a fixed source to a fixed destination. Assuming that packet velocity is independent from the source and the destination, the evolution of π follows a semilinear equation,

$$\pi_t + v(\rho)\pi_x = 0; \tag{1.2}$$

hence inside transmission lines the evolution of π is influenced by the average speed of packets.

The aim is then to consider networks in which many lines intersect. Riemann problems at junctions were solved in [10] proposing two different routing algorithms:

(RA1) Packets from incoming lines are sent to outgoing ones according to their final destination (without taking into account possible high loads of outgoing lines).

(RA2) Packets are sent to outgoing lines in order to maximize the flux through the node.

The main differences of the two algorithms are the following. The first one simply sends each packet to the outgoing line which is naturally chosen according to the final destination of the packet itself. The algorithm is blind to possible overloads of some outgoing lines and, by some abuse of notation, is similar to the behavior of a "switch." The second algorithm, on the contrary, send packets to outgoing lines taking into account the loads and thus possibly redirecting packets. Again by some abuse of notation, this is similar to a "router" behavior.

One of the drawbacks of the second algorithm is that it does not take into account the global path of packets, therefore leading to possible cycling. For example consider a telecommunication network in which some nodes are congested: if we use (RA2) alone, the packets are not routed towards the congested nodes, and so they can enter in loops (see Figure 1.1). These cyclings are avoided if we consider that the packets originated from a source and with an assigned destination have precise paths inside the network. Such paths are determined by the behavior at junctions via the coefficients π.

In this paper different distribution traffic functions describing different routing strategies have been considered:

- at a junction the traffic started at source s and with d as the final destination, coming from the transmission line i, is routed on an assigned line j;

- at a junction the traffic started at source s and with d as the final destination, coming from the transmission line i, is routed on every outgoing line or on some of them.

The first distribution traffic function has already been analyzed in [12] for road networks using algorithm (RA1); thus we focus on the second one. In particular, we define two ways according to which the traffic at a junction is split towards the outgoing lines.

Let us now comment further on the differences with the results of [12]. In that paper, only the routing algorithm (RA1) was considered, together with the first choice of distribution traffic functions (which can be seen as a particular case of the second choice). Since the algorithm (RA1) produces discontinuities in the map from traffic types to fluxes (and densities), a new Riemann solver was introduced, which considers the maximization of a quadratic cost. The latter produces as a drawback more difficulties in analysis and numerics. Finally, the present paper presents a more general approach and, using (RA2), the possibility of solving dynamics at nodes using linear functionals.

Starting from the distribution traffic function, and using the vector π, we assign the traffic distribution matrix, which describes the percentage of packets from an incoming line that are addressed to an outgoing one. Then we propose methods to solve Riemann problems considering the routing algorithms (RA1) and (RA2). The key point to construct a solution on the whole network, using a way-front tracking method, is to derive some BV estimates on the piecewise constant approximate solutions in order to pass to the limit. In the case in which the traffic at junctions is distributed on outgoing lines according to some probabilistic coefficients, estimates on packet density function and on traffic-type functions are derived for the algorithm (RA2) in order to prove the existence of solutions to Cauchy problems. More precisely, we prove the existence of solutions, locally in time, for perturbations of equilibria.

The paper is organized as follows. Section 2 gives a general definition of the network. Then, in section 3, we discuss possible choices of the traffic distribution functions and how to compute the traffic distribution matrix from the latter functions and the traffic-type function. We describe two routing algorithms in section 4, giving explicit unique solutions to Riemann problems. Finally, section 5 provides the needed estimates for constructing solutions to Cauchy problems.

2. Basic definitions.

We consider a telecommunication network that is a finite collection of transmission lines connected together by nodes, some of which are sources and destinations. Formally we introduce the following definition.

DEFINITION 2.1. *A telecommunication network is given by a 7-tuple* $(N, \mathcal{I}, \mathcal{F}, \mathcal{J}, \mathcal{S}, \mathcal{D}, \mathcal{R})$, *where we have the following:*

Cardinality. *N is the cardinality of the network, i.e., the number of lines in the network.*

Lines. *$\mathcal{I}$ is the collection of lines, modelled by intervals $I_i = [a_i, b_i] \subseteq \mathbb{R}$, $i = 1, \dots, N$.*

Fluxes. *$\mathcal{F}$ is the collection of flux functions $f_i : [0, \rho_i^{\max}] \mapsto \mathbb{R}$, $i = 1, \dots, N$.*

Nodes. *$\mathcal{J}$ is a collection of subsets of $\{\pm 1, \dots, \pm N\}$ representing nodes. If $j \in J \in \mathcal{J}$, then the transmission line $I_{|j|}$ is crossing at J as an incoming line (i.e., at point b_i) if $j > 0$ and as an outgoing line (i.e., at point a_i) if $j < 0$. For each junction $J \in \mathcal{J}$, we indicate by $\mathrm{Inc}(J)$ the set of incoming lines, which are I_i's such that $i \in J$, while by $\mathrm{Out}(J)$ the set of outgoing lines, which are I_i's such that $-i \in J$. We assume that each line is incoming for (at most) one node and outgoing for (at most) one node.*

Sources. *$\mathcal{S}$ is the subset of $\{1,\dots,N\}$ representing lines starting from traffic sources. Thus, $j \in \mathcal{S}$ if and only if j is not outgoing for any node. We assume that $\mathcal{S} \neq \emptyset$.*

Destinations. *$\mathcal{D}$ is the subset of $\{1,\dots,N\}$ representing lines leading to traffic destinations. Thus, $j \in \mathcal{D}$ if and only if j is not incoming for any node. We assume that $\mathcal{D} \neq \emptyset$.*

Traffic distribution functions. *$\mathcal{R}$ is a finite collection of functions $r_J : \mathrm{Inc}(J) \times \mathcal{S} \times \mathcal{D} \to \mathrm{Out}(J)$. For every J, $r_J(i,s,d)$ indicates the outgoing direction of traffic that started at source s, has d as the final destination, and reached J from the incoming road i. (We will also consider the case of r_J multivalued.)*

One usually assumes that the network is connected. However, this is not strictly necessary to develop our theory.

2.1. Dynamics on lines. Following [10], we recall the model used to define the dynamics of packet densities along lines. We make the following hypotheses:

(H1) Lines are composed of consecutive processors N_k, which receive and send packets. The number of packets at N_k is indicated by $R_k \in [0, R_{max}]$.

(H2) There are two time scales: Δt_0 represents the physical travel time of a single packet from node to node (assumed to be independent of the node for simplicity); T represents the processing time, during which each processor tries to operate the transmission of a given packet.

(H3) Each processor N_k tries to send all packets R_k at the same time. Packets are lost according to a loss probability function $p : [0, R_{max}] \to [0,1]$, computed at R_{k+1}, and lost packets are sent again for a time slot of length T.

The aim is to determine the fluxes on the network. Since the packet transmission velocity on the line is assumed constant, it is possible to compute an average velocity function and thus an average flux function.

Let us focus on two consecutive nodes N_k and N_{k+1}, assume a static situation, i.e., R_k and R_{k+1} are constant, and call δ the distance between the nodes. During a processing time slot of length T the following happens. All packets R_k are sent a first time: $(1 - p(R_{k+1}))\,R_k$ are sent successfully and $p(R_{k+1})\,R_k$ are lost. At the second attempt, among the lost packets $p(R_{k+1})\,R_k$, $(1-p(R_{k+1})\,p(R_{k+1})\,R_k$ are sent successfully and $p^2(R_{k+1})\,R_k$ are lost, and so on.

Let us indicate by Δt_{av} the average transmission time of packets, by $\bar{v} = \frac{\delta}{\Delta t_0}$ the packet velocity without losses, and by $v = \frac{\delta}{\Delta t_{av}}$ the average packet velocity. Then we can compute

$$\Delta t_{av} = \sum_{n=1}^{M} n \Delta t_0 (1 - p(R_{k+1})) p^{n-1}(R_{k+1}),$$

where $M = [T/\Delta t_0]$ (here $[\cdot]$ indicates the floor function) represents the number of attempts of sending a packet. We make a further assumption:

(H4) The number of packets not transmitted for a whole processing time slot is negligible.

Hypothesis (H4) corresponds to assuming that $\Delta t_0 \ll T$ or, equivalently, $M \sim +\infty$. Making the identification, $M = +\infty$, we get

$$\Delta t_{av} = \sum_{n=1}^{+\infty} n \Delta t_0 (1 - p(R_{k+1})) p^{n-1}(R_{k+1}) = \frac{\Delta t_0}{1 - p(R_{k+1})},$$

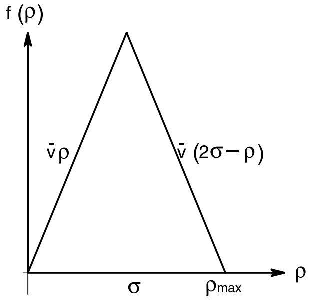


FIG. 2.1. *Example of flux function.*

and

$$v = \frac{\delta}{\Delta t_{av}} = \frac{\delta}{\Delta t_0}(1 - p(R_{k+1})) = \bar{v}(1 - p(R_{k+1})). \tag{2.1}$$

Let us now call ρ the averaged density and ρ_{max} its maximum. We can interpret the function p as a function of ρ and, using (2.1), determine the corresponding flux function, given by the averaged density times the average velocity. It is reasonable to assume that the probability loss function is null for some interval, which is a right neighborhood of zero. This means that at low densities no packet is lost. Then p should be increasing, reaching the value 1 at the maximal density, the situation of being completely stuck. A possible choice of the probability loss function is the following:

$$p\left(\rho\right) = \begin{cases} 0, & 0 \le \rho \le \sigma, \\ \frac{\rho_{max}\left(\rho - \sigma\right)}{\rho\left(\rho_{max} - \sigma\right)}, & \sigma \le \rho \le \rho_{\max}; \end{cases}$$

then it follows that

$$f\left(\rho\right) = \begin{cases} \bar{v}\rho, & 0 \le \rho \le \sigma, \\ \frac{\bar{v}\sigma(\rho_{max} - \rho)}{\rho_{max} - \sigma}, & \sigma \le \rho \le \rho_{\max}. \end{cases} \tag{2.2}$$

Setting, for simplicity, $\rho_{max} = 1$ and $\sigma = \frac{1}{2}$, we get the simple "tent" function of Figure 2.1. To simplify the treatment of the corresponding conservation laws, we will assume the following:

(F) Setting $\rho_{max} = 1$, on each line the flux $f_i : [0, 1] \to R$ is concave, $f(0) = f(1) = 0$, and there exists a unique maximum point $\sigma \in\,]0, 1[$.

Notice that the flux of Figure 2.1 or, more generally, the flux given in (2.2) satisfies assumption (F).

2.2. Dynamics on the network. On each transmission line I_i we consider the evolution equation

$$\partial_t \rho_i + \partial_x f_i\left(\rho_i\right) = 0, \tag{2.3}$$

where we use assumption (F). Therefore, the network load evolution is described by a finite set of functions $\rho_i : [0, +\infty[\, \times I_i \mapsto [0, \rho_i^{\max}]$.

On each transmission line I_i we want ρ_i to be a weak entropic solution of (2.3); that is, for every function $\varphi : [0, +\infty[\times I_i \to \mathbb{R}$ smooth, positive with compact support on $]0, +\infty[\times]a_i, b_i[$,

$$\int_0^{+\infty} \int_{a_i}^{b_i} \left(\rho_i \frac{\partial \varphi}{\partial t} + f_i(\rho_i) \frac{\partial \varphi}{\partial x} \right) dxdt = 0, \tag{2.4}$$

and for every $k \in \mathbb{R}$ and every $\tilde{\varphi} : [0, +\infty[\times I_i \to \mathbb{R}$ smooth, positive with compact support on $]0, +\infty[\times]a_i, b_i[$,

$$\int_0^{+\infty} \int_{a_i}^{b_i} \left(|\rho_i - k| \frac{\partial \tilde{\varphi}}{\partial t} + sgn(\rho_i - k) \left(f_i(\rho_i) - f_i(k) \right) \frac{\partial \tilde{\varphi}}{\partial x} \right) dxdt \geq 0. \tag{2.5}$$

For each $i \in \mathcal{S}$ (resp., $i \in \mathcal{D}$) we need an inflow function (resp., outflow) and thus consider measurable functions $\psi_i : [0, +\infty[\to [0, \rho_i^{\max}]$. Then the corresponding functions ρ_i must verify the boundary condition $\rho_i(t, a_i) = \psi_i(t)$ (resp., $\rho_i(t, b_i) = \psi_i(t)$) in the sense of [5].

Moreover, inside each line I_i we define a traffic-type function π_i, which measures the portion of the whole density coming from each source and travelling towards each destination.

Definition 2.2. *A traffic-type function on a line I_i is a function*

$$\pi_i : [0, \infty[\times [a_i, b_i] \times \mathcal{S} \times \mathcal{D} \mapsto [0, 1]$$

such that, for every $t \in [0, \infty[$ and $x \in [a_i, b_i]$,

$$\sum_{s \in \mathcal{S}, d \in \mathcal{D}} \pi_i(t, x, s, d) = 1.$$

In other words, $\pi_i(t, x, s, d)$ specifies the fraction of the density $\rho_i(t, x)$ that started from source s and is moving towards the final destination d.

We assumed, on the discrete model, that a FIFO policy is used at nodes. Then it is natural that the averaged velocity, obtained in the limit procedure, is independent from the original sources of packets and their final destinations. In other words, we make the following hypothesis:

(H5) On each line I_i, the average velocity of packets depends only on the value of the density ρ_i and not on the values of the traffic-type function π_i.

As a consequence of hypothesis (H5), we have the following. If $x(t)$ denotes a trajectory of a packet inside the line I_i, then we get

$$\pi_i(t, x(t), s, d) = \text{const.} \tag{2.6}$$

In fact, consider the packets that at time t are in position $x(t)$. All such packets have the same velocity by (H5); thus their trajectories coincide, independently of their sources and destinations. In other words, at a time $t' > t$ all packets will be in position $x(t')$. Then the fractions of the density, expressed by π, are the same at $(t, x(t))$ and at $(t', x(t'))$.

Taking the total differential with respect to the time of (2.6), we deduce the semilinear equation

$$\partial_t \pi_i(t, x, s, d) + \partial_x \pi_i(t, x, s, d) \cdot v_i(\rho_i(t, x)) = 0. \tag{2.7}$$

This equation is coupled with (2.3) on each line I_i. More precisely, (2.7) depends on the solution of (2.3), while in turn at junctions the values of π_i will determine the traffic distribution on outgoing lines as explained below.

For simplicity and without loss of generality, we assume from now on that the fluxes f_i are all the same, and we indicate them with f. Thus, the model for a single transmission line consists of the system of equations

$$\begin{cases} \rho_t + f(\rho)_x = 0, \\ \pi_t + \pi_x \cdot v(\rho) = 0. \end{cases}$$

To treat the evolution at a junction, let us introduce some notation. Fix a junction J with n incoming transmission lines, say $I_1, \dots, I_n$, and m outgoing transmission lines, say $I_{n+1}, \dots, I_{n+m}$. A weak solution at J is a collection of functions $\rho_l : [0, +\infty[\times I_l \mapsto \mathbb{R}$, $l = 1, \dots, n+m$, such that

$$\sum_{l=1}^{n+m} \left(\int_0^{+\infty} \int_{a_l}^{b_l} \left(\rho_l \frac{\partial \varphi_l}{\partial t} + f(\rho_l) \frac{\partial \varphi_l}{\partial x} \right) dx dt \right) = 0, \tag{2.8}$$

for every φ_l, $l = 1, \dots, n+m$, smooth having compact support in $]0, +\infty[\times]a_l, b_l]$ for $l = 1, \dots, n$ (incoming transmission lines) and in $]0, +\infty[\times [a_l, b_l[$ for $l = n+1, \dots, n+m$ (outgoing transmission lines), that are also smooth across the junction, i.e.,

$$\varphi_i(\cdot, b_i) = \varphi_j(\cdot, a_j), \quad \frac{\partial \varphi_i}{\partial x}(\cdot, b_i) = \frac{\partial \varphi_j}{\partial x}(\cdot, a_j), \quad i = 1, \dots, n, j = n+1, \dots, n+m.$$

REMARK 2.3. *Let $\rho = (\rho_1, \dots, \rho_{n+m})$ be a weak solution at the junction J such that each $x \to \rho_i(t, x)$ has bounded variation. We can deduce that ρ satisfies the Rankine–Hugoniot condition at J, namely*

$$\sum_{i=1}^{n} f(\rho_i(t, b_i-)) = \sum_{j=n+1}^{n+m} f(\rho_j(t, a_j+)),$$

for almost every $t > 0$.

For a scalar conservation law a Riemann problem is a Cauchy problem for an initial data of Heaviside type, that is, piecewise constant with only one discontinuity. One looks for centered solutions, i.e., $\rho(t, x) = \phi(\frac{x}{t})$ formed by simple waves, which are the building blocks to construct solutions to the Cauchy problem via a wave-front tracking algorithm. These solutions are formed by continuous waves called rarefactions and by travelling discontinuities called shocks. The speed of waves is related to the values of f'; see [6], [8], [19], [20].

Analogously, we call the Riemann problem for a junction the Cauchy problem corresponding to initial data $\rho_{1,0}, \dots, \rho_{n+m,0} \in [0, 1]$, and $\pi_1^{s,d}, \dots, \pi_{n+m}^{s,d} \in [0, 1]$, which are constant on each transmission line.

DEFINITION 2.4. *A Riemann solver (RS) for the junction J is a map that associates with Riemann data $\rho_0 = (\rho_{1,0}, \dots, \rho_{n+m,0})$ and $\Pi_0 = (\pi_{1,0}, \dots, \pi_{n+m,0})$ at J the vectors $\hat{\rho} = (\hat{\rho}_1, \dots, \hat{\rho}_{n+m})$ and $\hat{\Pi} = (\hat{\pi}_1, \dots, \hat{\pi}_{n+m})$ so that the solution on an incoming transmission line I_i, $i = 1, \dots, n$, is given by the wave $(\rho_{i,0}, \hat{\rho}_i)$ and on an outgoing one I_j, $j = n+1, \dots, n+m$, is given by the waves $(\hat{\rho}_j, \rho_{j,0})$ and $(\hat{\pi}_j, \pi_{j,0})$. We require the following consistency condition:*

(CC) $RS(RS(\rho_0, \Pi_0)) = RS(\rho_0, \Pi_0)$.

We will define an RS at a junction in the next sections. Once an RS is defined and the solution of the Riemann problem is obtained, we can define admissible solutions at junctions.

DEFINITION 2.5. *Assume an RS is assigned. Let $\rho = (\rho_1, \ldots, \rho_{n+m})$ and $\Pi = (\pi_1, \ldots, \pi_{n+m})$ be such that $\rho_i(t, \cdot)$ and $\pi_i(t, \cdot)$ are of bounded variation for every $t \geq 0$. Then (ρ, Π) is an admissible weak solution of* (1.1) *related to the RS at the junction J if and only if the following properties hold:*

1. *ρ is a weak solution at junction J;*
2. *Π is a weak solution at junction J;*
3. *for almost every t setting*

$$\rho_J(t) = (\rho_1(\cdot, b_1-), \ldots, \rho_n(\cdot, b_n-), \rho_{n+1}(\cdot, a_{n+1}+), \ldots, \rho_{n+m}(\cdot, a_{n+m}+)),$$
$$\Pi_J(t) = (\pi_1(\cdot, b_1-), \ldots, \pi_n(\cdot, b_n-), \pi_{n+1}(\cdot, a_{n+1}+), \ldots, \pi_{n+m}(\cdot, a_{n+m}+))$$

we have

$$RS(\rho_J(t), \Pi_J(t)) = (\rho_J(t), \Pi_J(t)).$$

Given an admissible network (see [12]) we have to specify how to define a solution.

DEFINITION 2.6. *Consider an admissible network $(N, \mathcal{I}, \mathcal{F}, \mathcal{J}, \mathcal{S}, \mathcal{D}, \mathcal{R})$. A set of initial-boundary conditions (IBC) is given assigning measurable functions $\bar{\rho}_i : I_i \mapsto [0, \rho_i^{\max}]$, $\bar{\pi}_i : [a_i, b_i] \times \mathcal{S} \times \mathcal{D} \mapsto [0, 1]$, $i = 1, \ldots, N$, and measurable functions $\psi_i : [0, +\infty[\mapsto [0, \rho_i^{\max}]$, $i \in \mathcal{S} \cup \mathcal{D}$, and $\vartheta_{i,j} : [0, +\infty[\mapsto [0, 1]$, $i \in \mathcal{S}$, $j \in \mathcal{D}$, with the property that $\sum_j \vartheta_{i,j}(t) = 1$.*

DEFINITION 2.7. *Consider an admissible network $(N, \mathcal{I}, \mathcal{F}, \mathcal{J}, \mathcal{S}, \mathcal{D}, \mathcal{R})$ and a set of IBC. A set of functions $\rho = (\rho_1, \ldots, \rho_N)$ with $\rho_i : [0, +\infty[\times I_i \mapsto [0, \rho_i^{\max}]$ continuous as functions from $[0, +\infty[$ into L^1, and $\Pi = (\pi_1, \ldots, \pi_N)$ with $\pi_i : [0, +\infty[\times I_i \times \mathcal{S} \times \mathcal{D} \mapsto [0, 1]$ continuous as functions from $[0, +\infty[$ into L^1 for every $s \in \mathcal{S}$, $d \in \mathcal{D}$, is an admissible solution if the following holds. Each ρ_i is a weak entropic solution to* (2.3) *on I_i, $\rho_i(0, x) = \bar{\rho}_i(x)$ for almost every $x \in [a_i, b_i]$, $\rho_i(t, a_i) = \psi_i(t)$ if $i \in S$ and $\rho_i(t, b_i) = \psi_i(t)$ if $i \in \mathcal{D}$ in the sense of* [5]. *Each π_i is a weak solution to the corresponding equation* (2.7), *$\pi_i(0, x) = \bar{\pi}_i(x)$ for almost every $x \in [a_i, b_i]$, and for every $i \in \mathcal{S}$, $j \in \mathcal{D}$ $\pi_i^{i,j}(t, a_i) = \vartheta_{i,j}$ in the sense of* [5]. *Finally, at each junction (ρ, Π) is a weak solution and is an admissible weak solution in the case of bounded variation.*

3. Traffic distribution at junctions.

3. Traffic distribution at junctions. Consider a junction J in which there are n transmission lines with incoming traffic and m transmission lines with outgoing traffic.

We denote by $\rho_i(t, x)$, $i = 1, \ldots, n$, and $\rho_j(t, x)$, $j = n+1, \ldots, n+m$, the traffic densities, respectively, on the incoming transmission lines and on the outgoing ones and by $(\rho_{1,0}, \ldots, \rho_{n+m,0})$ the initial datum.

Define $\gamma_i^{\max}$ and $\gamma_j^{\max}$ as follows:

$$\gamma_i^{\max} = \begin{cases} f(\rho_{i,0}) & \text{if } \rho_{i,0} \in [0, \sigma], \\ f(\sigma) & \text{if } \rho_{i,0} \in]\sigma, 1], \end{cases} \quad i = 1, \ldots, n, \tag{3.1}$$

and

$$\gamma_j^{\max} = \begin{cases} f(\sigma) & \text{if } \rho_{j,0} \in [0, \sigma], \\ f(\rho_{j,0}) & \text{if } \rho_{j,0} \in]\sigma, 1], \end{cases} \quad j = n+1, \ldots, n+m. \tag{3.2}$$

The quantities $\gamma_i^{\max}$ and $\gamma_j^{\max}$ represent the maximum flux that can be obtained by a single wave solution on each transmission line. Finally, denote by

$$\begin{aligned}\Omega_i &= [0, \gamma_i^{\max}], \quad i = 1, \dots, n,\\ \Omega_j &= [0, \gamma_j^{\max}], \quad j = n+1, \dots, n+m,\end{aligned}$$

and by $\hat{\gamma}_{inc} = (f(\hat{\rho}_i), \dots, f(\hat{\rho}_n))$, $\hat{\gamma}_{out} = (f(\hat{\rho}_{n+1}), \dots, f(\hat{\rho}_{n+m}))$, where $\hat{\rho} = (\hat{\rho}_1, \dots, \hat{\rho}_{n+m})$ is the solution of the Riemann problem at the junction.

Now, we discuss some possible choices for the traffic distribution function:

(1) $r_J : \mathrm{Inc}(J) \times \mathcal{S} \times \mathcal{D} \to \mathrm{Out}(J)$.

(2) $r_J : \mathrm{Inc}(J) \times \mathcal{S} \times \mathcal{D} \hookrightarrow \mathrm{Out}(J)$; i.e., r_J is a multifunction.

If r_J is of type (1), then each packet has a deterministic route; this means that, at the junction J, the traffic that started at source s and has d as the final destination, coming from the transmission line i, is routed on an assigned line j ($r_J(i, s, d) = j$).

Instead, if r_J is of type (2), at the junction J, the traffic with source s and destination d, coming from a line i, is routed on every line $I_j \in \mathrm{Out}(J)$ or on some lines $I_j \in \mathrm{Out}(J)$. We can define $r_J(i, s, d)$ in two different ways:

(2a) $r_J : \mathrm{Inc}(J) \times \mathcal{S} \times \mathcal{D} \hookrightarrow \mathrm{Out}(J)$,
$r_j(i, s, d) \subseteq \mathrm{Out}(J)$;

(2b) $r_J : \mathrm{Inc}(J) \times \mathcal{S} \times \mathcal{D} \to [0, 1]^{\mathrm{Out}(J)}$,
$r_J(i, s, d) = (\alpha_J^{i,s,d,n+1}, \dots, \alpha_J^{i,s,d,n+m})$
with $0 \le \alpha_J^{i,s,d,j} \le 1$, $j \in \{n+1, \dots, n+m\}$, $\sum_{j=n+1}^{n+m} \alpha_J^{i,s,d,j} = 1$.

In case (2a) we have to specify in which way the traffic at junction J is split towards the outgoing lines.

The definition (2b) means that, at the junction J, the traffic with source s and destination d, coming from line I_i, is routed on the outgoing line j, $j = n+1, \dots, n+m$, with probability $\alpha_J^{i,s,d,j}$.

Let us analyze how the distribution matrix A is constructed using π and r_J.

Definition 3.1. *A distribution matrix is a matrix*

$$A \doteq \{\alpha_{j,i}\}_{j=n+1,\dots,n+m, i=1,\dots,n} \in \mathbb{R}^{m \times n}$$

such that

$$0 < \alpha_{j,i} < 1, \qquad \sum_{j=n+1}^{n+m} \alpha_{j,i} = 1,$$

for each $i = 1, \dots, n$ and $j = n+1, \dots, n+m$, where $\alpha_{j,i}$ is the percentage of packets arriving from the ith incoming transmission line that take the jth outgoing transmission line.

In case (1) we can define the matrix A in the following way. Fix a time t, and assume that for all $i \in \mathrm{Inc}(J)$, $s \in \mathcal{S}$, and $d \in \mathcal{D}$, $\pi_i(t, \cdot, s, d)$ admits a limit at the junction J, i.e., the left limit at b_i. For $i \in \{1, \dots, n\}$, $j \in \{n+1, \dots, n+m\}$, we set

$$\alpha_{j,i} = \sum_{\substack{s \in \mathcal{S}, d \in \mathcal{D},\\ r_J(i,s,d) = j}} \pi_i(t, b_i-, s, d).$$

The fluxes $f_i(\rho_i)$ to be consistent with the traffic-type functions must satisfy the following relation:

$$f_j(\rho_j(\cdot, a_j+)) = \sum_{i=1}^{n} \alpha_{j,i} f_i(\rho_i(\cdot, b_i-))$$

for each $j = n+1, \dots, n+m$.

Let us analyze how to define the matrix A in case (2a). We may assign $\varphi(i,s,d) \in r_J(i,s,d)$ and set

$$\alpha_{j,i} = \sum_{\substack{s\in\mathcal{S},d\in\mathcal{D},\\ i:\varphi(i,s,d)=j}} \pi_i(t,b_i-,s,d),$$
$$\alpha_{j,i} = 0 \text{ if } j \notin r_J(i,s,d).$$

Example 3.2. *Fix a junction J with two incoming lines $\{1,2\}$ and two outgoing lines $\{3,4\}$, and suppose that $r_J(1,s,d) = \{3,4\}$ and $r_J(2,s,d) = \{3\}$. Since $\alpha_{4,2} = 0$, we have $\alpha_{3,2} = 1$. The coefficients $\alpha_{3,1}$ and $\alpha_{4,1}$ can assume the following values:*

$$\begin{cases} \alpha_{3,1} = 0, \\ \alpha_{4,1} = 1, \end{cases} \quad or \quad \begin{cases} \alpha_{3,1} = 1, \\ \alpha_{4,1} = 0. \end{cases}$$

We get a finite number of possible distribution matrices A:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Remark 3.3. *This model proposes an exclusive strategy; in fact all packet flow at the junction is routed from line 1 to line 3 or to line 4.*

However, it is more natural to assign a flexible strategy defining a set of admissible matrices A in the following way:

$$\mathcal{A} = \left\{ \begin{array}{c} A : \exists \alpha_J^{i,s,d,j} \in [0,1], \displaystyle\sum_{j=n+1}^{n+m} \alpha_J^{i,s,d,j} = 1, \alpha_J^{i,s,d,j} = 0 \text{ if } j \notin r_J(i,s,d) : \\ \alpha_{j,i} = \displaystyle\sum_{\substack{s\in\mathcal{S},d\in\mathcal{D},\\ j\in r_J(i,s,d)}} \pi_i(t,b_i-,s,d)\alpha_J^{i,s,d,j} \end{array} \right\}.$$

Finally, we now treat case (2b).

In this case the matrix A is unique and is defined by

$$\alpha_{j,i} = \sum_{s\in\mathcal{S},d\in\mathcal{D}} \pi_i(t,b_i-,s,d)\alpha_J^{i,s,d,j}. \tag{3.3}$$

4. Riemann solvers at junctions. In this section we define solutions to Riemann problems at junctions, since this is the basic ingredient to construct solutions to Cauchy problems via a wave-front tracking algorithm.

We describe two different Riemann solvers at a junction that represent two different routing algorithms:

(RA1) We assume that
  (A) the traffic from incoming transmission lines is distributed on outgoing transmission lines according to fixed coefficients;
  (B) respecting (A) the router chooses to send packets in order to maximize fluxes (i.e., the number of packets which are processed).

(RA2) We assume that the number of packets through the junction is maximized both over incoming and outgoing lines.

Remark 4.1. *In what follows we analyze the case in which the traffic distribution function is of type (2). Case (1) has been considered in [12] using the following rule:*

(RGP) *We assume that*

(A) *the traffic from incoming transmission lines is distributed on outgoing transmission lines according to fixed coefficients;*

(B) *respecting* (A) *the router chooses to send packets in order to maximize*

$$c_2 \sum_{i=1}^{n} f_i(\rho_i(\cdot, b_i-)) - c_1[dist((f_1(\rho_1(\cdot, b_1-)), \ldots, f_n(\rho_n(\cdot, b_n-))), r)]^2$$

subject to

$$f_j(\rho_j(\cdot, a_j+)) = \sum_{i=1}^{n} \alpha_{j,i} f_i(\rho_i(\cdot, b_i-)) \quad \textit{for each } j = n+1, \ldots, n+m,$$

where c_1 and c_2 are strictly positive constants, and $dist(\cdot, r)$ denotes the Euclidean distance in $\mathbb{R}^n$ from the line r, which is given by

$$\begin{cases} \gamma_2 = p_1\gamma_1, \\ \vdots \\ \gamma_n = p_{n-1}\gamma_{n-1}, \end{cases}$$

and $(p_1, \ldots, p_{n-1})$ determine a "level of priority" at the junctions of incoming lines. This maximization procedure takes into account priorities over incoming roads and ensures continuity of solutions with respect to the coefficients π.

4.1. Algorithm (RA1). We have to distinguish cases (2a) and (2b).

In case (2a) in order to solve the Riemann problem at the junction we have to prove that the admissible region is convex. First we prove the following lemma.

Lemma 4.2. *The set $\mathcal{A}$ is convex.*

Proof. Let us consider a convex combination $\lambda A_1 + (1-\lambda)A_2$ with $\lambda \in [0,1], A_1, A_2 \in \mathcal{A}$. We have

$$\begin{aligned}(\lambda A_1 + (1-\lambda)\mathcal{A}_2)_{i,j} &= \lambda \sum_{\substack{s\in\mathcal{S}, d\in\mathcal{D},\\ j\in r_J(i,s,d)}} \pi_i \alpha_{J,1}^{i,s,d,j} + (1-\lambda) \sum_{\substack{s\in\mathcal{S}, d\in\mathcal{D},\\ j\in r_J(i,s,d)}} \pi_i \alpha_{J,2}^{i,s,d,j} \\ &= \sum_{\substack{s\in\mathcal{S}, d\in\mathcal{D},\\ j\in r_J(i,s,d)}} \pi_i(\lambda\alpha_{J,1}^{i,s,d,j} + (1-\lambda)\alpha_{J,2}^{i,s,d,j}) = \sum_{\substack{s\in\mathcal{S}, d\in\mathcal{D},\\ j\in r_J(i,s,d)}} \pi_i \hat{\alpha}_J^{i,s,d,j},\end{aligned}$$

with $\hat{\alpha}_J^{i,s,d,j} \in [0,1]$. Moreover,

$$\sum_{j=n+1}^{n+m} \hat{\alpha}_J^{i,s,d,j} = \sum_{j=n+1}^{n+m} (\lambda\alpha_{J,1}^{i,s,d,j} + (1-\lambda)\alpha_{J,2}^{i,s,d,j}) = \lambda \sum_{j=n+1}^{n+m} \alpha_{J,1}^{i,s,d,j} + (1-\lambda) \sum_{j=n+1}^{n+m} \alpha_{J,2}^{i,s,d,j} = 1;$$

then $\lambda A_1 + (1-\lambda)A_2 \in \mathcal{A}$. □

Now recall that the admissible region is given by

$$\Omega_{adm} = \{\hat{\gamma} : \hat{\gamma} \in \Omega_1 \times \cdots \times \Omega_n, \exists A \in \mathcal{A}\ t.c. A\hat{\gamma} \in \Omega_{n+1} \times \cdots \times \Omega_{n+m}\}.$$

We can prove that this region is convex at least for the case of junctions with two incoming and two outgoing lines; more precisely, we have the following lemma.

Lemma 4.3. *Fix a junction J with $n = 2$ incoming lines and $m = 2$ outgoing ones, and assume that there is a unique source and a unique destination. Then the set Ω_{adm} is convex.*

Proof. We have to consider the following cases:

(i) $r_J(1,s,d)=3$ and $r_J(2,s,d)=3$;
(ii) $r_J(1,s,d)=3$ and $r_J(2,s,d)=4$;
(iii) $r_J(1,s,d)=\{3,4\}$ and $r_J(2,s,d)=4$;
(iv) $r_J(1,s,d)=\{3,4\}$ and $r_J(2,s,d)=\{3,4\}$.

All other cases can be obtained by relabelling lines. Cases (i) and (ii) are immediate, since $\hat\gamma\in\Omega_1\times\Omega_2$ satisfies $\hat\gamma\in\Omega_{adm}$ if and only if $\hat\gamma_1+\hat\gamma_2\le\gamma_3^{max}$ (case i) or $\hat\gamma_1\le\gamma_3^{max}$ and $\hat\gamma_2\le\gamma_4^{max}$ (case ii).

Now consider case (iii). Then $A\hat\gamma$, $A\in\mathcal{A}$, is the segment joining the point $(\hat\gamma_1,\hat\gamma_2)$ to the point $(0,\hat\gamma_1+\hat\gamma_2)$. Thus $\hat\gamma\in\Omega_1\times\Omega_2$ satisfies $\hat\gamma\in\Omega_{adm}$ if and only if $\hat\gamma_1+\hat\gamma_2\le\gamma_3^{max}+\gamma_4^{max}$ and $\hat\gamma_2\le\gamma_3^{max}$.

Finally, assume case (iv) holds true. Then $A\hat\gamma$, $A\in\mathcal{A}$, is the segment joining the point $(\hat\gamma_1+\hat\gamma_2,0)$ to the point $(0,\hat\gamma_1+\hat\gamma_2)$. Thus $\hat\gamma\in\Omega_1\times\Omega_2$ satisfies $\hat\gamma\in\Omega_{adm}$ if and only if $\hat\gamma_1+\hat\gamma_2\le\gamma_3^{max}+\gamma_4^{max}$. □

If the region Ω_{adm} is convex, then rules (A) and (B) amount to the linear programming problem:

$$\max_{\hat\gamma\in\Omega_{adm}}(\hat\gamma_1+\hat\gamma_2).$$

This problem clearly has a solution, which may not be unique.

Let us consider case (2b). We need some more notation.

Definition 4.4. *Let $\tau:[0,1]\to[0,1]$ be the map such that*

1. *$f(\tau(\rho))=f(\rho)$ for every $\rho\in[0,1]$;*
2. *$\tau(\rho)\neq\rho$ for every $\rho\in[0,1]\backslash\{\sigma\}$.*

Clearly, τ is well defined and satisfies

$$0\le\rho\le\sigma\Leftrightarrow\sigma\le\tau(\rho)\le1,$$
$$\sigma\le\rho\le1\Leftrightarrow0\le\tau(\rho)\le\sigma.$$

To state the main result of this section we need some assumption on the matrix A (satisfied under generic conditions for $m=n$). Let $\{e_1,\dots,e_n\}$ be the canonical basis of $\mathbb{R}^n$, and for every subset $V\subset\mathbb{R}^n$ indicate by $V^\perp$ its orthogonal. Define for every $i=1,\dots,n$, $H_i=\{e_i\}^\perp$, i.e., the coordinate hyperplane orthogonal to e_i, and for every $j=n+1,\dots,n+m$ let $\alpha_j=\{\alpha_{j1},\dots,\alpha_{jn}\}\in\mathbb{R}^n$, and define $H_j=\{\alpha_j\}^\perp$. Let $\mathcal{K}$ be the set of indices $k=(k_1,\dots,k_l)$, $1\le l\le n-1$, such that $0\le k_1<k_2<\dots<k_l\le n+m$, and for every $k\in\mathcal{K}$ set $H_k=\bigcap_{h=1}^{l}H_h$. Letting $\mathbf{1}=(1,\dots,1)\in\mathbb{R}^n$, we assume that

(C) for every $k\in\mathcal{K}$, $\mathbf{1}\notin H_k^\perp$.

In case (2b) the following result holds.

Theorem 4.5 (Theorem 3.1 in [7] and 3.2 in [12]). *Let $(N,\mathcal{I},\mathcal{F},\mathcal{J},\mathcal{S},\mathcal{D},\mathcal{R})$ be an admissible network and J a junction with n incoming lines and m outgoing ones. Assume that the flux $f:[0,1]\to\mathbb{R}$ satisfies* (F) *and the matrix A satisfies condition* (C). *For every $\rho_{1,0},\dots,\rho_{n+m,0}\in[0,1]$, and for every $\pi_1^{s,d},\dots,\pi_{n+m}^{s,d}\in[0,1]$, there exist densities $\hat\rho_1,\dots,\hat\rho_{n+m}$ and a unique admissible centered weak solution, $\rho=(\rho_1,\dots,\rho_{n+m})$, at the junction J such that*

$$\rho_1(0,\cdot)\equiv\rho_{1,0},\dots,\rho_{n+m}(0,\cdot)\equiv\rho_{n+m,0},$$
$$\pi^1(0,\cdot,s,d)=\pi_1^{s,d},\dots,\pi^{n+m}(0,\cdot,s,d)=\pi_{n+m}^{s,d}(s\in\mathcal{S},d\in\mathcal{D}).$$

We have

$$\hat{\rho}_i \in \left\{ \begin{array}{ll} \{\rho_{i,0}\} \cup]\tau(\rho_{i,0}), 1] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [\sigma, 1] & \text{if } \sigma \le \rho_{i,0} \le 1, \end{array} \right. \quad i = 1, \dots, n, \tag{4.1}$$

$$\hat{\rho}_j \in \left\{ \begin{array}{ll} [0, \sigma] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ \{\rho_{j,0}\} \cup [0, \tau(\rho_{j,0})[& \text{if } \sigma \le \rho_{j,0} \le 1, \end{array} \right. \quad j = n+1, \dots, n+m, \tag{4.2}$$

and on each incoming line I_i, $i = 1, \dots, n$, the solution consists of the single wave $(\rho_{i,0}, \hat{\rho}_i)$, while on each outgoing line I_j, $j = n+1, \dots, n+m$, the solution consists of the single wave $(\hat{\rho}_j, \rho_{j,0})$. Moreover, $\hat{\pi}_i(t, \cdot, s, d) = \pi_i^{s,d}$ for every $t \ge 0$, $i \in \{1, \dots, n\}$, $s \in \mathcal{S}$, $d \in \mathcal{D}$, and

$$\hat{\pi}_j(t, a_j+, s, d) = \frac{\sum_{i=1}^n \alpha_J^{i,s,d,j} \pi_i^{s,d}(t, b_i-, s, d) f(\hat{\rho}_i)}{f(\hat{\rho}_j)}$$

for every $t \ge 0$, $j \in \{n+1, \dots, n+m\}$, $s \in \mathcal{S}$, $d \in \mathcal{D}$.

4.2. Algorithm (RA2). To solve Riemann problems according to (RA2) we need some additional parameters called priority and traffic distribution parameters. For simplicity of exposition, consider first a junction J in which there are two transmission lines with incoming traffic and two transmission lines with outgoing traffic. In this case we have only one priority parameter $q \in]0, 1[$ and one traffic distribution parameter $\alpha \in]0, 1[$. We denote by $(\rho_{1,0}, \rho_{2,0}, \rho_{3,0}, \rho_{4,0})$ and $(\pi_{1,0}^{s,d}, \pi_{2,0}^{s,d}, \pi_{3,0}^{s,d}, \pi_{4,0}^{s,d})$ the initial data.

In order to maximize the number of packets through the junction over incoming and outgoing lines we define

$$\Gamma = \min \{\Gamma_{in}^{\max}, \Gamma_{out}^{\max}\},$$

where $\Gamma_{in}^{\max} = \gamma_1^{\max} + \gamma_2^{\max}$ and $\Gamma_{out}^{\max} = \gamma_3^{\max} + \gamma_4^{\max}$. Thus we want to have Γ as the flux through the junction.

One easily sees that to solve the Riemann problem, it is enough to determine the fluxes $\hat{\gamma}_i = f(\hat{\rho}_i)$, $i = 1, 2$. In fact, to have simple waves with the appropriate velocities, i.e., negative on incoming lines and positive on outgoing ones, we get the constraints (4.1), (4.2). Observe that we compute $\hat{\gamma}_i = f(\hat{\rho}_i)$, $i = 1, 2$, without taking into account the type of traffic distribution function.

We have to distinguish two cases:

I. $\Gamma_{in}^{\max} = \Gamma$;

II. $\Gamma_{in}^{\max} > \Gamma$.

In the first case we set $\hat{\gamma}_i = \gamma_i^{\max}$, $i = 1, 2$.

Let us analyze the second case in which we use the priority parameter q. Not all packets can enter the junction, and so let C be the amount of packets that can go through. Then qC packets come from the first incoming line and $(1 - q)C$ packets from the second. Consider the space (γ_1, γ_2), and define the following lines:

$$r_q : \gamma_2 = \frac{1 - q}{q} \gamma_1,$$
$$r_\Gamma : \gamma_1 + \gamma_2 = \Gamma.$$

Define P to be the point of intersection of the lines r_q and r_Γ. Recall that the final fluxes should belong to the region (see Figure 4.1):

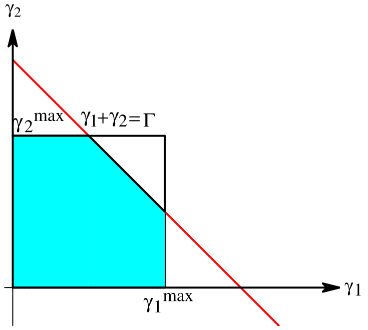


FIG. 4.1. *Case* $\Gamma_{in}^{\max} > \Gamma$.

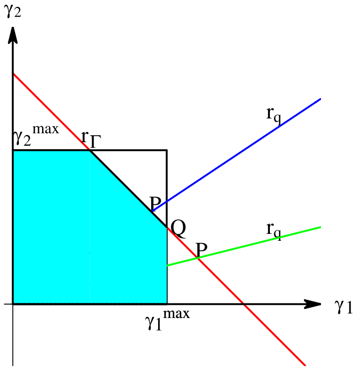


FIG. 4.2. *P belongs to* Ω, *and P is outside* Ω.

$$\Omega = \{(\gamma_1, \gamma_2) : 0 \leq \gamma_i \leq \gamma_i^{\max}, i = 1, 2\}.$$

We distinguish two cases:

(a) P belongs to Ω;

(b) P is outside Ω.

In the first case we set $(\hat{\gamma}_1, \hat{\gamma}_2) = P$, while in the second case we set $(\hat{\gamma}_1, \hat{\gamma}_2) = Q$, with $Q = proj_{\Omega \cap r_\Gamma}(P)$, where $proj$ is the usual projection on a convex set; see Figure 4.2.

The reasoning can be repeated also in the case of n incoming lines. In $\mathbb{R}^n$ the line r_q is given by $r_q = tv_q$, $t \in \mathbb{R}$, with $v_q \in \Delta_{n-1}$, where

$$\Delta_{n-1} = \left\{(\gamma_1, \dots, \gamma_n) : \gamma_i \geq 0, i = 1, \dots, n, \sum_{i=1}^{n} \gamma_i = 1\right\}$$

is the $(n-1)$ dimensional simplex and

$$H_\Gamma = \left\{(\gamma_1, \dots, \gamma_n) : \sum_{i=1}^{n} \gamma_i = \Gamma\right\}$$

is a hyperplane where $\Gamma = \min\{\sum_{in} \gamma_i^{\max}, \sum_{out} \gamma_j^{\max}\}$. Since $v_q \in \Delta_{n-1}$, there exists a unique point $P = r_q \cap H_\Gamma$. If $P \in \Omega$, then we set $(\hat{\gamma}_1, \dots, \hat{\gamma}_n) = P$. If $P \notin \Omega$, then we set $(\hat{\gamma}_1, \ldots, \hat{\gamma}_n) = Q = proj_{\Omega \cap H_\Gamma}(P)$, the projection over the subset $\Omega \cap H_\Gamma$. Observe that the projection is unique since $\Omega \cap H_\Gamma$ is a closed convex subset of H_Γ.

Remark 4.6. *A possible alternative definition in the case $P \notin \Omega$ is to set $(\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$ as one of the vertices of $\Omega \cap H_\Gamma$.*

As for algorithm (RA1) $\hat{\pi}_i^{s,d} = \pi_{i,0}^{s,d}$, $i = 1, 2$.

Let us now determine $\hat{\gamma}_j$, $j = 3, 4$.

As for the incoming transmission lines we have to distinguish two cases:

I. $\Gamma_{out}^{\max} = \Gamma$;

II. $\Gamma_{out}^{\max} > \Gamma$.

In the first case $\hat{\gamma}_j = \gamma_j^{\max}$, $j = 3, 4$. Let us determine $\hat{\gamma}_j$ in the second case.

Recall α, the traffic distribution parameter. Since not all packets can go on the outgoing transmission lines, we let C be the amount that goes through. Then αC packets go on the outgoing line I_3 and $(1-\alpha)C$ on the outgoing line I_4. Consider the space (γ_3, γ_4), and define the following lines:

$$r_\alpha : \gamma_4 = \frac{1-\alpha}{\alpha}\gamma_3,$$
$$r_\Gamma : \gamma_3 + \gamma_4 = \Gamma.$$

The line r_α can be computed from the matrix A. In fact, if we assume that a traffic distribution matrix A is assigned, then we compute $\hat{\gamma}_1, \ldots, \hat{\gamma}_n$ and choose $v_\alpha \in \Delta_{m-1}$ by

$$v_\alpha = \Delta_{m-1} \cap \{tA(\hat{\gamma}_1, \ldots, \hat{\gamma}_n) : t \in \mathbb{R}\},$$

where

$$\Delta_{m-1} = \left\{(\gamma_{n+1}, \ldots, \gamma_{m+n}) : \gamma_{n+i} \geq 0, i = 1, \ldots, m, \sum_{i=1}^{n} \gamma_{n+i} = 1\right\}$$

is the $(m-1)$ dimensional simplex.

We have to distinguish cases (2a) and (2b) for the traffic distribution function.

Case (2a). Let us introduce the set

$$\mathcal{G} = \left\{A\hat{\gamma}_{inc}^T : A \in \mathcal{A}\right\}.$$

Lemma 4.7. *The set $\mathcal{G}$ is connected.*

Proof. The set $\mathcal{G}$ is the image of a connected set through a continuous map. With fixed $(\hat{\gamma}_1, \hat{\gamma}_2)$ the map is defined by

$$(\tilde{\alpha}_J^{1,s,d,3}, \tilde{\alpha}_J^{2,s,d,3}) \in [0,1] \times [0,1] \to (\Sigma, \hat{\gamma}_1 + \hat{\gamma}_2 - \Sigma),$$

where $\Sigma = \sum_{s,d}(\hat{\gamma}_1 \pi_1^{s,d} \tilde{\alpha}_J^{1,s,d,3} + \hat{\gamma}_2 \pi_2^{s,d} \tilde{\alpha}_J^{2,s,d,3})$. □

Let us denote with G_1 and G_2 the endpoints of this set. Since in case (2a) we have an infinite number of matrices A, each one determining a line r_α, we choose the most “natural” line r_α, i.e., the one nearest to the statistic line determined by measurements on the network.

Recall that the final fluxes should belong to the region:

$$\Omega = \left\{(\gamma_3, \gamma_4) : 0 \leq \gamma_j \leq \gamma_j^{\max}, j = 3, 4\right\}.$$

Define $P = r_\alpha \cap r_\Gamma$, $R = (\Gamma - \gamma_4^{\max}, \gamma_4^{\max})$, $Q = (\gamma_3^{\max}, \Gamma - \gamma_3^{\max})$. We distinguish three cases:

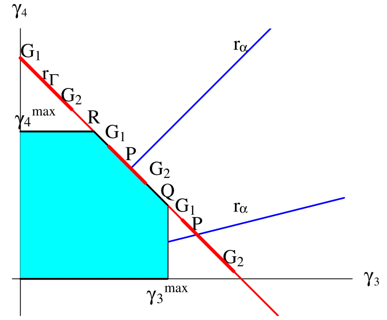


Fig. 4.3. *Traffic distribution function of type* (2a).

(a) $\mathcal{G} \cap \Omega \cap r_\Gamma \neq \varnothing$;
(b) $\mathcal{G} \cap \Omega \cap r_\Gamma = \varnothing$ and $\gamma_3(G_1) < \gamma_3(R)$;
(c) $\mathcal{G} \cap \Omega \cap r_\Gamma = \varnothing$ and $\gamma_3(G_1) > \gamma_3^{\max}$.

If the set $\mathcal{G}$ has a priority over the line r_Γ, we set $(\hat{\gamma}_3, \hat{\gamma}_4)$ in the following way. In case (a) we define $(\hat{\gamma}_3, \hat{\gamma}_4) = proj_{\mathcal{G}\cap\Omega\cap r_\Gamma}(P)$, in case (b) $(\hat{\gamma}_3, \hat{\gamma}_4) = R$, and finally in case (c) $(\hat{\gamma}_3, \hat{\gamma}_4) = Q$. The three cases are shown in Figure 4.3.

Otherwise, if r_Γ has a priority over $\mathcal{G}$, we set $(\hat{\gamma}_3, \hat{\gamma}_4) = \min_{\gamma\in\Omega} \mathcal{F}(\gamma, r_\alpha, \mathcal{G})$, where $\mathcal{F}$ is a convex functional which depends on γ, r_α, and the set $\mathcal{G}$ of the routing standards. A possible choice of $\mathcal{F}$ is $\mathcal{F} = d(\gamma, B)$, where $B = w_1 r_\alpha + w_2 \int_{\mathcal{G}} r dr$ with w_1, w_2 real numbers and d denotes a distance.

The reasoning can also be repeated in the case of m outgoing lines.

The vector $\hat{\pi}_i^{s,d}$, $j = 3, 4$, is computed in the same way as for algorithm (RA1).

Case (2b). In case (2b) we have a unique matrix A and a unique vector v_α, and so the fluxes on outgoing lines are computed as in the case without sources and destinations.

We distinguish two cases:
(a) P belongs to Ω;
(b) P is outside Ω.

In the first case we set $(\hat{\gamma}_3, \hat{\gamma}_4) = P$, while in the second case we set $(\hat{\gamma}_3, \hat{\gamma}_4) = Q$, where $Q = proj_{\Omega_{adm}}(P)$. Again, we can extend to the case of m outgoing lines as for the incoming lines defining the hyperplane $H_\Gamma = \{(\gamma_{n+1}, \ldots, \gamma_{n+m}) : \sum_{j=n+1}^{n+m} \gamma_j = \Gamma\}$ and choosing a vector $v_\alpha \in \Delta_{m-1}$.

Finally, we define $\hat{\pi}_i^{s,d}$, $j = 3, 4$, as in case (2a):

$$\hat{\pi}_j(t, a_j+, s, d) = \frac{\sum_{i=1}^{n} \alpha_J^{i,s,d,j} \pi_i^{s,d}(t, b_i-, s, d) f(\hat{\rho}_i)}{f(\hat{\rho}_j)}$$

for every $t \geq 0$, $j \in \{n+1, \ldots, n+m\}$, $s \in \mathcal{S}$, $d \in \mathcal{D}$.

Remark 4.8. *Note that in the case of algorithm* (RA2) *we find, separately, a solution on incoming and outgoing lines.*

Remark 4.9. *If* $\Gamma_{out}^{\max} < \Gamma_{in}^{\max}$*, we can define a different Riemann solver, considering a priority order of sending packets:* $(s_{k_1}, d_{l_1}) = c_1$, $(s_{k_2}, d_{l_2}) = c_2$, $(s_{k_3}, d_{l_3}) = c_3, \ldots$*. Packets are sent until the quantity of packets that has been sent is equal to*

$$\sum_{\iota=1}^{\bar{\iota}} \sum_{i=1}^{n} \pi_i^{c_\iota} \gamma_{i0},$$

where $\bar{\iota}$ is the minimum such that

$$\sum_{\iota=1}^{\bar{\iota}} \sum_{i=1}^{n} \pi_i^{c_\iota} \gamma_{i0} > \Gamma.$$

Let us define $d = \Gamma - \sum_{\iota=1}^{\bar{\iota}} \sum_{i=1}^{n} \pi_i^{c_\iota} \gamma_{i0}$*; then*

$$\hat{\gamma}_1 = \sum_{\iota=1}^{\bar{\iota}-1} \pi_1^{c_\iota} \gamma_{10} + d/2,$$

$$\hat{\gamma}_2 = \sum_{\iota=1}^{\bar{\iota}-1} \pi_2^{c_\iota} \gamma_{20} + d/2.$$

Once solutions to Riemann problems are given, one can use a wave-front tracking algorithm to construct a sequence of approximate solutions. To pass to the limit one has to bound the number of waves and the BV norm of approximate solutions; see [6, 7]. In the next section we prove a BV bound on the density for the case of junctions with two incoming and two outgoing transmission lines for both of the routing algorithms.

5. Estimates on density variation. In this section we derive estimates on the total variation of the densities along a wave-front tracking approximate solution (constructed as in [7]) for algorithm (RA2) with the traffic distribution function of type (2b). This allows us to construct the solutions to the Cauchy problem in the standard way; see [6].

Let us consider an admissible network $(N, \mathcal{I}, \mathcal{F}, \mathcal{J}, \mathcal{S}, \mathcal{D}, \mathcal{R})$. We assume that

(A1) every junction has at most two incoming and at most two outgoing lines.

This hypothesis is crucial, because the presence of more complicated junctions may provoke additional increases of the total variation of the flux and thus of the density. The case where junctions have at most two incoming transmission lines and at most two outgoing ones can be treated in the same way.

From now on we fix a telecommunication network $(\mathcal{I}, \mathcal{J})$, with each node having at most two incoming and at most two outgoing lines, and a wave-front tracking approximate solution ρ, Π, defined on the telecommunication network.

Our aim is to prove an existence result for a solution (ρ, Π) in the case of a small perturbation of the equilibrium $(\bar{\rho}, \bar{\Pi})$. We have to analyze the following types of interactions:

I1. interaction of ρ-waves with ρ-waves on lines;

I2. interaction of ρ-waves with Π-waves on lines;

I3. interaction of Π-waves with Π-waves on lines;

I4. interaction of ρ-waves with junctions;

I5. interaction of Π-waves with junctions.

Observe that interaction of type I1 is classical and the total variation of the density decreases. Interaction of type I3 cannot happen since Π-waves travel with speed depending only on the value of ρ.

5.1. Interaction of type I2. Let us consider a line I_i. We report some results proved in [12]. First we note that the characteristic speed of the density is smaller than the speed of a Π-wave, as follows from the next lemma.

Lemma 5.1. *Let $\rho \in [0, 1]$ be a density, and let $\lambda(\rho)$ be its characteristic speed. Then $\lambda(\rho) \leq v(\rho)$, and the equality holds if and only if $\rho = 0$.*

Lemma 5.2. *Let us consider a shock wave connecting ρ^- and ρ^+. Then*

1. $\lambda(\rho^-, \rho^+) < v(\rho^-)$;
2. $\lambda(\rho^-, \rho^+) \le v(\rho^+)$ *and the equality holds if and only if* $\rho^- = 0$.

LEMMA 5.3. *Let us consider a rarefaction shock fan connecting* ρ^- *and* ρ^+. *Then* $v(\rho^+) > v(\rho^-) > f'(\rho^-)$.

Putting together the previous lemmas we obtain the following result.

PROPOSITION 5.4. *An interaction of a* ρ*-wave with a* Π*-wave can happen only if the* Π*-wave interacts from the left with respect to the* ρ*-wave. Moreover, if this happens, then the* ρ*-wave does not change, while the* Π*-wave changes only its speed.*

5.2. Interaction of type I4.

We consider interactions of ρ-waves with the junctions. In general these interactions produce an increment of the total variation of the flux and of the density in all the lines and a variation of the values of traffic-type functions on outgoing lines.

Fix a junction J with two incoming transmission lines I_1 and I_2 and two outgoing ones I_3 and I_4. Suppose that at some time $\bar{t}$ a wave interacts with the junction J, and let $(\rho_1^-, \rho_2^-, \rho_3^-, \rho_4^-)$ and $(\rho_1^+, \rho_2^+, \rho_3^+, \rho_4^+)$ indicate the equilibrium configurations at the junction J before and after the interaction, respectively. Introduce the following notation:

$$\gamma_i^{\pm} = f(\rho_i^{\pm}), \quad \Gamma_{in}^{\pm} = \gamma_{1,\max}^{\pm} + \gamma_{2,\max}^{\pm}, \quad \Gamma_{out}^{\pm} = \gamma_{3,\max}^{\pm} + \gamma_{4,\max}^{\pm},$$
$$\Gamma^{\pm} = \min\{\Gamma_{in}^{\pm}, \Gamma_{out}^{\pm}\},$$

where $\gamma_{i,\max}^{\pm}$, $i = 1, 2$, and $\gamma_{j,\max}^{\pm}$, $j = 3, 4$, are defined as in (3.1) and (3.2). In general $-$ and $+$ denote the values before and after the interaction, while by Δ we indicate the variation, i.e., the value after the interaction minus the value before. For example $\Delta\Gamma = \Gamma^+ - \Gamma^-$. Let us denote by $TV(f)^{\pm} = TV(f(\rho(\bar{t}\pm, \cdot)))$ the flux variation of waves before and after the interaction, and

$$TV(f)_{in}^{\pm} = TV(f(\rho_1(\bar{t}\pm, \cdot))) + TV(f(\rho_2(\bar{t}\pm, \cdot))),$$
$$TV(f)_{out}^{\pm} = TV(f(\rho_3(\bar{t}\pm, \cdot))) + TV(f(\rho_4(\bar{t}\pm, \cdot))),$$

the flux variation of waves before and after the interaction, respectively, on incoming and outgoing lines.

Let us prove some estimates which are used later to control the total variation of the density function. For simplicity, from now on we assume that

(A2) the wave interacting at time $\bar{t}$ with J comes from line 1, and we let ρ_1 be the value on the left of the wave.

The case of a wave from an outgoing line can be treated similarly.

LEMMA 5.5. *We have*

$$sgn\,(\Delta\gamma_3) \cdot sgn\,(\Delta\gamma_4) \ge 0.$$

LEMMA 5.6. *We have*

$$sgn(\gamma_1^+ - \gamma_1) \cdot sgn(\Delta\gamma_2) \ge 0,$$

where $\gamma_1 = f(\rho_1)$.

LEMMA 5.7. *It holds that*

$$TV(f)_{out}^{+} = |\Delta\Gamma|.$$

Lemma 5.8. *We have*

$$TV(f)^-_{in} = TV(f)^+_{in} + |\Delta\Gamma|. \tag{5.1}$$

From Lemmas 5.7 and 5.8, we are ready to state the following.

Lemma 5.9. *It holds that*

$$TV(f)^+ = CTV(f)^-.$$

A ρ-wave produces a Π-wave, but the following lemma holds.

Lemma 5.10. *Let J be a junction with at most two incoming lines and two outgoing ones. Suppose that a ρ-wave (ρ_1, ρ_{10}) approaches the junction J. If there exists $\delta > 0$ such that $f(\rho_1) > \delta > 0$, $f(\rho_{1,0}) > \delta > 0$, then there exists $C > 0$, such that the variation of the traffic-type functions in outgoing lines is bounded by C times the flux variation of the interacting wave, i.e.,*

$$TV(\Pi)^+ \leq \frac{C}{\delta} TV(f)^-.$$

Proof. Fix a source $s \in \mathcal{S}$ and a destination $d \in \mathcal{D}$. We denote by $\pi_{i,0}, \rho_{i,0}$ and $\hat{\pi}_i, \hat{\rho}_i$ $(i \in \{1,2,3,4\})$ the values of the densities and of the traffic-type functions for s and d at J, respectively, before and after the interaction of the ρ-wave with J. We have for $j \in \{3,4\}$

$$\begin{aligned}
&|\pi^{s,d}_{j,0} - \hat{\pi}^{s,d}_j| \\
&= \left| \frac{\alpha_J^{1,s,d,j} \pi^{s,d}_{1,0} f(\rho_{1,0})}{f(\rho_{j,0})} + \frac{\alpha_J^{2,s,d,j} \pi^{s,d}_{2,0} f(\rho_{2,0})}{f(\rho_{j,0})} - \frac{\alpha_J^{1,s,d,j} \pi^{s,d}_{2,0} f(\hat{\rho}_1)}{f(\hat{\rho}_j)} - \frac{\alpha_J^{2,s,d,j} \pi^{s,d}_{2,0} f(\hat{\rho}_2)}{f(\hat{\rho}_j)} \right| \\
&\leq \frac{\alpha_J^{1,s,d,j} \pi^{s,d}_{1,0}}{f(\rho_{j,0}) f(\hat{\rho}_j)} |f(\rho_{1,0}) f(\hat{\rho}_j) - f(\hat{\rho}_1) f(\rho_{j,0})| + \frac{\alpha_J^{2,s,d,j} \pi^{s,d}_{2,0}}{f(\rho_{j,0}) f(\hat{\rho}_j)} |f(\rho_{2,0}) f(\hat{\rho}_j) - f(\hat{\rho}_2) f(\rho_{j,0})| \\
&\qquad \leq \frac{C'}{\delta^2} |f(\rho_{1,0})(f(\hat{\rho}_j) - f(\rho_{j,0})) + f(\rho_{j,0})(f(\rho_{1,0}) - f(\hat{\rho}_1))| \\
&\qquad\quad + \frac{C'}{\delta^2} |f(\rho_{2,0})(f(\hat{\rho}_j) - f(\rho_{j,0})) + f(\rho_{j,0})(f(\rho_{2,0}) - f(\hat{\rho}_2))| \\
&\qquad \leq \frac{C'}{\delta^2} f(\rho_{1,0}) |f(\hat{\rho}_j) - f(\rho_{j,0})| + \frac{C'}{\delta^2} f(\rho_{j,0}) |f(\rho_{1,0}) - f(\hat{\rho}_1)| \\
&\qquad\quad + \frac{C'}{\delta^2} f(\rho_{2,0}) |f(\hat{\rho}_j) - f(\rho_{j,0})| + \frac{C'}{\delta^2} f(\rho_{j,0}) |f(\rho_{2,0}) - f(\hat{\rho}_2)| \\
&\qquad\qquad = \frac{C'}{\delta} |f(\hat{\rho}_j) - f(\rho_{j,0})| + \frac{C'}{\delta} |f(\rho_{1,0}) - f(\hat{\rho}_1)| \\
&\qquad\qquad\quad + \frac{C'}{\delta} |f(\hat{\rho}_j) - f(\rho_{j,0})| + \frac{C'}{\delta} |f(\rho_{2,0}) - f(\hat{\rho}_2)| \\
&= \frac{C'}{\delta} (|f(\hat{\rho}_j) - f(\rho_{j,0})| + |f(\rho_{1,0}) - f(\hat{\rho}_1)| + |f(\rho_{2,0}) - f(\hat{\rho}_2)|) \leq 2 \frac{C'}{\delta} TV(f)^-
\end{aligned}$$

with a suitable constant C'. Set $C = 2C'$. □

5.3. Interaction of type I5. We consider interactions of Π-waves with the junctions. Since Π-waves always have positive speed, they can interact with the junction only from an incoming line.

Lemma 5.11. *Let us consider a junction J and a Π-wave on an incoming line I_i interacting with J. If A is the distributional matrix for J, whose entries are given by*

(3.3)*, then the interaction of the* Π*-wave with* J *modifies only the* i*th column of* A*. Moreover, the variation of the* i*th column is bounded by the* Π*-wave variation.*

Proof. For each $s \in \mathcal{S}$ and a destination $d \in \mathcal{D}$, we denote by $\pi_i^{s,d}$ and $\pi_{i,0}^{s,d}$, respectively, the left and the right states of the Π-wave. Moreover, for every $j \in \{3,4\}$, we denote by $\alpha_{j,i}^-$ and $\alpha_{j,i}^+$, respectively, the entries of the matrix A before and after the interaction of the Π-wave with J. By (3.3), it is clear that if $l \neq i$, then the entries $\alpha_{j,l}$ are not modified. For $l = i$, we have

$$\begin{aligned}|\alpha_{j,i}^+ - \alpha_{j,i}^-| &= \left| \sum_{s\in\mathcal{S},d\in\mathcal{D}} \pi_i^{s,d}\alpha_J^{i,s,d,j} - \sum_{s\in\mathcal{S},d\in\mathcal{D}} \pi_{i,0}^{s,d}\alpha_J^{i,s,d,j} \right| \\ &\leq \sum_{s\in\mathcal{S},d\in\mathcal{D}} |\pi_i^{s,d} - \pi_{i,0}^{s,d}|\alpha_J^{i,s,d,j}.\end{aligned}$$

This completes the proof. □

LEMMA 5.12. *Let us consider a junction* J *and a* Π*-wave on an incoming line* I_i *interacting with* J*. Then there exists* $C > 0$*, such that the variation of the fluxes is bounded by* C *times the* Π*-wave variation, i.e.,*

$$TV(f)^+ \leq CTV(\Pi)^-.$$

Proof. For simplicity let us consider the case $P \in \Omega$, where

$$\Omega = \left\{(\gamma_1,\gamma_2) \in \Omega_1 \times \Omega_2 : A(\gamma_1,\gamma_2)^T \in \Omega_3 \times \Omega_4\right\}.$$

Since the solution of the Riemann problem depends on the position of the traffic distribution line r_α we consider

$$\begin{aligned}|A(\pi)\gamma_{inc}^T - A(\hat{\pi})\gamma_{inc}^T| &= |(A(\pi) - A(\hat{\pi}))\gamma_{inc}^T| \\ &= \left|\begin{pmatrix} \alpha_{3,1}(\pi) - \alpha_{3,1}(\hat{\pi}) & \alpha_{3,2}(\pi) - \alpha_{3,2}(\hat{\pi}) \\ \alpha_{4,1}(\pi) - \alpha_{4,1}(\hat{\pi}) & \alpha_{4,2}(\pi) - \alpha_{4,2}(\hat{\pi}) \end{pmatrix}\begin{pmatrix}\gamma_1 \\ \gamma_2\end{pmatrix}\right| \\ &= |(\alpha_{3,1}(\pi) - \alpha_{3,1}(\hat{\pi}))\gamma_1 + (\alpha_{3,2}(\pi) - \alpha_{3,2}(\hat{\pi}))\gamma_2, \\ &\quad (\alpha_{4,1}(\pi) - \alpha_{4,1}(\hat{\pi}))\gamma_1 + (\alpha_{4,2}(\pi) - \alpha_{4,2}(\hat{\pi}))\gamma_2|\end{aligned}$$

$$= |(\alpha_{3,1}(\pi) - \alpha_{3,1}(\hat{\pi}), \alpha_{4,1}(\pi) - \alpha_{4,1}(\hat{\pi}))\gamma_1 + (\alpha_{3,2}(\pi) - \alpha_{3,2}(\hat{\pi}), \alpha_{4,2}(\pi) - \alpha_{4,2}(\hat{\pi}))\gamma_2|$$

$$\leq \gamma_1|(\alpha_{3,1}(\pi) - \alpha_{3,1}(\hat{\pi}), \alpha_{4,1}(\pi) - \alpha_{4,1}(\hat{\pi}))| + \gamma_2|(\alpha_{3,2}(\pi) - \alpha_{3,2}(\hat{\pi}), \alpha_{4,2}(\pi) - \alpha_{4,2}(\hat{\pi}))|$$

$$\begin{aligned}&= \gamma_1\left|\left(\sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d})\alpha_J^{1,s,d,3}, \sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d})\alpha_J^{1,s,d,4}\right)\right| \\ &\quad + \gamma_2\left|\left(\sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d})\alpha_J^{2,s,d,3}, \sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d})\alpha_J^{2,s,d,4}\right)\right| \\ &= \gamma_1\left|\sum_{s\in\mathcal{S},d\in\mathcal{D}} \left((\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d})\alpha_J^{1,s,d,3}, (\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d})\alpha_J^{1,s,d,4}\right)\right| \\ &\quad + \gamma_2\left|\sum_{s\in\mathcal{S},d\in\mathcal{D}} \left((\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d})\alpha_J^{2,s,d,3}, (\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d})\alpha_J^{2,s,d,4}\right)\right|\end{aligned}$$

$$\begin{aligned}
&\gamma_1 \left| \sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d})(\alpha_J^{1,s,d,3}, \alpha_J^{1,s,d,4}) \right| \\
&+\gamma_2 \left| \sum_{s\in\mathcal{S},d\in\mathcal{D}} (\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d})(\alpha_J^{2,s,d,3}, \alpha_J^{2,s,d,4}) \right| \\
\le \gamma_1 & \sum_{s\in\mathcal{S},d\in\mathcal{D}} |\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d}||(\alpha_J^{1,s,d,3}, \alpha_J^{1,s,d,4})| + \gamma_2 \sum_{s\in\mathcal{S},d\in\mathcal{D}} |\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d}||(\alpha_J^{2,s,d,3}, \alpha_J^{2,s,d,4})| \\
= & \sum_{s\in\mathcal{S},d\in\mathcal{D}} (\gamma_1 |\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d}||(\alpha_J^{1,s,d,3}, \alpha_J^{1,s,d,4})| + \gamma_2 |\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d}||(\alpha_J^{2,s,d,3}, \alpha_J^{2,s,d,4})|) \\
&\le C \sum_{s\in\mathcal{S},d\in\mathcal{D}} (|\pi_{1,0}^{s,d} - \hat{\pi}_1^{s,d}| + |\pi_{2,0}^{s,d} - \hat{\pi}_2^{s,d}|)
\end{aligned}$$

for some constant C. □

5.4. Existence of solutions for equilibria perturbations. Let us consider an admissible network $(N, \mathcal{I}, \mathcal{F}, \mathcal{J}, \mathcal{S}, \mathcal{D}, \mathcal{R})$. We have the following proposition.

PROPOSITION 5.13. *Let $(\bar{\rho}, \bar{\Pi})$ be an equilibrium on the whole network such that $f(\bar{\rho}) > \delta > 0$. Define $\hat{\lambda} = \max\{f'(0), -f'(1)\}$ and*

$$\Delta t = \frac{\min_i(b_i - a_i)}{\hat{\lambda}},$$

which represents the minimum time for a wave to go from one junction to another. For $0 < \varepsilon < \delta/\hat{\lambda}$ there exists $\tilde{t} = \tilde{t}(\delta, \varepsilon)$ such that the following holds. For every perturbation $(\tilde{\rho}, \tilde{\Pi})$ of the equilibrium with

$$\|\tilde{\rho}\|_{BV} \le \varepsilon, \left\|\tilde{\Pi}\right\|_{BV} \le \varepsilon$$

and

$$\|\tilde{\rho} - \bar{\rho}\|_\infty \le \varepsilon, \left\|\tilde{\Pi} - \bar{\Pi}\right\|_\infty \le \varepsilon$$

there exists an admissible solution (ρ, Π) defined for every $t \in [0, \tilde{t}]$ with initial datum $(\tilde{\rho}, \tilde{\Pi})$.

Proof. Denote with (ρ_ν, Π_ν) a sequence of approximate wave-front tracking solutions with initial data approximating $(\tilde{\rho}, \tilde{\Pi})$. Let us introduce the following notation:

$$\begin{aligned}
TV(f(\rho_\nu(k\Delta t, \cdot))) &= Tf_k, \\
TV(\Pi_\nu(k\Delta t, \cdot)) &= T\Pi_k.
\end{aligned}$$

For every interaction of a wave with a junction we have the estimates of Lemmas 5.9, 5.10, and 5.12; therefore

$$\begin{aligned}
Tf_k &\le Tf_{k-1} + CT\Pi_{k-1}, \\
T\Pi_k &\le T\Pi_{k-1} + \frac{C}{\delta_k} Tf_{k-1},
\end{aligned}$$

where $\delta_k = \delta_{k-1} - TVf_{k-1}$ and δ_0 is such that $f(\tilde{\rho}) > \delta_0 > 0$ (notice that $\delta_0 > \delta - \hat{\lambda}\epsilon$). Setting

$$\begin{aligned}
T_k &= \max_k(Tf_k, T\Pi_k), \\
\delta_k &= \delta_{k-1} - T_k,
\end{aligned}$$

we obtain

$$T_k \leq \left(\frac{C}{\delta_k}+1\right) T_{k-1}.$$

The exact computation of the existence time interval $[0,\tilde{t}]$ is a bit involved, thus we assume, for simplicity, that δ is small, and consider a continuous evolution. Defining $\delta(t)=\delta_0-T(t)$ we obtain

$$\dot{T}(t) \leq \frac{C}{\delta}T(t)=\frac{CT(t)}{\delta_0-T(t)},$$

from which we get $\delta_0 \ln(T)-T=\delta_0\ln(T_0)+CT-T_0$, which implicitly defines $T=T(t,\delta_0,T_0)$. Define $\hat{t}$ such that $T(\hat{t},\delta_0,T_0)=+\infty$; then for $t\leq\tilde{t}=\hat{t}/2$ there exists a constant $C_1>0$ such that

$$TV(f(\rho_\nu(t,\cdot)))\leq C_1$$
$$TV(\Pi_\nu(t,\cdot))\leq C_1$$

uniformly in ν.

Now, by the Helly theorem, Π_ν and $f(\rho_\nu)$ converge by subsequences strongly in L^1. Moreover, again by subsequences, ρ_ν converges weakly in L^1_{loc}. We then can complete the proof as in [7]. □

REFERENCES

[1] D. Alderson, H. Chang, M. Roughan, S. Uhlig, and W. Willinger, *The many facets of internet topology and traffic*, Netw. Heterog. Media, 1 (2006), pp. 569–600.

[2] D. Armbruster, P. Degond, and C. Ringhofer, *A model for the dynamics of large queuing networks and supply chains*, SIAM J. Appl. Math., 66 (2006), pp. 896–920.

[3] F. Baccelli, A. Chaintreau, D. De Vleeschauwer, and D. McDonald, *HTTP turbulence*, Netw. Heterog. Media, 1 (2006), pp. 1–40.

[4] F. Baccelli, D. Hong, and Z. Liu, *Fixed Point Methods for the Simulation of the Sharing of a Local Loop by Large Number of Interacting TCP Connections*, Technical report RR-4154, INRIA, Le Chesnay Cedex, France, 2001.

[5] C. Bardos, A. Y. Le Roux, and J. C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[6] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[7] G. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[8] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 1999.

[9] C. Daganzo, *Fundamentals of Transportation and Traffic Operations*, Pergamon-Elsevier, Oxford, UK, 1997.

[10] C. D'Apice, R. Manzo, and B. Piccoli, *Packet flow on telecomunication networks*, SIAM J. Math. Anal., 38 (2006), pp. 717–740.

[11] M. Garavello and B. Piccoli, *Traffic Flow on Networks*, AIMS Series Appl. Math. 1, American Institute of Mathematical Sciences, Springfield, MO, 2006.

[12] M. Garavello and B. Piccoli, *Source-destination flow on a road network*, Commun. Math. Sci., 3 (2005), pp. 261–283.

[13] M. Herty and A. Klar, *Modeling, simulation, and optimization of traffic flow networks*, SIAM J. Sci. Comput., 25 (2003), pp. 1066–1087.

[14] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[15] F. Kelly, A. K. Maulloo, and D. K. H. Tan, *Rate control in communication networks: Shadow prices, proportional fairness and stability*, J. Oper. Res. Soc., 49 (1998), pp. 237–252.

[16] M. J. Lighthill and G. B. Whitham, *On kinetic waves.* II. *Theory of traffic flows on long crowded roads*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 229 (1955), pp. 317–345.

[17] G. F. Newell, *Traffic Flow on Transportation Networks*, MIT Press, Cambridge, MA, 1980.

[18] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.

[19] D. Serre, *Systems of Conservation Laws* I, Cambridge University Press, Cambridge, UK, 1999.

[20] D. Serre, *Systems of Conservation Laws* II, Cambridge University Press, Cambridge, UK, 2000.

[21] A. S. Tanenbaum, *Computer Networks*, Prentice–Hall, Upper Saddle River, NJ, 2003.

[22] W. Willinger and V. Paxson, *Where mathematics meets the Internet*, Notices Amer. Math. Soc., 45 (1998), pp. 961–970.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1004–1025

STATIONARY SOLUTIONS OF A MODEL FOR THE GROWTH OF TUMORS AND A CONNECTION BETWEEN THE NONNECROTIC AND NECROTIC PHASES*

H. BUENO†, G. ERCOLE†, AND A. ZUMPANO†

Abstract. Results of existence of stationary solutions are proved for a problem modeling the growth of a spheroid tumor in absence of inhibitor agents, for both the nonnecrotic and necrotic cases. The results obtained for the nonnecrotic case are used to prove the existence of stationary solutions for the necrotic case, thus clarifying the connection between both cases. Some bounds for the inner and external radii of the necrotic tumor are given. We also discuss the critical nutrient concentration that determines the necrotic phase.

Key words. necrotic and nonnecrotic tumors, stationary solution

AMS subject classifications. 35B40, 35B35, 35K57, 35R35, 92C50

DOI. 10.1137/060654815

1. Introduction.

In contrast to the exceptional development in the modeling aspects of tumor growth observed in recent years, rigorous mathematical analysis of the behavior of the models considered is rare in the literature. This is by no means surprising: the more the interrelated processes presented in cancer growth are taken into account by the model, the harder the mathematical analysis of the model becomes. (For a historical development of mathematical cancer modeling, see the review article by Araujo and McElwain [2].)

The aim of this paper is to present a rigorous study of a variation on an early model proposed by Byrne and Chaplain, introduced in [5] for the nonnecrotic case and then in [6] for its necrotic version. Here we investigate the transition from the nonnecrotic phase to the necrotic phase of the tumor and show how the stationary solutions of the nonnecrotic model impart information about the stationary solutions of the necrotic model.

In order to better situate the aim and scope of this paper, we describe briefly the model we consider. It consists of a spherical mass of cells of radius $R(t)$ (growing with time t), whose center, or necrotic core, contains only dead cells and is bounded by an inner sphere of radius $\rho(t) \geq 0$, while its surrounding symmetric annulus is composed of proliferating cells. In radial coordinates with $r = |x|$, this intermediary region $\rho(t) < r < R(t)$ receives nutrients not only by diffusion, but also through a developed network of capillary vessels, which is typical of in vivo cancer growth.

The *first equation* of the model is a reaction-diffusion equation for the nutrient concentration $\sigma(r, t)$, which we present in dimensionless form:

$$\epsilon\sigma_t = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\sigma}{\partial r}\right) - f(\sigma)H(r-\rho(t)); \quad 0 < r < R(t), \quad t > 0. \tag{1.1}$$

*Received by the editors May 22, 2006; accepted for publication (in revised form) November 12, 2007; published electronically February 6, 2008.
http://www.siam.org/journals/siap/68-4/65481.html

†Departamento de Matemática - ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31.270-901, Brazil (hamilton@mat.ufmg.br, grey@mat.ufmg.br, zumpano@mat.ufmg.br). The authors were supported in part by CNPq-Brazil.

The coefficient $\epsilon = T_d/T_g$ is the ratio between the time necessary for the diffusion of nutrients T_d and the time elapsed until the tumor doubles its size. Typical values for T_d and T_g are, respectively, on the order of minutes and days. Thus, $\epsilon \approx 0$. Since we suppose that there is no blood supply within the necrotic core, the reaction term has the form $f(\sigma)H(r-\rho(t))$, where H is the Heaviside function (that is, $H(\xi) = 0$ for all $\xi \leq 0$ and $H(\xi) = 1$ for all $\xi > 0$). The function $f(\sigma) = C(\sigma) - V(\sigma)$ is the difference between the consumption of nutrients by the cells of the tumor, which is given by $C(\sigma)$, and the transference of nutrients $V(\sigma)$ between the tumor and the vasculature. This transference usually occurs from the vasculature to the tumor; that is, the tumor receives an extra supply of nutrients. (A detailed discussion on the absorption rate can be found, e.g., in [4]. Our hypotheses and further discussion on f are at the end of this section.)

In order to present the initial and boundary data for (1.1), we distinguish two cases. The first one, $\rho(t) \equiv 0$, corresponds to the *nonnecrotic* phase of the tumor, when there is no necrotic core. In this case, the initial conditions for the unknowns $R(t)$ and $\sigma(r,t)$ are, respectively,

$$R(0) = R_0 \quad \text{and} \quad \sigma(r,0) = \sigma_0(r), \quad 0 < r < R_0, \tag{1.2}$$

which describe the initial radius of the tumor and the initial concentration of nutrients on the tumor; the boundary conditions for σ are

$$\sigma_r(0,t) = 0 \quad \text{and} \quad \sigma(R(t),t) \equiv \bar{\sigma}, \quad t > 0. \tag{1.3}$$

Since $\sigma = \sigma(r,t)$ with $r = |x|$, $\nabla_x \sigma = \frac{\partial \sigma}{\partial r}\frac{x}{|x|}$, the boundary condition $\sigma_r(0,t) = 0$ is natural, meaning that the radial component σ_r of the field σ is null at the center of the tumor; mathematically, this condition is necessary to guarantee differentiability at the origin. We also make the natural supposition that the concentration of nutrients external to the tumor is constant, denoted by $\bar{\sigma}$. For compatibility, the initial data must satisfy $\sigma_0'(0) = 0$ and $\sigma_0(R_0) = \bar{\sigma}$.

The second case describes the *necrotic* phase of the tumor: we suppose that $\rho(0) > 0$. Aggressive tumors usually pass from the nonnecrotic phase to the necrotic phase. There are several causes for this behavior: we will suppose that the local concentration of nutrients is insufficient to sustain an individual cell.

The initial and boundary conditions for the necrotic phase of the tumor are

$$\begin{cases} 0 \leq \rho(0) < R(0), \\ \sigma(r,0) = \sigma_{\text{nec}} \ \text{ if } \ 0 \leq r \leq \rho(0), \\ \sigma(R(0),0) = \bar{\sigma}, \\ \sigma(r,t) \equiv \sigma_{\text{nec}} \ \text{ if } \ 0 \leq r \leq \rho(t), \\ \sigma_r(\rho(t),t) = 0, \ \sigma(\rho(t),t) \equiv \sigma_{\text{nec}}, \ \sigma(R(t),t) \equiv \bar{\sigma} \ \text{ for } \ t > 0. \end{cases} \tag{1.4}$$

Since all the cells in the necrotic core are dead, we admit that there is no absorption of nutrients in that region and that, on its border, the concentration of nutrients satisfies $\sigma(\rho(t),t) = \sigma_{\text{nec}} < \bar{\sigma}$. The experimentally obtained constant σ_{nec} is a threshold level below which the cells cannot survive. So, necrosis occurs when $\sigma(r,t) < \sigma_{\text{nec}}$ and cell proliferation is possible only if $\sigma(r,t) > \sigma_{\text{nec}}$. Because the initial data is constant, from the maximum principle we infer that $\sigma(r,t) = \sigma_{\text{nec}}$ for all $0 \leq r \leq \rho(t)$, $t \geq 0$.

The *second equation* of the model

$$\begin{aligned} R^2\frac{dR}{dt} &= \int_0^{R(t)} \big[S\big(\sigma(r,t)\big)H\big(r-\rho(t)\big) - N\big(\sigma(r,t)\big)H\big(\rho(t)-r\big)\big]r^2dr \\ &= \int_{\rho(t)}^{R(t)} S\big(\sigma(r,t)\big)r^2dr - \int_0^{\rho(t)} N\big(\sigma(r,t)\big)r^2dr \\ &= \int_{\rho(t)}^{R(t)} S\big(\sigma(r,t)\big)r^2dr - \mu\frac{\rho^3(t)}{3} \end{aligned} \tag{1.5}$$

describes the evolution of the tumor radius $R(t)$ and is obtained by applying mass balance equations with adequate constitutive laws and simplifying hypotheses to obtain a closed system (see [1] for a detailed deduction). It incorporates two terms: the *proliferation rate* $S(\sigma)$, which is the balance between birth (mitosis) and natural death rates of the cells (in the region $\rho(t) < r < R(t)$), and the term $N(\sigma)$, which measures necrosis and hypoxic apoptosis, i.e., death rates caused by deficiency of nutrients in the necrotic core. Since we have $\sigma \equiv \sigma_{\rm nec}$ on the necrotic core, it is natural to assume that $N(\sigma) \equiv \mu$, where μ stands for a constant. (Other types of necrosis and apoptosis are not taken into account by the model.) Of course, in the nonnecrotic phase of the tumor, we have $\rho(t) \equiv 0$. (In the nonnecrotic case, a broader discussion on the proliferation rate can be found in [4]. Our hypotheses on S are at the end of this section.)

A *quasi-stationary* solution of both problems (nonnecrotic and necrotic) is a solution of the respective problem in the case $\epsilon = 0$, maintaining dependence on time t (see [4]).

In the nonnecrotic phase of the tumor ($\rho(t) \equiv 0$), an *evolutionary solution* is a C^2-function $\sigma(r,t)$ that satisfies (1.1)–(1.3), (1.5). On its turn, a *stationary solution* is an equilibrium configuration in which the radius of the tumor stabilizes; it is obtained as a pair $(\sigma_R(r), R)$ such that both the concentration of nutrients $\sigma_R(r)$ and the stabilizing radius R do not depend on the time t.

A triple $(\sigma(r,t), \rho(t), R(t))$ is an evolutionary solution in the necrotic phase if $\sigma(r,t)$ is a C^2-function in each region (necrotic and nonnecrotic) which satisfies (1.1), (1.4), and (1.5) and has a C^1-contact on the boundary $r = \rho(t)$; as before, a stationary solution is a triple $(\sigma(r), \rho_{\rm nec}, R_{\rm nec})$ such that the concentration of nutrients $\sigma(r)$ and both the inner and the outer radii $\rho_{\rm nec}$ and $R_{\rm nec}$ do not depend on time.

In this paper we deal only with stationary solutions. Initially, our approach consists in establishing realistic hypotheses on the absorption rate $f(\sigma)$ and on the proliferation rate $S(\sigma)$ that imply the existence of a stationary solution. Then, we obtain bounds for the outer and inner radii of the tumor. In the nonnecrotic case, our investigation deals with rather general nonlinear absorption and proliferation rates, and we focus our analysis on different types of proliferation rates. In the necrotic case, for simplicity, we assume that the proliferation rate $S(\sigma)$ is a generic, continuous, and increasing function. As mentioned before, our ultimate intention is to investigate how the stationary solutions of the nonnecrotic model give information about the stationary solutions of the necrotic model. In both the nonnecrotic and necrotic cases, our approach naturally induces a simple numerical analysis of the problem, conjugating an iterative process with a finite difference method.

In spite of the simplicity of the model considered here, its mathematical analysis is interesting for various reasons: (a) detailed examination of phenomena not yet considered in tumor growth often start with this model [9, 13], which is, in a certain sense, generic to the area; (b) more elaborated models reduce to it under appropriate hypotheses [1, 7, 8, 11]; (c) the nonnecrotic model reproduces patterns observed in tumors cultured in vitro; (d) since no explicit stationary solution is available in the case of general absorption and proliferation rates, a rigorous proof of the existence of stationary solutions is an open question.

The outline of the article is as follows. The initial phase of the tumor (i.e., nonnecrotic) is considered in section 2. In the case of *linear* absorption and proliferation rates, the first theoretical results about the nonnecrotic model treated here were pointed out by Byrne and Chaplain [5] and rigorously proved by Friedman and Reitich [14]. However, for linear absorption and proliferation rates, the existence of a stationary solution is not a problem in itself, since an explicit stationary solution is available. (We stress that [14] is mainly devoted to the stability of the stationary solution.) The first subsection of section 2 establishes results of existence, uniqueness, and localization of stationary solutions for more realistic absorption and proliferation rates. Subsection 2.2 comments on some results obtained by Byrne and Chaplain [5], and subsection 2.3 compares our outcomes, by means of a numerical implementation, with those in [14].

This section poses, however, a simple question: Is the stationary solution (when it exists) a limit of solutions of the evolutionary problem? It is proved in [4] that solutions $\sigma(r,t)$ of the quasi-stationary problem (i.e., if we consider $\epsilon = 0$) converge monotonically to the stationary solution, which is also presented in this paper. This result gives insight about what can be expected for small $\epsilon > 0$ (a condition satisfied by real data). For example, let us consider a nonnecrotic tumor which has stabilizing radius R^* (a value that can be numerically computed, when the absorption and proliferation rates are known). If ϵ is small enough, this would imply that, if the tumor has radii $R_0 = R(t_0) < R_*$ for some time $t = t_0$ and $R_1 = R(t_1) > R_*$ for $t = t_1 > t_0$, then either no stationary solution will be achieved or the tumor will pass to the necrotic phase.

Section 3 studies the relationship between the stationary solutions of the nonnecrotic and necrotic models. This connection has been set aside in the mathematical handling of tumor growth. However, it turns out that this relationship allows us to predict if it is possible to achieve a stationary solution for the necrotic tumor. We show the existence of a value σ_*, which determines the existence of stationary solutions. Consequently, if the tumor is in the necrotic phase, the stationary solution exists if and only if $\sigma_* < \sigma_{\text{nec}} < \bar{\sigma}$; it exists in the nonnecrotic phase if and only if $\sigma_0 < \sigma_{\text{nec}} \leq \sigma_*$, where σ_0 stands for the zero of the absorption rate. In other words, if the phase of the tumor and the value of σ_{nec} are known, we can say whether it is possible to achieve a stationary solution. In both cases these solutions are unique. Under additional conditions on the behavior of the proliferation rate, we also present some estimates on the radii of the stabilizing tumor. (We emphasize that the value σ_{nec} can be empirically established.)

Our conclusions are presented in section 4, while most of the proofs of our results are deferred to section 5.

We now describe succinctly our hypotheses on the absorption rate $f(\sigma)$ and on the proliferation rate $S(\sigma)$. A more detailed discussion of these functions can be found in [4].

In the majority of works, the functions f and S depend linearly on σ. Under these hypotheses, existence and global stability of the nontrivial stationary solutions are proved, respectively, in [14] and [10] for the nonnecrotic and the necrotic cases. The approach used there rests on the existence of an explicit form for the stationary solution. Here our assumptions on both f and S are more general, and such an explicit solution is not available.

In this work we consider nonlinear absorption and proliferation rates. This was done not merely for generality reasons. Linearization of the absorption rate is just an approximation of real data, while the choice of an adequate proliferation rate is necessary for the evaluation of the mathematical model considered. Furthermore, nonlinearities of both functions might be considered as indirect effects of inhibitors on the absorption and proliferation rates if we are not able to mathematically describe their action; a hypothesis now in discussion concerns the inhibitory effects of certain nutrients on cancer growth.

We assume that the function $f(\sigma) = C(\sigma) - V(\sigma)$ is such that the consumption of nutrients by the tumor is an increasing function of σ, while the transference rate, as mentioned before, is a positive function. It is natural to assume that $V'(\sigma) \leq 0$, because the diminution of the concentration of nutrients in the tumor increases the transference rate as a compensation mechanism. We also assume that $f(\sigma) \geq 0$, which means that $C(\sigma) \geq V(\sigma)$, and we accept that $f(\sigma) = 0$ happens at a single value σ_0. So, we suppose that the absorption rate is increasing for $\sigma \in (\sigma_0, \bar{\sigma}]$. Therefore, if $\sigma > \sigma_0$, the consumption rate is greater than the transference rate, and the difference between these rates decreases when the concentration of nutrients decreases. Our assumptions on f appear to be natural and are similar to those of [3] if one considers only proliferating cells.

In the nonnecrotic phase, we acknowledge that S is continuous and either that it is monotonic on the interval $[\sigma_0, \bar{\sigma}]$ or that it assumes only one local extreme on this interval. But, supposing that the number of cells is growing when the nutrient concentration is $\bar{\sigma}$, it follows that $S(\bar{\sigma}) > 0$. And assuming that S increases in a neighborhood of $\bar{\sigma}$, our analysis is done when $S(\sigma)$ is either increasing on $[\sigma_0, \bar{\sigma}]$ or decreasing on $[\sigma_0, \Lambda]$ and increasing on $[\Lambda, \bar{\sigma}]$, Λ denoting the local minimum of S. These hypotheses include those made in the papers already cited (in some cases, if one considers only proliferating cells).

In the necrotic case, to render the presentation less complex, we assume that S is continuous and increasing on the interval $[\sigma_0, \bar{\sigma}]$.

2. The nonnecrotic problem. Putting $\lambda := R^2$, the change of variables $\sigma_\lambda(r) = \sigma_R(rR)$ transforms the system (1.1)–(1.3), (1.5) into the boundary value problem (BVP)

$$\begin{cases} (r^2\sigma_\lambda')' = \lambda r^2 f(\sigma_\lambda), \quad 0 < r < 1, \\ \sigma_\lambda'(0) = 0, \quad \sigma_\lambda(1) = \bar{\sigma}, \end{cases} \tag{2.1}$$

and the integral equation

$$\lambda \int_0^1 S(\sigma_\lambda(r)) r^2 dr = 0. \tag{2.2}$$

Hence, a stationary solution for the system (1.1)–(1.3), (1.5) is equivalent to a pair $(\sigma_\lambda, \lambda)$ satisfying (2.1)–(2.2). We note that $(\bar{\sigma}, 0)$ is a trivial stationary solution.

We will first solve (2.1) for all $\lambda > 0$ and then consider λ as a parameter in order to find a solution for (2.2).

Remark 1. Since σ_{nec} is an experimentally obtained value, the relation between σ_0 and σ_{nec} is unclear: it may be, for example, $\sigma_0 < \sigma_{\text{nec}} < \bar{\sigma}$. Let the pair $(\sigma_{\lambda_*}, \lambda_*)$

denote a solution of the problem (2.1)–(2.2). Since (2.1) and (2.2) are handled separately, the stationary solution for the nonnecrotic model makes sense only when $\sigma_{\lambda_*}(r)$ is above σ_{nec}. This condition is fulfilled, for instance, if $\sigma_{\text{nec}} \le \sigma_0 < \bar{\sigma}$; as we shall see, $\sigma_\lambda(r) > \sigma_0$ for all $\lambda \ge 0$.

2.1. Stationary solutions for the nonnecrotic model. In this subsection we collect some properties of the solution of (2.1)–(2.2) and present some estimates on the stabilizing radius. The technical proofs are found in section 5.

PROPOSITION 2.1. *Suppose that* $f(\sigma) \in C^1([\sigma_0, \bar{\sigma}])$ *is increasing and vanishes only at* $\sigma = \sigma_0$. *Then, for each* $\lambda \ge 0$, *there exists exactly one solution* σ_λ *of the BVP* (2.1), *which satisfies*

$$\sigma_0 < \sigma_\lambda(r) \le \bar{\sigma} \quad \textit{for all } r \in [0,1]. \tag{2.3}$$

Furthermore,

(i) σ_λ *is strictly increasing;*

(ii) σ_λ *is strictly convex, and*

$$\sigma''_\lambda(r) \ge \frac{\sigma'_\lambda(r)}{r} > 0 \quad \textit{for all } r \in (0,1]; \tag{2.4}$$

(iii) *if* $0 \le \lambda_1 < \lambda_2$, *then* $\sigma_{\lambda_2}(r) \le \sigma_{\lambda_1}(r)$ *for all* $r \in [0,1]$*;*

(iv) *the map* $\lambda \mapsto \sigma_\lambda$ *is continuous; and*

(v) $\lim_{\lambda\to\infty} \sigma_\lambda(r) = 0$ *pointwise in* $[0,1)$ *and uniformly in closed intervals contained in* $[0,1)$.

The properties of u_λ (monotonicity with respect to r and λ, convexity, and convergence to σ_0) are displayed in Figure 2.1.

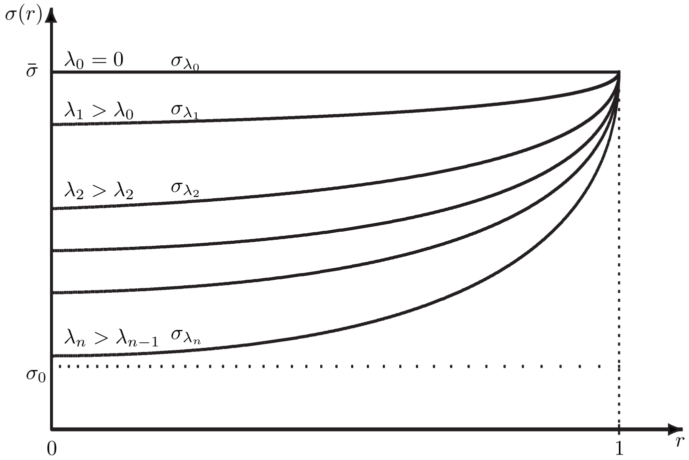


FIG. 2.1. *For each* $\lambda \ge 0$, *the functions* σ_λ *are convex and increasing. The sequence* $\{\sigma_{\lambda_n}\}$ *converges uniformly to* σ_0 *in compact subsets of* $[0,1)$ *when* $\lambda_n \to \infty$.

The uniqueness assertion in Proposition 2.1 rests on a form of the maximum principle (see section 5), while existence is based on the method of sub- and supersolutions. A numerical implementation of this method is very simple (see the remark below and subsection 2.3). The convexity of the solutions u_λ makes it possible to handle nonmonotonic proliferation rates.

Remark 2. The method of sub- and supersolutions (see [12]) is valid under more general hypotheses on f—for example, if $f \in C([\sigma_0, \bar{\sigma}])$ and if there exists a constant k such that $f(\sigma) - k\sigma$ is decreasing (a condition fulfilled by Lipschitz continuous f). Its numerical implementation is simple (see also subsection 2.3): the method produces solutions of (2.1) given as limits of two iterative sequences of the form $\{\sigma_n\}_{n=1}^{\infty}$, where

$$\begin{aligned} &(r^2\sigma'_{n+1})' - k\lambda r^2 \sigma_{n+1} = \lambda r^2 \left(f(\sigma_n) - k\sigma_n\right), \\ &\sigma'_{n+1}(0) = 0, \qquad \sigma_{n+1}(1) = \bar{\sigma}, \end{aligned}$$

starting at $\sigma_1(r) \equiv \sigma_0$ and $\sigma_1(r) = \bar{\sigma}$, respectively. (Taking into account the uniqueness of solutions, both sequences have the same limit.)

Let the pair $(\sigma_{\lambda_*}, \lambda_*)$ denote a solution of the problem (2.1)–(2.2) such that $\lambda_* > 0$. Then λ_* is obtained as a zero of the equation $I(\lambda) = 0$, where the continuous function $I : [0, \infty) \to \mathbb{R}$ is defined by

$$I(\lambda) = \int_0^1 S(\sigma_\lambda(r)) r^2 dr \tag{2.5}$$

and σ_λ is given by Proposition 2.1.

In this case, the pair (σ_{R_*}, R_*) is a stationary solution for the system (1.1)–(1.3), (1.5), where $R_* = \sqrt{\lambda_*}$ is the stabilizing radius and $\sigma_{R_*}(r) = \sigma_{\lambda_*}(r/R_*)$ for all $r \in [0, R_*]$.

The simplest way to ensure the existence of a zero for $I(\lambda)$ is given by the following theorem.

THEOREM 2.2. *If the continuous function $S(\sigma)$ satisfies the condition*

$$S(\sigma_0)S(\bar{\sigma}) < 0, \tag{2.6}$$

then there exists at least one positive value λ_* such that $I(\lambda_*) = 0$. Moreover, if $S(\sigma)$ is also monotonic, then λ_* is unique.

Proof. Since

$$I(0) = \int_0^1 S(\bar{\sigma}) r^2 dr = \frac{S(\bar{\sigma})}{3}$$

and

$$\lim_{\lambda \to \infty} I(\lambda) = \int_0^1 S\left(\lim_{\lambda \to \infty} \sigma_\lambda(r)\right) r^2 dr = \frac{S(\sigma_0)}{3},$$

the existence of λ_* is a consequence of the continuity of the function $I(\lambda)$. Uniqueness follows from the monotonicity of $I(\lambda)$. □

We now handle more general cases. The proof of the next result is based on a geometrical construction that strongly depends on the convexity of σ_λ. (See section 5.)

THEOREM 2.3. *Suppose that, for some $\alpha \in (\sigma_0, \bar{\sigma})$, S is nondecreasing on $[\alpha, \bar{\sigma}]$ and*

$$\int_\alpha^{\bar{\sigma}} S(\sigma)(\sigma - \alpha)^2 d\sigma = 0. \tag{2.7}$$

Then there exists $\lambda_* > 0$ such that $I(\lambda_*) = 0$. Moreover,

$$3\frac{(\bar{\sigma} - \alpha)}{f(\bar{\sigma})} \le \lambda_* \le 6\frac{(\bar{\sigma} - \alpha)}{f(\alpha)}. \tag{2.8}$$

Remark 3. Theorem 2.3 is also valid if $S \in C([\sigma_0, \bar{\sigma}])$ attains a minimum value at $\sigma = \Lambda \in (\sigma_0, \bar{\sigma})$, decreasing if $\sigma \in [\sigma_0, \Lambda]$ and increasing if $\sigma \in [\Lambda, \bar{\sigma}]$. In this situation, if $S(\sigma_0) \leq 0 < S(\bar{\sigma})$, we change (2.7) to

$$\int_{\Lambda}^{\bar{\sigma}} S(\sigma)(\sigma - \Lambda)^2 d\sigma \leq 0.$$

The technique applied in the proof of Theorem 2.3 may also be used to obtain bounds for λ_*, even when S does not satisfy (2.7). As an example we mention the following proposition

PROPOSITION 2.4. *Suppose that the continuous function $S(\sigma)$ is increasing and satisfies $S(\sigma_0) < 0 < S(\bar{\sigma})$. Then*

$$\lambda_* \geq \max\left\{\frac{3(\bar{\sigma} - \alpha)}{f(\bar{\sigma})}, \frac{6(\bar{\sigma} - \beta)}{f(\bar{\sigma})}\right\}, \tag{2.9}$$

where

$$\alpha = \min\left\{\xi \in [\sigma_0, \bar{\sigma}] : \int_{\xi}^{\bar{\sigma}} S(\sigma)(\sigma - \xi)^2 d\sigma \geq 0\right\} \tag{2.10}$$

and

$$\beta = \min\left\{\eta \in [\sigma_0, \bar{\sigma}] : \int_{\eta}^{\bar{\sigma}} S(\sigma)\sqrt{\sigma - \eta} d\sigma \geq 0\right\}. \tag{2.11}$$

Now we state the existence of solutions for (2.1)–(2.2) when neither (2.6) nor (2.7) is satisfied. From now on we assume that the proliferation rate $S \in C([\sigma_0, \bar{\sigma}])$ attains a minimum value at $\sigma = \Lambda \in (\sigma_0, \bar{\sigma})$, decreasing if $\sigma \in [\sigma_0, \Lambda]$ and increasing if $\sigma \in [\Lambda, \bar{\sigma}]$. The proof of the result below is obtained by refining the technique used to demonstrate Theorem 2.3. (See section 5.)

THEOREM 2.5. *Let S be as above. In addition, suppose that $S(\sigma_0) = 0 < S(\bar{\sigma})$ and*

$$\int_{\sigma_0}^{\bar{\sigma}} S(\sigma) d\sigma \leq 0.$$

Then there exists at least one λ_* such that $I(\lambda_*) = 0$. Moreover,

$$\lambda_* \geq 6\frac{(\bar{\sigma} - \Lambda)}{f(\bar{\sigma})}. \tag{2.12}$$

Remark 4. In the last theorem, we can exchange S for a function that increases from σ_0 to a maximum value Λ and then decreases from Λ to $\bar{\sigma}$. But the form of S stated in the theorem appears to be more natural under the model considered. (See the comments below.)

2.2. Comments. If $\widetilde{\sigma}$ stands for a constant, the logistic form

$$S(\sigma) = s\sigma(1 - \sigma/\widetilde{\sigma}) \tag{2.13}$$

was suggested by Byrne and Chaplain [5], the terms $s\sigma$ ($s > 0$) and $s\sigma^2/\widetilde{\sigma}$ meaning the birth and death rates, respectively.

In this case we have $S(\sigma) > 0$ if $\sigma \in (0, \widetilde{\sigma})$ and $S(\sigma) < 0$ if $\sigma > \bar{\sigma}$. If we have $0 < \sigma_0 < \widetilde{\sigma} < \bar{\sigma}$, there exists a stationary solution, since (2.6) is fulfilled. Assuming that $\sigma_0 = 0$ (which corresponds to an avascular tumor in [5]) and $\bar{\sigma} > \widetilde{\sigma}$, Remark 4 leads to the existence of λ_*, provided that $\int_{\sigma_0}^{\bar{\sigma}} S(\sigma) d\sigma > 0$.

The proliferation rate (2.13) was not analytically studied in [5], but some numerical implementations were done just for comparison with the linear case. In one of these implementations Byrne and Chaplain considered $f(\sigma) = 2\sigma - 1.2$ and $\sigma_0 = \widetilde{\sigma} = 0.6$ and obtained that the evolutionary solutions do not converge to any stationary solution. (See Figure 3 in that paper.)

According to our analysis,

$$0 \leq \sigma_0 < \widetilde{\sigma} < \bar{\sigma} \tag{2.14}$$

is a necessary condition for the existence of stationary solutions, since $\sigma_\lambda(r) \in (\sigma_0, \bar{\sigma}]$ for all $\lambda \geq 0$ and $r \in [0, 1]$. So, our results show that it was predictable that no stationary solution was to be found as a limit of the evolutionary solutions in the numerical implementation done in [5].

But (2.13)–(2.14) does not appear to be natural: it implies that $S(\bar{\sigma}) < 0$, that is, that the proliferation rate is negative when the concentration of nutrients is maximal, and $S(\sigma)$ grows when this concentration decreases, until it reaches the maximum value $S(\widetilde{\sigma}/2)$.

On the other hand, if we suppose a proliferation rate of the form

$$S(\sigma) = s\sigma \left(\frac{\sigma}{\widetilde{\sigma}} - 1 \right),$$

then (2.14) is still a necessary condition for the existence of a stationary solution. Remark 3 then implies the existence of at least one stationary solution if

$$0 < \sigma_0 < \widetilde{\sigma} < \bar{\sigma} \quad \text{and} \quad \int_{\widetilde{\sigma}/2}^{\bar{\sigma}} \sigma \left(\frac{\sigma}{\widetilde{\sigma}} - 1 \right) \left(\sigma - \frac{\widetilde{\sigma}}{2} \right)^2 d\sigma \leq 0$$

or

$$\sigma_0 = 0 < \widetilde{\sigma} < \bar{\sigma} \quad \text{and} \quad \int_{\sigma_0}^{\bar{\sigma}} \sigma \left(\frac{\sigma}{\widetilde{\sigma}} - 1 \right) d\sigma \leq 0.$$

In the first case, we obtain from Theorem 2.3 the estimates

$$3 \frac{(\bar{\sigma} - \alpha)}{f(\bar{\sigma})} < \lambda_* < 6 \frac{(\bar{\sigma} - \alpha)}{f(\alpha)}$$

if $\alpha \in (0, \bar{\sigma})$ is such that

$$\int_{\alpha}^{\bar{\sigma}} \sigma \left(\frac{\sigma}{\widetilde{\sigma}} - 1 \right) (\sigma - \alpha)^2 d\sigma = 0. \tag{2.15}$$

In the second case, we obtain from Theorem 2.5 the lower bound

$$\lambda_* > 6 \frac{(\bar{\sigma} - \widetilde{\sigma}/2)}{f(\bar{\sigma})}.$$

2.3. Numerical implementation. The approach used to solve problem (2.1)–(2.2) naturally induces a numerical method to deal with the problem. The numerical procedure described below was first intended to test our approach and to compare our outputs with results given in other papers for linear rates. Since they were coherent, we have in mind a rigorous analysis of our numerical method.

To exemplify how our theoretical approach leads to a numerical treatment of (2.1)–(2.2), we take $\bar{\sigma} = 1$ and consider quadratic absorption and proliferation rates:

$$f(\sigma) = (\sigma - 0.3)(\sigma + 0.5), \qquad S(\sigma) = 4\sigma \left(\frac{\sigma}{\widetilde{\sigma}} - 1 \right).$$

The values of $\widetilde{\sigma}$ are calculated from the given values of α by the equation

$$\widetilde{\sigma} = \frac{2(\alpha^2 + 3\alpha + 6)}{5(\alpha + 3)}, \quad 0 \le \alpha \le 1,$$

which is equivalent to (2.15). So, to each $\alpha \in [0, 1)$ corresponds a unique $\widetilde{\sigma} \in [0.8, 1)$.

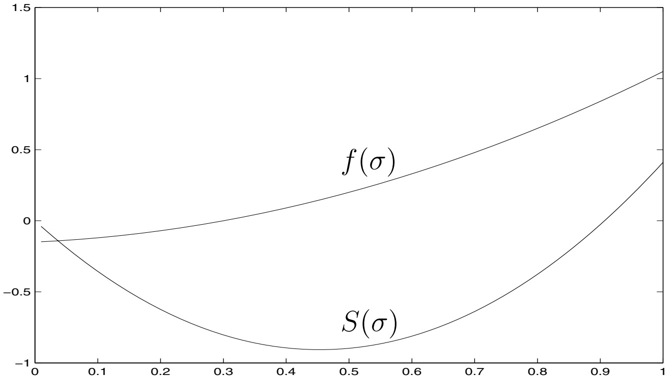


FIG. 2.2. *In the case of $\widetilde{\sigma}$ corresponding to $\alpha = 0.6$, the graphs of $f(\sigma)$ and $S(\sigma)$ are displayed.*

In Figure 2.2, the graphs of $f(\sigma)$ and $S(\sigma)$ are shown, while Figure 2.3 displays $I(\lambda)$ for $\alpha = 0.6$, in the interval $[\lambda_1, \lambda_2]$, where

$$\lambda_1 = \frac{3(\bar{\sigma} - \alpha)}{f(\bar{\sigma})} = \frac{3(1 - \alpha)}{1.05}$$

and

$$\lambda_2 = \frac{6(\bar{\sigma} - \alpha)}{f(\alpha)} = \frac{6(1 - \alpha)}{(\alpha - 0.3)(\alpha + 0.5)}.$$

The zero λ_* of $I(\lambda)$ is to be found in this interval, according to Theorem 2.3, and is confirmed by our numerical outputs. The same coherent behavior was observed for various values of α we have tested.

To approximate the functions $\sigma_* = \sigma_{\lambda_*}$ we have used an iterative method based on super- and subsolutions (see Remark 2), where two numerical sequences are generated, one departing from the subsolution and the other from the supersolution. Both sequences converge to the unique stationary solution (see Proposition 2.1). One

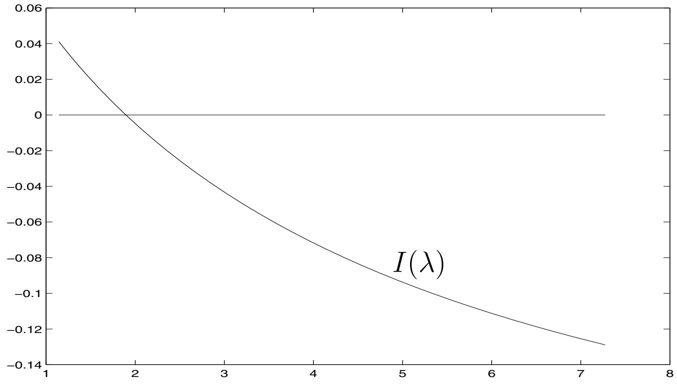


FIG. 2.3. *For $\widetilde{\sigma}$ corresponding to $\alpha = 0.6$, the graph of $I(\lambda)$ is displayed in the interval* $[\lambda_1, \lambda_2]$, *with $\lambda_1 \approx 1.1428$, $\lambda_2 \approx 7.2727$, and $\lambda_* \approx 1.8967$.*

function of these sequences was chosen to represent the solution, when the distance between the functions of each individual sequence and also between correspondent elements of both sequences was less than 10^{-3}. In each iteration, centered finite difference was used in a uniform 100-point grid. A maximum of 9 iterations was necessary in the "critical case" corresponding to the least possible value for α. We point out, however, that we have made no rigorous error analysis for this iterative method.

A simple method was chosen to evaluate $I(\lambda)$—the trapezoidal rule in a 100-point uniform grid. To find the zero λ_* of $I(\lambda)$, the bisection method with approximation of the order of 10^{-3} was used.

We now compare our results with those obtained in [14]. In order to do that, we consider, as in that paper, a linear case:

$$f(\sigma) = \sigma \quad \text{and} \quad S(\sigma) = \sigma - \widetilde{\sigma},$$

with $\sigma_0 = 0$, $\bar{\sigma} = 1$, and $0 \leq \widetilde{\sigma} < 1$. It holds that

$$\alpha = 4\widetilde{\sigma} - 3, \quad \lambda_1 = 12(1 - \widetilde{\sigma}), \quad \text{and} \quad \lambda_2 = \frac{24(1 - \widetilde{\sigma})}{4\widetilde{\sigma} - 3} \quad \text{for} \quad 0.75 < \widetilde{\sigma} < 1.$$

Theorem 2.3 ensures that

$$2.64 \leq \lambda_* \leq 44 \quad \text{if} \quad \widetilde{\sigma} = 0.78$$

and

$$0.48 \leq \lambda_* \leq 1.1429 \quad \text{if} \quad \widetilde{\sigma} = 0.96.$$

Using the numerical approach described above, we find

$$\lambda_* \approx 4.7697 \quad \text{if} \quad \widetilde{\sigma} = 0.78 \tag{2.16}$$

and

$$\lambda_* \approx 0.6363 \quad \text{if} \quad \widetilde{\sigma} = 0.96. \tag{2.17}$$

The value $\eta = \sqrt{\lambda_*}$ is obtained as a solution of an algebraic equation (see [14]), namely,

$$\tanh(\eta) = \frac{\eta}{1+\Lambda\eta^2}, \tag{2.18}$$

where $\Lambda = \widetilde{\sigma}/3 \in (0, 1/3)$. The approximate solutions of (2.18) are

$$\lambda_* = \eta^2 \approx 4.7681 \quad \text{if} \quad \widetilde{\sigma} = 0.78$$

and

$$\lambda_* = \eta^2 \approx 0.6362 \quad \text{if} \quad \widetilde{\sigma} = 0.96,$$

which are very close to our results, given by (2.16) and (2.17).

Friedman and Reitich [14] observed that the constant $\Lambda_{\text{crit}} = 0.2727\ldots$ satisfies

$$\eta(\Lambda) < \frac{1}{\sqrt{\Lambda}} \quad \text{if} \quad \Lambda_{\text{crit}} < \Lambda < 1/3, \tag{2.19}$$

where $\eta(\Lambda)$ denotes the solution of (2.18) for $\Lambda \in (0, 1/3)$. The estimate we have obtained is

$$\eta(\Lambda) \leq 2\sqrt{\frac{3(1-3\Lambda)}{4\Lambda - 1}} \quad \text{if} \quad 0.25 < \Lambda < \frac{1}{3}.$$

Our estimate is not only valid in a larger interval than that described in (2.19) but also better when Λ is very close to $1/3$. For example, when $\Lambda = 0.32$, we have

$$2\sqrt{\frac{3(1-3\Lambda)}{4\Lambda - 1}} \approx 1.3093 < \frac{1}{\sqrt{\Lambda}} \approx 1.76781.$$

3. The necrotic model. In the analysis of the necrotic phase of the tumor, as in the nonnecrotic phase, we suppose that $f(\sigma) > 0$ for $\sigma \in [\sigma_{\text{nec}}, \bar{\sigma}]$. (This means that the rate of consumption overcomes the rate of transference of nutrients from vasculature in that region.) Furthermore, we also assume that f vanishes only at $\sigma_0 \in [0, \sigma_{\text{nec}})$ and is nondecreasing on the interval $[\sigma_0, \bar{\sigma}]$. For simplicity, we assume here that the proliferation rate $S(\sigma)$ is continuous and *increasing* on the interval $[\sigma_0, \bar{\sigma}]$.

A stationary solution for the necrotic model is a triple $(\sigma, \rho_{\text{nec}}, R_{\text{nec}})$ such that $\sigma \in C^1[0,1] \cap C^2[0,\rho] \cap C^2[\rho, 1]$ satisfies $\sigma(r) \equiv \sigma_{\text{nec}}$ for all $r \in (0, \rho)$ and

$$\begin{cases} (r^2\sigma')' = r^2 f(\sigma), \ \rho_{\text{nec}} < r < R_{\text{nec}}, \\ \sigma'(\rho_{\text{nec}}) = 0, \ \sigma(\rho_{\text{nec}}) = \sigma_{\text{nec}}, \ \sigma(R_{\text{nec}}) = \bar{\sigma}, \\ \int_{\rho_{\text{nec}}}^{R_{\text{nec}}} S(\sigma(r)) r^2 dr = \mu(\rho_{\text{nec}})^3. \end{cases} \tag{3.1}$$

(We assume $\sigma_0 < \sigma_{\text{nec}} \leq \bar{\sigma}$; see Remark 1.)

As before, by making the change of variables $r \to rR_{\text{nec}}$, we transform (3.1) into the equivalent boundary value problem

$$\begin{cases} (r^2\sigma')' = \lambda r^2 f(\sigma), \ \rho < r < 1, \\ \sigma'(\rho) = 0, \ \sigma(\rho) = \sigma_{\text{nec}}, \ \sigma(1) = \bar{\sigma}, \end{cases} \tag{3.2}$$

and the integral equation

$$\int_{\rho}^{1} S(\sigma(r))r^2 = \mu\rho^3, \tag{3.3}$$

where $\lambda = R_{\text{nec}}^2$ and $\rho = \rho_{\text{nec}}/R_{\text{nec}} < 1$. Therefore, a stationary solution is still given by a triple $(\sigma(r), \rho, \lambda)$, which solves (3.2)–(3.3).

3.1. Stationary solutions for the necrotic model. In this subsection we give conditions for the existence of solutions of (3.2). Our approach is basically the one used to handle the nonnecrotic case, but now we treat both λ and the inner radius ρ as parameters. The proofs of the next two results are similar to that of Proposition 2.1.

LEMMA 3.1. *Let $\rho \in [0, 1)$ be fixed. Then the following hold:*

(i) *For each $\lambda \in (0, \infty)$, there exists a unique $\sigma_\lambda \in C^2[\rho, 1]$ satisfying*

$$\begin{cases} (r^2\sigma_\lambda')' = \lambda r^2 f(\sigma_\lambda), & \rho < r < 1, \\ \sigma_\lambda'(\rho) = 0, \ \sigma_\lambda(1) = \bar{\sigma}. \end{cases} \tag{3.4}$$

Furthermore, if $\lambda > 0$, then σ_λ is (strictly) increasing and strictly convex and satisfies

$$\sigma_0 < \sigma_\lambda(r) \leq \bar{\sigma} \text{ for all } r \in [\rho, 1]. \tag{3.5}$$

(ii) *The application $\lambda \longmapsto \sigma_\lambda$ from $[0, \infty)$ to $C[\rho, 1]$ is continuous and nonincreasing; that is, $\sigma_{\lambda_1} \geq \sigma_{\lambda_2}$ if $0 \leq \lambda_1 < \lambda_2$.*

(iii) $\lim_{\lambda\to\infty} \sigma_\lambda = \sigma_0$ *uniformly on each closed interval contained in $[\rho, 1)$.*

Since $\sigma_0 < \sigma_{\text{nec}} \leq \bar{\sigma}$, property (iii) of Lemma 3.1 associates each $\rho \in [0, 1)$ to a unique $\lambda = \lambda_\rho$ so that $\sigma_{\lambda_\rho}(\rho) := \sigma_\rho(\rho) = \sigma_{\text{nec}}$. We stress that the localization of the zero σ_0 of the function f is critical for the existence of a stationary solution: it is impossible to find a solution satisfying $\sigma(\rho) = \sigma_{\text{nec}}$, if $\sigma_{\text{nec}} \leq \sigma_0$.

THEOREM 3.2. *For each $\rho \in [0, 1)$ there exist a unique $\lambda_\rho > 0$ and a unique function $\sigma_\rho \in C^2[0, 1] \cap C^2[0, \rho] \cap C^2[\rho, 1]$ such that $\sigma_\rho(r) := \sigma_{\text{nec}}$ for all $r \in [0, \rho)$ and*

$$\begin{cases} (r^2\sigma_\rho')' = \lambda_\rho r^2 f(\sigma_\rho), & \rho < r < 1, \\ \sigma_\rho'(\rho) = 0, \ \sigma_\rho(\rho) = \sigma_{\text{nec}}, \ \sigma_\rho(1) = \bar{\sigma}. \end{cases} \tag{3.6}$$

Furthermore, the maps

$$\rho \in [0, 1) \mapsto \sigma_\rho \in C[0, 1] \quad \textit{and} \quad \rho \in [0, 1) \mapsto \lambda_\rho \in (0, \infty)$$

are continuous and satisfy the following:

(i) *$\rho \mapsto \sigma_\rho$ is nonincreasing:*

$$0 \leq \rho_1 < \rho_2 \quad \Rightarrow \quad \sigma_{\rho_1}(r) \geq \sigma_{\rho_2}(r) \quad \textit{for all } r \in [0, 1].$$

(ii) *$\rho \mapsto \lambda_\rho$ is (strictly) increasing and*

$$\lim_{\rho\to 1} \lambda_\rho = \infty. \tag{3.7}$$

Remark 5. When $\rho = 0$, we will denote λ_ρ by λ_{nec} and the corresponding solution σ_ρ by $\sigma_{\lambda_{\text{nec}}}$. So,

$$(r^2\sigma_{\lambda_{\text{nec}}}')' = \lambda_{\text{nec}} r^2 f(\sigma_{\lambda_{\text{nec}}}), \quad 0 < r < 1,$$
$$\sigma_{\lambda_{\text{nec}}}'(0) = 0, \ \sigma_{\lambda_{\text{nec}}}(0) = \sigma_{\text{nec}}, \ \sigma_{\lambda_{\text{nec}}}(1) = \bar{\sigma}.$$

We stress that $\sigma_{\lambda_{\mathrm{nec}}}$ is also a solution of the boundary value problem (2.1) for $\lambda = \lambda_{\mathrm{nec}}$.

Since the differential equation (3.2) is solved, let us now consider the integral equation (3.3). For this, we define

$$J(\rho) = \int_\rho^1 S(\sigma_\rho(r))r^2 dr - \mu\rho^3 \text{ for } \rho \in [0,1).$$

We will show how the nonnecrotic model brings information to the necrotic model. For this, we assume[1] that the proliferation rate $S(\sigma)$ is a *continuous, nondecreasing function* of the nutrient concentration σ.

Theorem 3.2 implies that the function J is continuous and nonincreasing. The existence of a uniform bound for $S(\sigma_\rho)$ with respect to ρ implies that

$$\lim_{\rho\to 1^-} J(\rho) = -\mu < 0. \tag{3.8}$$

According to Remark 5, we have

$$J(0) = I(\lambda_{\mathrm{nec}}), \tag{3.9}$$

where the function I is defined by (2.5). So, (3.2)–(3.3) has a solution $(\sigma(r), \rho, \lambda)$ with $\rho > 0$ if and only if $I(\lambda_{\mathrm{nec}}) > 0$ (since λ_{nec} corresponds to $\rho = 0$, a nonnecrotic solution).

As a consequence, we have the following result on the existence and uniqueness of stationary solutions.

THEOREM 3.3. *Suppose that* $(\sigma_{\lambda_*}, \lambda_*)$ *is a solution of* (2.1)–(2.2). *Denote* $\sigma_{\lambda_*}(0)$ *by* σ_*. *There exists a (unique) solution of* (3.2)–(3.3) *if and only if* $\sigma_* < \sigma_{\mathrm{nec}}$.

Proof. We have $J(0) = I(\lambda_{\mathrm{nec}}) > 0 = I(\lambda_*)$ if and only if $\lambda_{\mathrm{nec}} < \lambda_*$, an inequality that is equivalent to $\sigma_* < \sigma_{\mathrm{nec}}$. □

Theorem 3.3 depends on the solution of (2.1)–(2.2), which can be numerically obtained. A more intrinsic criterion is given by the following proposition.

PROPOSITION 3.4. *Suppose that* $S(\sigma_0) < 0 < S(\bar\sigma)$. *Let* $\widetilde\sigma$ *denote the zero of the function* S *(that is, the value of the nutrient concentration at which the rates of birth and death are equal). If* $\sigma_{\mathrm{nec}} \geq \widetilde\sigma$, *then there exists a unique solution of* (3.2)–(3.3).

Proof. Theorem 2.2 ensures the existence of a (unique) pair $(\sigma_{\lambda_*}, \lambda_*)$ that solves (2.1)–(2.2). The inequality $S(\sigma) \geq 0$ for all $\sigma \in [\widetilde\sigma, \bar\sigma]$ implies $I(\lambda) > 0$ for all λ such that $\sigma_\lambda(0) \geq \widetilde\sigma$. Therefore, $I(\lambda_{\mathrm{nec}}) > 0$. □

Remark 6. Comparing Proposition 3.4 with Theorem 3.3, we see that existence of stationary solutions is possible even when $\sigma_{\mathrm{nec}} < \widetilde\sigma$.

(a) Imitating the proof of Theorem 2.3, it can be shown that if there exists $\alpha \in (\sigma_0, \bar\sigma)$ such that

$$\int_\alpha^{\bar\sigma} S(\sigma)(\sigma - \alpha)^2 d\sigma = 0, \tag{3.10}$$

then $\alpha \leq \sigma_*$ (thus there is no stationary solution for the necrotic model if $\sigma_{\mathrm{nec}} \leq \alpha$) and there is a (unique) stationary solution when $\sigma_{\mathrm{nec}} \geq k_\alpha := \bar\sigma - \frac{(\bar\sigma-\alpha)}{2}\frac{f(\alpha)}{f(\bar\sigma)}$. Moreover, we obtain a lower bound for the outer radius

[1] The same technique also copes with more general proliferation rates.

$R_{\rm nec} = \sqrt{\lambda_{\rho_*}}$ of the stationary solution in terms of the parameters $\sigma_{\rm nec}$ and $\bar{\sigma}$, which are intrinsic to the tumor:

$$\lambda_{\rho_*} \geq \frac{6(\bar{\sigma} - \sigma_{\rm nec})}{f(\bar{\sigma})}.$$

In addition, if the constant μ in (3.3) fulfills the condition

$$\mu \geq \frac{1}{(\sigma_{\rm nec} - \alpha)^3} \int_\alpha^{\sigma_{\rm nec}} |S(\sigma)|(\sigma - \alpha)^2 d\sigma,$$

the following bounds for the inner and outer radii are true:

$$\rho_* \leq \rho_\alpha := \frac{\sigma_{\rm nec} - \alpha}{\bar{\sigma} - \alpha}$$

and

$$\lambda_{\rho_*} \leq \frac{6(\bar{\sigma} - \alpha)}{f(\alpha)(1 - \rho_\alpha^3)}.$$

Let us now consider the linear case $f(\sigma) = \sigma$ and $S(\sigma) = \sigma - \widetilde{\sigma}$, as studied in [6, 10]. Then k_α is an improvement on the result given by Proposition 3.4. More precisely, if

$$\widetilde{\sigma} \in \left(\frac{7\bar{\sigma}}{8}, \bar{\sigma}\right),$$

then $k_\alpha < \widetilde{\sigma}$ and we still have existence of stationary solutions for $\sigma_{\rm nec} \geq k_\alpha$. A result of this type was obtained in [10].

(b) We can also improve Proposition 3.4 by following closely the proofs of Theorem 2.3 and Proposition 2.4. In fact, suppose that $S(\sigma_0) < 0 < S(\bar{\sigma})$. Let us denote

$$\bar{\lambda} := \max\left\{\frac{3(\bar{\sigma} - \alpha)}{f(\bar{\sigma})}, \frac{6(\bar{\sigma} - \beta)}{f(\bar{\sigma})}\right\},$$

where α and β are given by (2.10) and (2.11), respectively. If we define $\gamma \in [\sigma_0, \bar{\sigma}]$ as the unique solution of

$$\gamma + \frac{\bar{\lambda}}{6} f(\gamma) = \bar{\sigma},$$

it can be shown that there exists a (unique) stationary solution if $\sigma_{\rm nec} \geq \gamma$. Besides, the outer radius is bounded by

$$\lambda_{\rho_*} \geq \frac{6(\bar{\sigma} - \sigma_{\rm nec})}{f(\bar{\sigma})}.$$

If the constant μ is such that

$$\mu \geq \frac{1}{(\sigma_{\rm nec} - \gamma)^3} \int_{\sigma_{\rm nec}}^{\bar{\sigma}} S(\sigma)(\sigma - \gamma)^2 d\sigma,$$

then

$$\rho_* \leq \rho_\gamma := \frac{\sigma_{\rm nec} - \gamma}{\bar{\sigma} - \gamma}$$

and

$$\lambda_{\rho_*} \leq \frac{\bar{\lambda}}{6(1-\rho_\gamma)^2}.$$

Considering, as before, $f(\sigma) = \sigma$ and $S(\sigma) = \sigma - \widetilde{\sigma}$, this constant γ is an enhancement in Proposition 3.4 if $\widetilde{\sigma} \in \left(\frac{3}{5}\bar{\sigma}, \bar{\sigma}\right)$.

4. Conclusions.

Rigorous analysis of the model given by the system (1.1)–(1.3), (1.5) was made considering only its stationary solutions. A realistic class of nonlinear absorption and proliferations rates has also been regarded. Our procedure also takes into account the results proved in this work and seems to be justified.

Dealing with nonlinear absorption and proliferation rates is not a simple task: even the existence of stationary solutions for the model is questionable, while an explicit stationary solution is available for its linear version.

Considering the nonnecrotic model, the chosen approach proved to be effective. It was shown that a stationary solution is possible in many situations, and bounds for the stabilizing radius were obtained. Comparing our result with that in other papers, where only linear absorption and proliferations rates were considered, our bounds are noteworthy.

Studying a nonnecrotic tumor, Byrne and Chaplain [5] considered a linear absorption rate and a nonlinear proliferation rate in a numerical implementation and looked for a stationary solution as a limit of evolutionary solutions. Taking into account the bounds for the stabilizing radius given in this paper, we easily concluded that the numerical implementation done in that paper was destined to fail. This shows that prior knowledge of the possibility of a stationary solution has practical implications. Of course, similar circumstances are plausible in the mathematical analysis of a situation in biomedical praxis.

On the nonnecrotic model, however, the principal contribution of our paper is given by the ease of its computational implementation, as exemplified in section 2.3.

The major contribution of this paper involves the stationary solutions of the necrotic model (3.2)–(3.3). It concerns the analysis of the intertwining of the nonnecrotic and necrotic models. Results from the nonnecrotic model are decisive for the study of the necrotic model: the (numerically obtainable) constant σ_* conclusively settles the possibility of existence of a stationary solution. Once more, the biomedical consequences of this result are remarkable.

Of course, since the evolution equation has not been considered for nonlinear rates, the results of this paper are given as possibilities and not as predictions. However, in [4] we prove the stability of the stationary solution in the quasi-stationary case. This shows that the outcomes of this paper are most likely valid in praxis.

5. Proofs.

For the convenience of the reader, we state the following form of the maximum principle, which is a fundamental tool in the demonstration of our results. Its proof is a simple exercise.

LEMMA 5.1. *Suppose that* $w \in C^2([\rho,1],\mathbb{R})$ *satisfies*

$$(r^2w')' = \phi(r) + h(r)w, \quad \rho < r < 1, \qquad w'(\rho) \geq 0, \tag{5.1}$$

where ϕ *and* h *are continuous functions, with* $h \geq 0$ *on* $[\rho,1]$.

(i) *If* $w(1) \leq 0$ *and* $\phi \geq 0$ *on* $[\rho,1]$*, then* $\max w \leq 0$*;*

(ii) *if* $w(\rho) \geq 0$ *and* $\phi > 0$ *on* $[\rho,1]$*, then* $\min w = w(\rho) \geq 0$.

Proof of Proposition 2.1. For each $\lambda \geq 0$, the constant functions $\underline{\sigma} \equiv \sigma_0$ and $\overline{\sigma} \equiv \bar{\sigma}$ are sub- and supersolutions of the BVP (2.1), respectively. Hence, there exists at least one solution $\sigma_\lambda \in C^2([0,1])$ for this problem, satisfying

$$\sigma_0 \leq \sigma_\lambda(r) \leq \bar{\sigma} \quad \text{for all } r \in [0,1].$$

Lemma 5.1 brings in the uniqueness of σ_λ. In fact, if $\sigma_1, \sigma_2 \in C^2([0,1])$ are two solutions of this problem, $w := \sigma_1 - \sigma_2$ and $-w$ satisfy

$$\begin{aligned} &(r^2 w')' = h(r) w, \quad 0 < r < 1, \\ &w'(0) = 0 = w(1), \end{aligned}$$

where

$$h(r) = \lambda r^2 \int_0^1 f'(\xi \sigma_1(r) + (1-\xi)\sigma_2(r)) d\xi \geq 0.$$

By integrating (2.1) once, we obtain (i). Uniqueness for the initial value problem implies that $\sigma_0 < \sigma_\lambda(0)$. Combining this with (i), we obtain $\sigma_0 < \sigma_\lambda(r)$ for all $r \in [0,1]$. Inequality (ii) results by substituting (i) into (2.1).

We obtain (iii) by using again Lemma 5.1. The continuity of the map $\lambda \mapsto u_\lambda$ is proved by integrating (2.1) twice and then considering an increasing sequence (λ_n) in $[0,\infty)$ such that $\lambda_n \nearrow \lambda_\infty \in [0,\infty]$. We define

$$\sigma_\infty(r) = \lim_{n\to\infty} \sigma_n(r) = \inf_{\lambda \geq 0} \sigma_\lambda(r) \quad \text{for all } r \in [0,1].$$

Application of the Arzela–Ascoli theorem produces the result if $\lambda_\infty < \infty$ (considering also a decreasing subsequence). If $\lambda_\infty = \infty$, the dominated convergence theorem shows that $\sigma_\infty = \sigma_0$ almost everywhere in $[0,1]$. A simple argument proves then that the convergence is uniform on each closed interval of $[0,1)$. □

Proof of Theorem 2.3. Taking into account Proposition 2.1, the function $\psi(\lambda) = \sigma_\lambda(0)$ is continuous and nonincreasing. It follows easily that ψ is onto $(\sigma_0, \bar{\sigma}]$. Therefore, there exists at least one $\lambda_\alpha > 0$ such that $\psi(\lambda_\alpha) = \alpha$.

The graph of the function

$$v(r) = \alpha + (\bar{\sigma} - \alpha) r$$

is the straight line through the points $(0,\alpha)$ and $(1,\bar{\sigma})$. Since σ_{λ_α} is strictly convex, for all $r \in [0,1]$ we have

$$\alpha \leq \sigma_{\lambda_\alpha}(r) \leq v(r) \leq \bar{\sigma}. \tag{5.2}$$

Now, let us consider the auxiliary function $\phi(\lambda) = \sigma'_\lambda(1)$ for $\lambda \geq 0$. The continuous function ϕ clearly satisfies $\phi(0) = 0$. We claim that $\phi(\infty) = \infty$. In fact, from (2.4) we obtain

$$r \sigma'_\lambda(1) \geq \sigma'_\lambda(r) \quad \text{for all} \quad 0 < r < 1.$$

Integration of the last inequality produces

$$\sigma'_\lambda(1) \left(\frac{1 - r^2}{2} \right) \geq \bar{\sigma} - \sigma_\lambda(r), \quad 0 < r < 1. \tag{5.3}$$

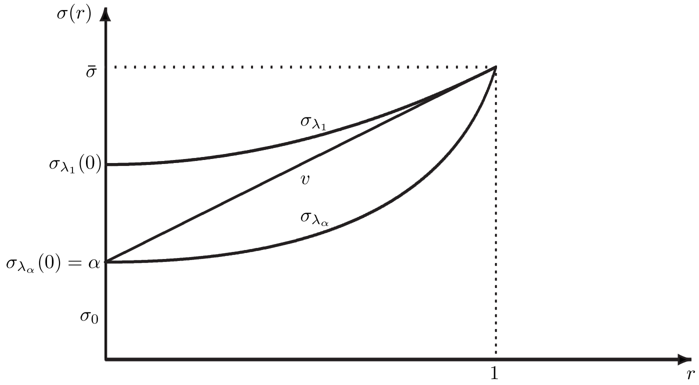


FIG. 5.1. *The geometrical construction in Theorem* 2.3: $\alpha \le \sigma_{\lambda_\alpha}(r) \le v(r) \le \sigma_{\lambda_1}(r) \le \bar{\sigma}$.

Consequently,

$$\lim_{\lambda\to\infty} \phi(\lambda) \ge \lim_{\lambda\to\infty} \frac{2(\bar{\sigma}-\sigma_\lambda(r))}{1-r^2} = \frac{2(\bar{\sigma}-\sigma_0)}{1-r^2} \quad \text{for each } r \in [0,1).$$

The claim results by making $r \to 1^-$.

Ergo, there exists $\lambda_1 > 0$ such that $\phi(\lambda_1) = \bar{\sigma} - \alpha$. The convexity of σ_{λ_1} implies that

$$\sigma_{\lambda_1}(r) \ge (\bar{\sigma}-\alpha)r + \alpha = v(r), \quad r \in [0,1]. \tag{5.4}$$

Combining (5.2) and (5.4) yields (see Figure 5.1)

$$\alpha \le \sigma_{\lambda_\alpha}(r) \le v(r) \le \sigma_{\lambda_1}(r) \le \bar{\sigma} \quad \text{for all } r \in [0,1].$$

Hence,

$$\int_0^1 S(\sigma_{\lambda_\alpha}(r))r^2 dr \le \int_0^1 S(v(r))r^2 dr \le \int_0^1 S(\sigma_{\lambda_1}(r))r^2 dr. \tag{5.5}$$

Since

$$\int_0^1 S(v(r))r^2 dr = \frac{1}{(\bar{\sigma}-\alpha)^3} \int_\alpha^{\bar{\sigma}} S(\sigma)(\sigma-\alpha)^2 d\sigma = 0,$$

inequality (5.5) means that

$$I(\lambda_\alpha) \le 0 \le I(\lambda_1).$$

The existence of $\lambda_* \in [\lambda_1, \lambda_\alpha]$ such that $I(\lambda_*) = 0$ follows then by continuity.

Now we obtain the estimates (2.8). Since $\alpha = \sigma_{\lambda_\alpha}(0) \le \sigma_{\lambda_\alpha}(r)$ and

$$\sigma_{\lambda_\alpha}(0) = \bar{\sigma} - \lambda_\alpha \int_0^1 \int_0^\theta \left(\frac{s}{\theta}\right)^2 f(\sigma_{\lambda_\alpha}(s)) ds d\theta,$$

we have

$$\alpha \le \bar{\sigma} - \frac{\lambda_\alpha}{6} f(\alpha) \le \bar{\sigma} - \frac{\lambda_*}{6} f(\alpha),$$

thus showing the upper bound for λ_*.

On the other hand, since

$$\sigma'_{\lambda_1}(0) = \phi(\lambda_1) = \bar{\sigma} - \alpha = \lambda_1 \int_0^1 s^2 f(\sigma_{\lambda_1}(s))ds,$$

we find

$$\bar{\sigma} - \alpha \leq \lambda_1 f(\bar{\sigma}) \int_0^1 s^2 ds < \frac{\lambda_*}{3} f(\bar{\sigma}),$$

from which we obtain the lower bound. □

Proof of Proposition 2.4. Proposition 2.2 brings in existence and uniqueness of λ_*. Since $S(\bar{\sigma}) > 0$,

$$\int_\xi^{\bar{\sigma}} S(\sigma)(\sigma - \xi)^2 d\sigma \geq 0 \quad \text{for all } \xi \text{ near } \bar{\sigma}. \tag{5.6}$$

So, for each ξ as above, we repeat the construction of the last proof to find

$$I(\lambda_1) \geq \int_0^1 S(v(r))r^2 dr = \frac{1}{(\bar{\sigma} - \xi)^3} \int_\xi^{\bar{\sigma}} S(\sigma)(\sigma - \xi)^2 dt \geq 0,$$

showing that $\lambda_1 \leq \lambda_*$. Because of that,

$$\bar{\sigma} - \xi \leq \lambda_1 f(\bar{\sigma}) \int_0^1 s^2 ds \leq \frac{\lambda_*}{3} f(\bar{\sigma})$$

implies the lower bound

$$\lambda_* \geq 3 \frac{(\bar{\sigma} - \xi)}{f(\bar{\sigma})}.$$

But the same function S also satisfies

$$\int_\eta^{\bar{\sigma}} S(\sigma)\sqrt{\sigma - \eta} d\sigma \geq 0 \quad \text{for all } \eta \text{ near } \bar{\sigma}. \tag{5.7}$$

Let $\lambda_\eta > 0$ be such that $\sigma'_{\lambda_\eta}(1) = 2(\bar{\sigma} - \eta)$. From (5.3) we obtain

$$\sigma_{\lambda_\eta}(r) \geq \bar{\sigma} - \frac{\sigma'_{\lambda_\eta}(1)}{2}(1 - r^2) = \eta + (\bar{\sigma} - \eta)r^2,$$

and, since $S(\sigma)$ is increasing and

$$\int_0^1 S(\eta + (\bar{\sigma} - \eta)r^2)r^2 dr = \frac{1}{2\,(\bar{\sigma} - \eta)^{3/2}} \int_\eta^{\bar{\sigma}} S(\sigma)\sqrt{\sigma - \eta} d\sigma \geq 0,$$

we conclude that $I(\lambda_\eta) \geq 0$. The monotonicity of S then implies that $\lambda_* \geq \lambda_\eta$. Accordingly,

$$2(\bar{\sigma} - \eta) = \sigma'_{\lambda_\eta}(1) = \lambda_\eta \int_0^1 s^2 f(\sigma_{\lambda_\eta}(s))ds \leq \frac{\lambda_* f(\bar{\sigma})}{3}$$

implies the bound

$$\lambda_* \geq \frac{6(\bar{\sigma}-\eta)}{f(\bar{\sigma})}.$$

Collecting the two estimates, we obtain (2.9). □

Proof of Theorem 2.5. Since $I(0) = S(\bar{\sigma})/3 > 0$ and $\lim_{\lambda\to\infty} I(\lambda) = \int_0^1 S(\sigma_0)r^2 dr = 0$, it suffices to verify that $I(\lambda)$ tends to zero by negative values.

Fix $\lambda_\Lambda > 0$ such that $\sigma_{\lambda_\Lambda}(0) = \Lambda$. The monotonicity of $\sigma_\lambda(r)$ with respect to λ and to r ensures the existence of a unique $r_\lambda \in (0,1)$ such that $\sigma_\lambda(r_\lambda) = \Lambda$ for all $\lambda \geq \lambda_\Lambda$. Denote by Γ the function whose graph is the straight line through the points (r_λ, Λ) and $(1, \bar{\sigma})$. Since the functions σ_λ are strictly convex, $\Gamma(r) > \sigma_\lambda(r)$ if $r > r_\lambda$ and $\Gamma(r) < \sigma_\lambda(r)$ if $r < r_\lambda$. Now, for each $\lambda \geq \lambda_\Lambda$, let $(s_\lambda, \sigma_\lambda(0))$ be the point where the horizontal line $\sigma = \sigma_\lambda(0) \leq \Lambda$ intersects Γ (see Figure 5.2).

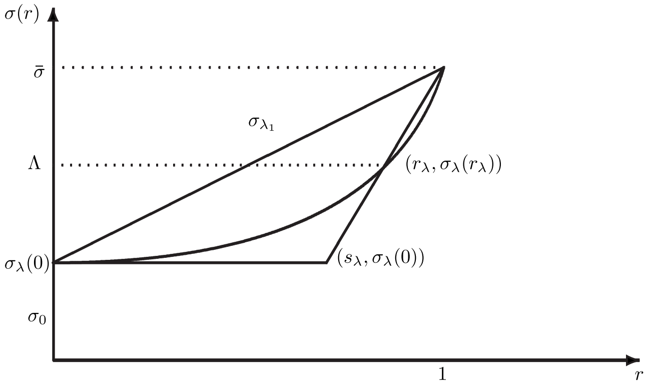


FIG. 5.2. *The geometrical construction in Theorem* 2.5*:* $s_\lambda = \frac{\bar{\sigma}-\sigma_\lambda(0)}{\bar{\sigma}-\Lambda} r_\lambda - \frac{\Lambda-\sigma_\lambda(0)}{\bar{\sigma}-\Lambda} < r_\lambda$.

So, Γ is given by

$$\Gamma(r) = \sigma_\lambda(0) + \frac{\bar{\sigma}-\sigma_\lambda(0)}{1-s_\lambda}(r - s_\lambda),$$

where

$$s_\lambda = \frac{\bar{\sigma}-\sigma_\lambda(0)}{\bar{\sigma}-\Lambda} r_\lambda - \frac{\Lambda-\sigma_\lambda(0)}{\bar{\sigma}-\Lambda} < r_\lambda.$$

From Proposition 2.1 we deduce that

$$\lim_{\lambda\to\infty} s_\lambda = \lim_{\lambda\to\infty} r_\lambda = 1.$$

Let us define the auxiliary function

$$v_\lambda(r) = \begin{cases} \sigma_\lambda(0) & \text{if } 0 \leq r \leq s_\lambda, \\ \sigma_\lambda(0) + \frac{\bar{\sigma}-\sigma_\lambda(0)}{1-s_\lambda}(r - s_\lambda) & \text{if } s_\lambda \leq r \leq 1. \end{cases} \tag{5.8}$$

Thus, v_λ coincides with the horizontal line $\sigma = \sigma_\lambda(0)$ for $0 \leq r \leq s_\lambda$ and with Γ for $s_\lambda \leq r \leq 1$.

For $0 \leq r \leq r_\lambda$ we have $v_\lambda(r) \leq \sigma_\lambda(r)$. Since S is decreasing in $[0, \Lambda]$, we have $S(\sigma_\lambda(r)) < S(v_\lambda(r))$ for all $r \in [0, r_\lambda]$. On the other hand, we have $v_\lambda(r) \geq \sigma_\lambda(r)$ if $r_\lambda \leq r \leq 1$. Since S is increasing in the interval $[\Lambda, \bar{\sigma}]$, we have that $S(\sigma_\lambda(r)) \leq S(v_\lambda(r))$ for all $r \in [r_\lambda, 1]$. We deduce that

$$S(\sigma_\lambda) \leq S(v_\lambda) \text{ on } [0,1].$$

For all $\lambda \geq \lambda_\Lambda$, it follows that

$$\begin{aligned} I(\lambda) &\leq \int_0^1 S(v_\lambda(r))r^2 dr \\ &= \int_0^{s_\lambda} S(v_\lambda(r))r^2 dr + \int_{s_\lambda}^1 S(v_\lambda(r))r^2 dr \\ &= S(\sigma_\lambda(0))\frac{s_\lambda^3}{3} + \int_{s_\lambda}^1 S(v_\lambda(r))r^2 dr. \end{aligned}$$

Since S is decreasing for $[\sigma_0, \sigma_\lambda(0)] \subset [\sigma_0, \Lambda]$ and $S(\sigma_0) = 0$, we have

$$S(\sigma_\lambda(0))s_\lambda^3/3 \leq 0 \quad \text{for all } \lambda \geq \lambda_\Lambda.$$

Therefore,

$$I(\lambda) \leq \int_{s_\lambda}^1 S(v_\lambda(r))r^2 dr \quad \text{for all } \lambda \geq \lambda_\Lambda.$$

Making use of (5.8), we obtain

$$\int_{s_\lambda}^1 S(v_\lambda(r))r^2 dr = (1 - s_\lambda)p(\lambda),$$

where

$$p(\lambda) = \frac{1}{(\bar{\sigma} - \sigma_\lambda(0))^3} \int_{\sigma_\lambda(0)}^{\bar{\sigma}} S(\sigma)[(1 - s_\lambda)\,\sigma + \bar{\sigma} s_\lambda - \sigma_\lambda(0)^2] d\sigma.$$

Since $s_\lambda \to 1$ and $\sigma_\lambda(0) \to \sigma_0^+$ when $\lambda \to \infty$, we have

$$\lim_{\lambda \to \infty} p(\lambda) = \frac{1}{\bar{\sigma} - \sigma_0} \int_{\sigma_0}^{\bar{\sigma}} S(\sigma) d\sigma \leq 0.$$

Consequently,

$$I(\lambda) \leq (1 - s_\lambda)p(\lambda) \leq 0$$

for all $\lambda \geq \lambda_\Lambda$ large enough, since $(1 - s_\lambda) > 0$ for all $\lambda \geq \lambda_\Lambda$. This shows the existence of $\lambda_* > \lambda_\Lambda$ such that $I(\lambda_*) = 0$.

Furthermore, since $\Lambda = \sigma_{\lambda_\Lambda}(0)$ and

$$\Lambda = \sigma_{\lambda_\Lambda}(0) = \bar{\sigma} - \lambda_\Lambda \int_0^1 \int_0^\theta (s/\theta)^2 f(\sigma_{\lambda_\Lambda}(s)) ds d\theta \geq \bar{\sigma} - \frac{\lambda_*}{6} f(\bar{\sigma}),$$

we obtain the lower bound (2.12) for λ_*. □

REFERENCES

[1] D. Ambrosi and L. Preziosi, *On the closure of mass balance models for tumor growth*, Math. Models Methods Appl. Sci., 12 (2002), pp. 737–754.

[2] R. Araujo and D. L. S. McElwain, *A history of the study of solid tumour growth: The contribution of mathematical modelling*, Bull. Math. Biol., 66 (2004), pp. 1039–1091.

[3] A. Bertuzzi, A. Fasano, and A. Gandolfi, *A free boundary problem with unilateral constraints describing the evolution of a tumor cord under the influence of cell killing agents*, SIAM J. Math. Anal., 36 (2004), pp. 882–915.

[4] H. Bueno, G. Ercole, and A. Zumpano, *Asymptotic behavior of quasi-stationary solutions of a nonlinear problem modeling the growth of tumors*, Nonlinearity, 18 (2005), pp. 1629–1642.

[5] H. M. Byrne and M. A. J. Chaplain, *Growth of non-necrotic tumors in the presence and absence of inhibitors*, Math. Biosci., 130 (1995), pp. 151–181.

[6] H. M. Byrne and M. A. J. Chaplain, *Growth of necrotic tumors in the presence and absence of inhibitors*, Math. Biosci., 135 (1996), pp. 187–217.

[7] H. M. Byrne, J. R. King, D. L. S. McElwain, and L. Preziosi, *A two-phase model of tumour growth*, Appl. Math. Lett., 16 (2003), pp. 567–573.

[8] H. M. Byrne and L. Preziosi, *Modelling solid tumour growth using the theory of mixtures*, Math. Med. Biol., 20 (2003), pp. 341–366.

[9] S. Cui, *Analysis of a mathematical model for the growth of tumors under the action of external inhibitors*, J. Math. Biol., 44 (2002), pp. 395–426.

[10] S. Cui and A. Friedman, *Analysis of a mathematical model of the growth of necrotic tumors*, J. Math. Anal. Appl., 255 (2001), pp. 636–677.

[11] S. Cui and A. Friedman, *A free boundary problem for a singular system of differential equations: An application to a model of tumor growth*, Trans. Amer. Math. Soc., 355 (2002), pp. 3537–3590.

[12] D. G. de Figueiredo, *Positive solutions of semilinear elliptic problems*, in Differential Equations, Lecture Notes in Math. 957, Springer-Verlag, Berlin, 1982, pp. 34–87.

[13] J. I. Díaz and J. I. Tello, *On the mathematical controllability in a simple growth tumours model by the internal localized action of inhibitors*, Nonlinear Anal. Real World Appl., 4 (2002), pp. 109–125.

[14] A. Friedman and F. Reitich, *Analysis of a mathematical model for the growth of tumors*, J. Math. Biol., 38 (1999), pp. 262–284.

SIAM J. Appl. Math.
Vol. 68, No. 4, pp. 1026–1044

RECONSTRUCTING DISCONTINUITIES USING COMPLEX GEOMETRICAL OPTICS SOLUTIONS[*]

GUNTHER UHLMANN[†] AND JENN-NAN WANG[‡]

Abstract. In this paper we provide a framework for constructing general complex geometrical optics solutions for several systems of two variables that can be reduced to a system with the Laplacian as the leading order term. We apply these special solutions to the problem of reconstructing inclusions inside a domain filled with known conductivity from local boundary measurements. Computational results demonstrate the versatility of these solutions to determine electrical inclusions.

Key words. geometrical optics, inclusions, numerical algorithm

AMS subject classifications. 35R30, 65N21

DOI. 10.1137/060676350

1. Introduction.

Inverse boundary value problems are a class of inverse problems where one attempts to determine the internal parameters of body by making measurements only at the surface of the body. A prototypical example that has received a lot of attention is electrical impedance tomography (EIT). In this inverse method one would like to determine the conductivity distribution inside a body by making voltage and current measurements at the boundary.

There are many applications of EIT ranging from early breast cancer detection [32] to geophysical sensing for underground objects; see [18], [24], [25], [27]. The article [28] and the ones reviewed in [29] assume that the measurements are made on the whole boundary. However, it is often possible to make the measurements only on part of the boundary; this is the partial data problem. This is the case for the applications in breast cancer detection and geophysical sensing mentioned above.

The boundary information is encoded into the Dirichlet-to-Neumann map associated with the conductivity equation. More precisely, let Ω be an open bounded domain with smooth boundary in $\mathbb{R}^d$ with $d = 2$ or 3. Assume that $\gamma(x) > 0$ in Ω possesses a suitable regularity. The conductivity equation is described by the following elliptic equation:

$$\nabla \cdot (\gamma(x)\nabla u) = 0 \quad \text{in} \quad \Omega. \tag{1.1}$$

For an appropriate function f defined on $\partial\Omega$, there exists a unique solution $u(x)$ to the boundary value problem for (1.1) with Dirichlet condition $u|_{\partial\Omega} = f$. Thus, one can define a map Λ_γ sending the Dirichlet data to the Neumann data by

$$\Lambda_\gamma(f) = \gamma\frac{\partial u}{\partial \nu}\Big|_{\partial\Omega}.$$

[*] Received by the editors November 30, 2006; accepted for publication (in revised form) October 16, 2007; published electronically February 29, 2008.
http://www.siam.org/journals/siap/68-4/67635.html

[†] Department of Mathematics, University of Washington, Box 354305, Seattle, WA 98195-4350 (gunther@math.washington.edu). This author's research was partially supported by NSF and a Walker Family Endowed Professorship.

[‡] Department of Mathematics, Taida Institute for Mathematical Sciences, and NCTS (Taipei), National Taiwan University, Taipei 106, Taiwan (jnwang@math.ntu.edu.tw). This author's research was supported in part by the National Science Council of Taiwan (NSC 95-2115-M-002-003).

The map Λ_γ is the Dirichlet-to-Neumann map associated with the conductivity equation (1.1). It is worth mentioning that even though (1.1) is linear, the map Λ_γ depends nonlinearly on γ. The famous Calderón problem is to determine γ from the knowledge of Λ_γ.

In [3], Calderón studied this inverse problem by linearizing the fully nonlinear problem around a constant conductivity function. To attack this linearized problem, Calderón introduced harmonic functions of the form $e^{x\cdot\rho}$ with $\rho \in \mathbb{C}^n$ and $\rho\cdot\rho = 0$, which is the genesis of *complex geometrical optics* (CGO) *solutions* since the phase function $x\cdot\rho$ is complex-valued. Inspired by Calderón's approach, Sylvester and Uhlmann [28] solved the uniqueness question of Calderón's problem for smooth conductivities by constructing CGO solutions for (1.1). Since the conductivity equation (1.1) is closely related to the Schrödinger equation (see (2.2)), it suffices to construct CGO solutions for the Schrödinger which are of the form $u(x) = e^{x\cdot\rho}(1 + r(x,\rho))$, where r is decaying in $|\rho|$. To motivate the name of the solution, we write

$$u(x) = e^{ih^{-1}x\cdot(\omega_1+i\omega_2)}(1 + h\tilde{r}), \tag{1.2}$$

where $h = |\rho|^{-1}$, $i(\omega_1 + i\omega_2) = |\rho|^{-1}(\mathrm{Re}\rho + i\mathrm{Im}\rho)$, and $\tilde{r} = h^{-1}r = |\rho|r$. The form (1.2) is analogous to the geometrical optics solution for the wave propagation equation in which the phase function is real-valued. Here the phase function in (1.2) is complex-valued. Nevertheless, it is linear. CGO solutions have been used in EIT and have been instrumental in solving several inverse problems. We will not review these developments in detail here; see [30] and [29] for references; other reviews in EIT are [1], [2], and [4].

Recently, new CGO solutions that are useful for the partial data problem were constructed in [20] for the conductivity equation and zeroth order perturbations of the Laplacian. The real parts of the phase of these solutions are limiting Carleman weights. They have been generalized to first order perturbation of the Laplacian for scalar equations or systems in [5], [9], [26], and [31]. Constructions of CGO solutions for the conductivity equation and zeroth order perturbations of the Laplacian using hyperbolic geometry can be found in [16], [17]; these have been applied to determine electrical inclusions in [10].

In two dimensions, when the underlying equation has the Laplacian as the leading part, due to the rich conformal structure, we have more freedom of choosing the complex phases for the CGO solutions. In particular any harmonic function is a limiting Carleman weight and can be the real part of a CGO solution. The aim of the paper is to provide a framework for constructing these solutions for several systems of two variables that can be reduced to a system with the Laplacian as the leading term. We apply these special solutions to the problem of reconstructing inclusions inside a domain filled with known conductivity from local boundary measurements. We also provide numerical results to demonstrate the applicability and flexibility of these special solutions.

From now on, we consider the case $d = 2$, i.e., the $\mathbb{R}^2$ plane. Let $n \in \mathbb{N}$ and denote $U(x) = (u_1(x_1,x_2),\ldots,u_n(x_1,x_2))^\top$. We consider the following system of equations:

$$PU := \Delta_x U + A_1(x)\partial_{x_1}U + A_2(x)\partial_{x_2}U + Q(x)U = 0 \quad \text{in} \quad \Omega, \tag{1.3}$$

where $\Delta_x = \partial_{x_1}^2 + \partial_{x_2}^2$ and A_1, A_2, Q are $n\times n$ matrices whose regularities will be specified later. The system (1.3) contains all scalar or two-dimensional physical systems that can be reduced to a system with the Laplacian as the leading part. Those systems include the conductivity equation, the magnetic Schrödinger equation, the two-dimensional isotropic elasticity system, the two-dimensional Stokes system, etc. In this paper we first study CGO solutions with special phase functions for (1.3).

In the papers [20], [5], [9], [10], [17], [26], and [31], the real parts of the phase functions are radial functions. These can be used to probe the region with spherical fronts, the so-called complex spherical waves. Even though these solutions are better suited for the local data problem than the usual CGO solutions with linear phase functions, they are still quite restrictive. Fortunately, in the two-dimensional case, we have many more choices of phase functions. For example, let $\varphi(x)$ be a harmonic function with nonvanishing gradient in Ω; then $\varphi + i\psi$ can be the phase function of the CGO solutions when ψ is a harmonic conjugate of φ. In other words, $\rho(x) := \varphi(x) + i\psi(x)$ is holomorphic in Ω. Our method in this paper is developed based on this idea.

Using the CGO solutions, we can consider the problem of finding embedded inclusions in a known medium. This is the object identification problem. The method developed here shares the same spirit as Ikehata's *enclosure method* [11], [12]. For the two-dimensional problem, we would like to mention a very interesting result by Ikehata in [14], where he introduced the Mittag–Leffler function in the object identification problem. This has the property that its modulus grows exponentially in some cone and decays to zero algebraically outside the same cone. Using the Mittag–Leffler function and shrinking the opening angle of the cone, one can reconstruct precisely the shapes of some embedded objects such as star-shaped objects. The numerical implementation of the Mittag–Leffler functions was carried out by Ikehata and Siltanen in [15]. The main restriction of the method using the Mittag–Leffler function is that it can be applied only to scalar equations with a homogeneous background. That is, they probe the region with harmonic functions. The novelty of our method is its flexibility in treating scalar equations, or even two-dimensional *systems*, with an *inhomogeneous background*. Furthermore, for the object identification problem in such general systems, using our special CGO solutions, we are able to reconstruct the precise information of some embedded objects including star-shaped regions by boundary measurements. This identification result is similar to that in [14] and [15], where only the Laplace equation is treated. So, in theory, our reconstruction method with these CGO solutions is in greater generality. In this paper, we are developing the foundational work to treat the case of an inhomogeneous background and also to deal with the case of systems. Moreover, we give numerical evidence that the method works in the homogeneous case.

Before going further, we also would like to compare our method with that in [10]. As we have pointed out above, the real parts of the phase functions of CGO solutions in [10] are radially symmetric. So their probing fronts are circles or spheres. Moreover, the construction of CGO solutions in [10] is based on the hyperbolic geometry. It has not been developed for studying more general equations or systems. The advantage of our method lies in the freedom of choosing the phase functions of CGO solutions. One useful example is to take $\rho(x)$ as a polynomial. By increasing the degree of the polynomial, we can narrow our probing fronts. Consequently, we are able to determine more information in the object identification problem in the two-dimensional case than [10] does. On the other hand, since the real parts of the phase functions in our CGO solutions are not necessarily radially symmetric, we can create different probing fronts by simply rotating the phase functions. Like [10], we can also localize the measurements in an arbitrarily small region on the boundary. Here the local data means that the Dirichlet condition is nonzero only on a small part of the boundary.

On the same region, we measure the Neumann condition. In theory, the nonzero part of the Dirichlet data can be taken as small as we wish.

Our construction of CGO solutions with more general phases is rather elementary. The main idea is to transform CGO solutions with linear phases by suitable conformal mappings. The construction of CGO solutions with linear phases for (1.3) was first given by Nakamura and Uhlmann in [21], [22], where they introduced the intertwining technique in handling the first order terms (also see [7] for similar results). Here we shall use Carleman's technique to construct CGO solutions with linear phases for (1.3).

This paper is organized a follows. In section 2, we give concrete examples of (1.3). In section 3, we review of the construction of CGO solutions with linear phases for (1.3). CGO solutions with more general phases will be discussed in section 4. For an application of CGO solutions with general phases, we consider the problem of reconstructing inclusions embedded into a domain with known conductivity by boundary measurements. Numerical experiments of our method are presented in section 6.

2. Physical examples of (1.3).

2.1. Conductivity equation. Our first example is the well-known conductivity equation already given in the previous section. Let $\gamma(x) \in C^2(\bar{\Omega})$ and $\gamma(x) > 0$ for all $x \in \bar{\Omega}$. We consider the equation

$$\nabla \cdot (\gamma \nabla u) = 0 \quad \text{in} \quad \Omega. \tag{2.1}$$

Introducing the new variable $v = \gamma^{1/2} u$, (2.1) is equivalent to

$$(\Delta + q)v = 0 \quad \text{in} \quad \Omega \tag{2.2}$$

with $q = -\Delta\gamma^{1/2}/\gamma^{1/2} \in L^\infty(\Omega)$. Equation (2.2) is a Schrödinger-type equation. We can also consider a more general Schrödinger-type equation with a convection term:

$$(\Delta + a(x) \cdot \nabla + q)v = 0 \quad \text{in} \quad \Omega, \tag{2.3}$$

where $a = (a_1, a_2)$.

2.2. Isotropic elasticity. The domain Ω is now modeled as an inhomogeneous, isotropic, elastic medium characterized by the Lamé parameters $\lambda(x)$ and $\mu(x)$. Assume that $\lambda(x) \in C^2(\overline{\Omega})$, $\mu(x) \in C^4(\overline{\Omega})$, and the following inequalities hold:

$$\mu(x) > 0 \quad \text{and} \quad \lambda(x) + 2\mu(x) > 0 \quad \forall\, x \in \overline{\Omega} \quad \text{(strong ellipticity).} \tag{2.4}$$

We consider the static isotropic elasticity system without sources

$$\nabla \cdot (\lambda(\nabla \cdot u)I + 2\mu S(\nabla u)) = 0 \quad \text{in} \quad \Omega. \tag{2.5}$$

Here and below, $S(A) = (A + A^T)/2$ denotes the symmetric part of the matrix $A \in \mathbb{C}^{2\times 2}$. Equivalently, if we denote $\sigma(u) = \lambda(\nabla \cdot u)I + 2\mu S(\nabla u)$ the stress tensor, then (2.5) becomes

$$\nabla \cdot \sigma = 0 \quad \text{in} \quad \Omega.$$

On the other hand, since the Lamé parameters are differentiable, we can also write (2.5) in the nondivergence form

$$\mu\Delta u + (\lambda + \mu)\nabla(\nabla \cdot u) + \nabla\lambda\nabla \cdot u + 2S(\nabla u)\nabla\mu = 0 \quad \text{in} \quad \Omega. \tag{2.6}$$

We will use the reduced system derived by Ikehata [13]. This reduction was also mentioned in [29]. Let $\binom{w}{g}$ satisfy

$$\Delta\begin{pmatrix} w\\ g\end{pmatrix}+A(x)\begin{pmatrix}\nabla g\\ \nabla\cdot w\end{pmatrix}+Q(x)\begin{pmatrix} w\\ g\end{pmatrix}=0, \tag{2.7}$$

where

$$A(x)=\begin{pmatrix}2\mu^{-1/2}(-\nabla^2+\Delta)\mu^{-1} & -\nabla\log\mu\\ 0 & \frac{\lambda+\mu}{\lambda+2\mu}\mu^{1/2}\end{pmatrix}$$

and

$$Q(x)=\begin{pmatrix}-\mu^{-1/2}(2\nabla^2+\Delta)\mu^{1/2} & 2\mu^{-5/2}(\nabla^2-\Delta)\mu\ \nabla\mu\\ -\frac{\lambda-\mu}{\lambda+2\mu}(\nabla\mu^{1/2})^T & -\mu\Delta\mu^{-1}\end{pmatrix}.$$

Here $\nabla^2 f$ is the Hessian of the scalar function f. Then

$$u:=\mu^{-1/2}w+\mu^{-1}\nabla g-g\nabla\mu^{-1}$$

satisfies (2.6). A similar form was also used in [7] for studying the inverse boundary value problem for the isotropic elasticity system.

2.3. Stokes system. Let $\mu(x)\in C^4(\bar\Omega)$ and $\mu(x)>0$ for all $x\in\bar\Omega$. Here μ is called the viscosity function. Suppose that $u=(u_1,u_2)$ and p satisfy the Stokes system

$$\begin{cases}\nabla\cdot(\mu S(\nabla u))-\nabla p=0 \quad \text{in} \quad \Omega,\\ \nabla\cdot u=0 \quad \text{in} \quad \Omega.\end{cases} \tag{2.8}$$

Here u and p represent the velocity field and the pressure, respectively. Motivated by the isotropic elasticity, we set $u=\mu^{-1/2}w+\mu^{-1}\nabla g-(\nabla\mu^{-1})g$ and

$$p=\nabla\mu^{1/2}\cdot w+\mu^{1/2}\nabla\cdot w+2\Delta g=\nabla\cdot(\mu^{1/2}w)+2\Delta g; \tag{2.9}$$

then (u,p) is a solution of (2.8), provided $\binom{w}{g}$ satisfies

$$\Delta\begin{pmatrix} w\\ g\end{pmatrix}+A(x)\begin{pmatrix}\nabla g\\ \nabla\cdot w\end{pmatrix}+Q(x)\begin{pmatrix} w\\ g\end{pmatrix}=0 \tag{2.10}$$

with

$$A(x)=\begin{pmatrix}-2\mu^{1/2}\nabla^2\mu^{-1} & -\mu^{-1}\nabla\mu\\ 0 & \mu^{1/2}\end{pmatrix}$$

and

$$Q=\begin{pmatrix}-2\mu^{-1/2}\nabla^2\mu^{1/2}-\mu^{-1/2}\Delta\mu^{1/2} & -4\nabla^2\mu^{-1}\nabla\mu^{1/2}-2\mu^{1/2}\nabla\cdot(\nabla\mu^{-1})\\ \mu(\nabla\mu^{-1/2})^T & -\mu\Delta\mu^{-1}\end{pmatrix}.$$

3. CGO solutions with linear phases. In this section we review the method of constructing CGO solutions with linear phases using Carleman estimates. We consider a slightly different system here. Let $\tilde{\Omega}$ be an open bounded domain in $\mathbb{R}^2$. Let $V(y) = V(y_1, y_2)$ satisfy

$$\Delta_y V + \tilde{A}_1 \partial_{y_1} V + \tilde{A}_2 \partial_{y_2} V + \tilde{Q} V = 0 \quad \text{in} \quad \tilde{\Omega}. \tag{3.1}$$

Assume that $\tilde{A}_1, \tilde{A}_2 \in C^2(\bar{\tilde{\Omega}})$ and $\tilde{Q} \in L^\infty(\tilde{\Omega})$. Given $\omega \in \mathbb{R}^2$ with $|\omega| = 1$, we look for $V(y)$ of (3.1) having the form

$$V(y) = e^{y\cdot(\omega+i\omega^\perp)/h}(\tilde{L} + \tilde{R}), \tag{3.2}$$

where $\tilde{L}$ is independent of h and $\tilde{R}$ satisfies

$$\|\partial^\alpha \tilde{R}\|_{L^2(\tilde{\Omega})} \le Ch^{1-\alpha} \quad \forall\ |\alpha| \le 2. \tag{3.3}$$

To construct V having the form (3.2), (3.3), we follow the approach in [9] and [31], which are based on [5] and [20]. Note that the real part of the phase function $y \cdot \omega$ is a limiting Carleman estimate. So if we define the semiclassical operator

$$P_h = h^2\Delta + h\tilde{A}_1(h\partial_{y_1}) + h\tilde{A}_2(h\partial_{y_2}) + h^2\tilde{Q},$$

then we can derive, by combining a Carleman estimate and the Hahn–Banach theorem, the following.

THEOREM 3.1 (see [9], [31]). *For h sufficiently small, for any $F \in L^2(\tilde{\Omega})$, there exists $W \in H^2_h(\tilde{\Omega})$ such that*

$$e^{-y\cdot\omega/h}P_h(e^{y\cdot\omega/h}W) = F$$

and $h\|W\|_{H^2_h(\tilde{\Omega})} \le C\|F\|_{L^2(\tilde{\Omega})}$, where $\|W\|^2_{H^2_h(\tilde{\Omega})} = \sum_{|\alpha|\le 2} \|(h\partial)^\alpha W\|^2_{L^2(\tilde{\Omega})}$ is the semiclassical H^2 norm.

This theorem will be needed below. Finding V of the form (3.2) is equivalent to solving

$$e^{-y\cdot(\omega+i\omega^\perp)/h}P_h(e^{y\cdot(\omega+i\omega^\perp)/h}(\tilde{L}+\tilde{R})) = 0 \quad \text{in} \quad \tilde{\Omega}.$$

We can compute that

$$e^{-y\cdot(\omega+i\omega^\perp)/h}P_h e^{y\cdot(\omega+i\omega^\perp)/h} = hT_\omega + P_h,$$

where $T_\omega = 2(\omega+i\omega^\perp)\cdot\nabla + (\omega+i\omega^\perp)\cdot(\tilde{A}_1, \tilde{A}_2)$. Hence we want to find $\tilde{L}$, independent of h, so that

$$T_\omega \tilde{L} = 0 \quad \text{in} \quad \tilde{\Omega}. \tag{3.4}$$

Equation (3.4) is a system of Cauchy–Riemann type. In fact, introducing the new variable $z = (z_1, z_2) = (\omega + i\omega^\perp) \cdot y$ and setting $\tilde{A}(\omega, z) = (\omega + i\omega^\perp) \cdot (\tilde{A}_1, \tilde{A}_2)$, (3.4) becomes

$$(4\partial_{\bar{z}} + \tilde{A})\tilde{L} = 0, \tag{3.5}$$

where $\partial_{\bar{z}} = (\partial_{z_1} + i\partial_{z_2})/2$. The existence of nontrivial $\tilde{L}$ can be found in, for example, [6], [8], and [23]. Having found $\tilde{L}$, $\tilde{R}$ is required to satisfy

$$e^{-y\cdot\omega/h}P_h(e^{y\cdot(\omega+i\omega^\perp)/h}\tilde{R}) = -e^{iy\cdot\omega^\perp/h}P_h\tilde{L}. \tag{3.6}$$

Note that $\|e^{iy\cdot\omega^\perp/h}P_h\tilde{L}\|_{L^2(\tilde{\Omega})}=O(h^2)$. Thus Theorem 3.1 implies that

$$\|e^{iy\cdot\omega^\perp/h}\tilde{R}\|_{H^2_h(\tilde{\Omega})}\le Ch, \tag{3.7}$$

which leads to

$$\|\partial^\alpha\tilde{R}\|_{L^2(\tilde{\Omega})}\le Ch^{1-|\alpha|}\quad\text{for}\quad|\alpha|\le 2. \tag{3.8}$$

Remark 3.2. *The leading term $\tilde{L}$ of the CGO solution* (3.2) *is obtained by solving* (3.5). *It is possible to solve* (3.5) *by an iteration scheme, which is numerically feasible. Theorem* 3.1 *is a general theorem to guarantee the existence of the remainder term $\tilde{R}$ in* (3.2). *It may be a nontrivial task to actually find $\tilde{R}$ for general systems. However, since $\tilde{R}$ is $O(h)$ for small h, it could be omitted in numerical computations.*

4. CGO solutions with general phases. In this section we will construct CGO solutions with more general phases for (1.3) from CGO solutions with linear phases given in the previous section. Without loss of generality, we choose $\omega=(1,0)$ and $\omega^\perp=(0,1)$, i.e., $y\cdot(\omega+i\omega^\perp)=y_1+iy_2$. Denote $y=y_1+iy_2$ and $x=x_1+ix_2$. Let Ω_0 be an open subdomain of Ω. Suppose that $A_1,A_2\in C^2(\bar{\Omega}_0)$ and $Q\in L^\infty(\Omega_0)$. Let $y=\rho(x)=y_1(x_1,x_2)+iy_2(x_1,x_2)$ be a conformal map in Ω_0, i.e., $\rho'(x)\neq 0$ for all $x\in\Omega_0$. Define $U(x)=V(y(x))$ and $\tilde{\Omega}=\rho(\Omega_0)$. By straightforward computations, we have

$$\begin{pmatrix}\partial_{x_1}\\ \partial_{x_2}\end{pmatrix}U=J(x)\begin{pmatrix}\partial_{y_1}\\ \partial_{y_2}\end{pmatrix}V\Big|_{y=\rho(x)}\quad\text{and}\quad\Delta_xU=\Delta_yV|\rho'(x)|^2,$$

where

$$J(x)=\begin{pmatrix}\partial_{x_1}y_1 & \partial_{x_1}y_2\\ \partial_{x_2}y_1 & \partial_{x_2}y_2\end{pmatrix}.$$

Suppose that ρ^{-1} exists on $\tilde{\Omega}$. Let $\hat{A}_1(y)=(A_1\partial_{x_1}y_1+A_2\partial_{x_2}y_1)\circ\rho^{-1}(y)$, $\hat{A}_2(y)=(A_1\partial_{x_1}y_2+A_2\partial_{x_2}y_2)\circ\rho^{-1}(y)$, and $\hat{Q}(y)=(Q\circ\rho^{-1})(y)$ and $g(y)=|(\rho'\circ\rho^{-1})(y)|^2$. Now if we choose $V(y)$ satisfying

$$\Delta_yV+g(y)^{-1}\hat{A}_1(y)\partial_{y_1}V+g(y)^{-1}\hat{A}_2(y)\partial_{y_2}V+g(y)^{-1}\hat{Q}V=0\quad\text{in }\tilde{\Omega}, \tag{4.1}$$

then $U(x)$ satisfies (1.3) in Ω_0. According to the construction given previously, let $V(y)$ be a solution of (4.1) having the form

$$V(y)=e^{(y_1+iy_2)/h}(\tilde{L}+\tilde{R}),$$

where

$$\|\partial^\alpha\tilde{R}\|_{L^2(\tilde{\Omega})}\le Ch^{1-\alpha}\quad\forall\ |\alpha|\le 2.$$

Denote $y_1(x_1,x_2)=\varphi(x_1,x_2)$ and $y_2(x_1,x_2)=\psi(x_1,x_2)$. We then obtain CGO solutions for (1.3) in Ω_0:

$$U(x)=e^{(\varphi+i\psi)/h}(L+R)$$

with $L=\tilde{L}\circ\rho$, $R=\tilde{R}\circ\rho$, and

$$\|\partial^\alpha R\|_{L^2(\Omega_0)}\le Ch^{1-\alpha}\quad\forall\ |\alpha|\le 2. \tag{4.2}$$

Due to the conformality of ρ, φ and ψ are harmonic functions in Ω_0. Conversely, given any φ harmonic in Ω_0 with $\nabla\varphi \neq 0$ in Ω_0, we can find a harmonic conjugate ψ of φ in Ω_0 so that $\rho = \varphi + i\psi$ is conformal in Ω_0. The freedom of choosing φ plays a key role in our reconstruction method for the object identification problem. Actually, we will mainly focus on the level curves of φ. We give some concrete examples here.

Pick a point $x_0 \notin \bar{\Omega}$. It is no restriction to assume that $x_0 = 0$. We now consider $\varphi_N = \mathrm{Re}(c_N x^N)$ for $N \geq 2$, where $c_N \in \mathbb{C}$ with $|c_N| = 1$. In the polar coordinates, $\varphi_N(r, \theta) = r^N \cos N(\theta - \theta_N)$ for some θ_N determined by c_N. We observe that $\varphi_N > 0$ in some open cone Γ_N with an opening angle π/N. The freedom of choosing θ_N (or, equivalently, c_N) allows us to "sweep" the domain Ω by Γ_N without moving the point x_0. This is quite useful in practice. Now assume that $\Gamma_N \cap \Omega \neq \emptyset$. The complex function $\rho_N(x) = c_N x^N = \varphi_N + i\psi_N$ is clearly conformal in Ω, where $\psi_N = \mathrm{Im}(c_N x^N)$. In order to apply to the inverse problem, we want to shrink the opening angle of Γ_N by taking $N \to \infty$. However, there are two serious problems in doing so. On one hand, φ_N is periodic in the angular variable, which means that it is positive in some other cones with the same opening angle which also intersect Ω when N is large. Some level curves of φ_N for different N's are shown in Figure 4.1. This property of φ_N prohibits us from using corresponding CGO solutions with large N to the object identification problem. On the other hand, the complex function $\rho_N(x)$ fails to be injective in the whole domain Ω when N is large. To overcome those difficulties and construct useful CGO solutions in the whole domain Ω, we shall carry out the construction described above in a suitable Ω_0 and extend the constructed solutions to Ω by cut-off functions.

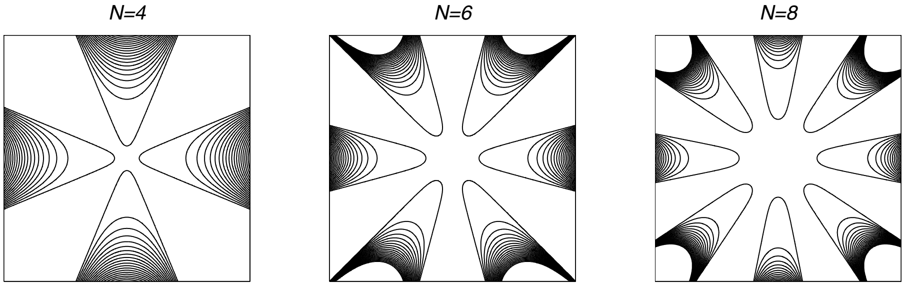


FIG. 4.1. *Some level curves of ϕ_N.*

We now set

$$\Omega_0 := \Gamma_N \cap \Omega.$$

Then ρ_N is conformal in Ω_0 and is bijective from Ω_0 onto $\rho_N(\Omega_0)$. Therefore, we can find CGO solutions for (1.3) in Ω_0,

$$U_{N,h}(x) = e^{(\varphi_N + i\psi_N)/h}(L + R),$$

and the estimate (4.2) holds. So far we have constructed only special solutions for (1.3) in some particular subdomain of Ω. To get solutions in the whole domain Ω, we use a cut-off technique. For $s > 0$, let $\ell_s = \{x \in \Gamma_N : \varphi_N = s^{-1}\}$. This is the level curve of φ_N in Γ_N. Let $0 < t < t_0$ such that

$$\left(\bigcup_{s\in(0,t)} \ell_s\right) \cap \Omega \neq \emptyset$$

and choose a small $\varepsilon > 0$. Define a cut-off function $\phi_{N,t}(x) \in C^\infty(\mathbb{R}^2)$ so that $\phi_{N,t}(x) = 1$ for $x \in \overline{(\cup_{s\in(0,t+\varepsilon/2)}\ell_s) \cap \Omega}$ and is zero for $x \in \bar{\Omega} \setminus (\cup_{s\in(0,t+\varepsilon)}\ell_s)$. We now define

$$U_{N,t,h}(x) = \phi_{N,t}e^{-t^{-1}/h}U_N = \phi_{N,t}e^{(\varphi_N - t^{-1}+i\psi_N)/h}(L+R)$$

for $x \in (\cup_{s\in(0,t+\varepsilon)}\ell_s) \cap \Omega$. So $U_{N,t,h}$ can be regarded as a function in Ω which is zero outside of Ω_0. We now take $f_{N,t,h} = U_{N,t,h}|_{\partial\Omega}$. We remark that $f_{N,t,h}$ can be used as the boundary data in the inverse problem. An obvious reason for using $f_{N,t,h}$ is that they are local.

Now we define a function $W := W_{N,t,h}$ satisfying

$$\begin{cases} \Delta W + A_1(x)\partial_{x_1}W + A_2(x)\partial_{x_2}W + Q(x)W = 0 \quad \text{in} \quad \Omega, \\ W = f_{N,t,h} \quad \text{on} \quad \partial\Omega. \end{cases} \tag{4.3}$$

We would like to compare $W_{N,t,h}$ with $U_{N,t,h}$. It turns out they differ only by an exponentially small term under some minor condition. This property plays an essential role in our method for the inverse problem.

Lemma 4.1. *Assume that the boundary value problem*

$$\begin{cases} PU = 0 \quad \textit{in} \quad \Omega, \\ U = 0 \quad \textit{on} \quad \partial\Omega \end{cases} \tag{4.4}$$

has only a trivial solution. Then there exist $C > 0$ and $\varepsilon' > 0$ such that

$$\|W_{N,t,h} - U_{N,t,h}\|_{H^2(\Omega)} \le Ce^{-\varepsilon'/h} \tag{4.5}$$

for $h \ll 1$.

Proof. By setting $G := W_{N,t,h} - U_{N,t,h}$, we get that

$$\begin{aligned} PG &= P(W_{N,t,h} - U_{N,t,h}) \\ &= -\phi_{N,t}e^{-t^{-1}/h}PU_N + [\phi_{N,t}, P]e^{-t^{-1}/h}U_N \\ &= [\phi_{N,t}, P]e^{-t^{-1}/h}U_N \\ &= [\phi_{N,t}, P]e^{(\varphi_N - t^{-1}+i\psi_N)/h}(L+R) \end{aligned}$$

since $PU_N = 0$ in $(\cup_{s\in(0,t_0)}\ell_s) \cap \Omega$. Now we observe that $[\phi_{N,t}, P]$, the commutator of $\phi_{N,t}$ and P, is a first order differential operator with coefficients supported in

$$\overline{\left(\bigcup_{s\in(t+\varepsilon/2,t+\varepsilon)} \ell_s\right) \cap \Omega}.$$

So we have that

$$\|[\phi_{N,t}, P]e^{(\varphi_N - t^{-1}+i\psi_N)/h}(L+R)\|_{L^2(\Omega)} \le C'e^{-\varepsilon'/h} \tag{4.6}$$

for some $C' > 0$ and $\varepsilon' > 0$. Note that $G = 0$ on $\partial\Omega$. Combining the regularity theorem, the triviality of (4.4), and (4.6) yields (4.5). □

Even though the solutions $W_{N,t,h}$ of (1.3) are not exactly in the form of complex geometrical optics, with the help of Lemma 4.1, they are exponentially close to $U_{N,t,h}$. Now we describe how to construct special solutions for some concrete systems given in section 2 from $W_{N,t,h}$. For the conductivity equation (2.1), (1.3) is reduced to (2.2). For (2.2), we denote the corresponding $U_{N,t,h} = u_{N,t,h}$ and

$$u_{N,t,h} = \phi_{N,t} e^{(\varphi_N - t^{-1} + i\psi_N)/h}(1 + r),$$

where r satisfies (4.2). With $u_{N,t,h}$, we can solve for $w_{N,t,h}$ satisfying

$$\begin{cases} (\Delta + q)w = 0 \quad \text{in} \quad \Omega, \\ w = u_{N,t,h} \quad \text{on} \quad \partial\Omega. \end{cases} \tag{4.7}$$

The problem (4.7) has a unique solution since the boundary value problem for the corresponding conductivity equation has a unique solution. So Lemma 4.1 implies that

$$\|w_{N,t,h} - u_{N,t,h}\|_{H^1(\Omega)} \le Ce^{-\varepsilon'/h}. \tag{4.8}$$

Returning to the conductivity equation, we see that $\gamma^{-1/2} w_{N,t,h}$ are solutions of (2.1).

For the isotropic elasticity and the Stokes system, we have that $n = 3$ and (1.3) become, respectively, (2.7) and (2.10). We discuss only the isotropic elasticity here. The Stokes system can be treated similarly. Assume that the homogeneous boundary value problem (4.4) associated with (2.7) has only the trivial solution. Thus Lemma 4.1 yields

$$\|W_{N,t,h} - U_{N,t,h}\|_{H^2(\Omega)} \le Ce^{-\varepsilon'/h}.$$

We now express $U_{N,t,h} = \binom{v_{N,t,h}}{b_{N,t,h}}$ and $W_{N,t,h} = \binom{w_{N,t,h}}{g_{N,t,h}}$, where $v_{N,t,h}$, $w_{N,t,h}$ are two-dimensional vectors and $b_{N,t,h}$, $g_{N,t,h}$ are scalars. Hence, we obtain that

$$u_{N,t,h} = \mu^{-1/2} w_{N,t,h} + \mu^{-1}\nabla g_{N,t,h} - g_{N,t,h}\nabla\mu^{-1}$$

are solutions of (2.6) or (2.5) and $u_{N,t,h}$ satisfies

$$\|u_{N,t,h} - (\mu^{-1/2} v_{N,t,h} + \mu^{-1}\nabla b_{N,t,h} - b_{N,t,h}\nabla\mu^{-1})\|_{H^1(\Omega)} \le Ce^{-\varepsilon'/h}.$$

5. Inverse problems. In this section we demonstrate how to use CGO solutions constructed previously in the object identification problem. To simplify our presentation, we will discuss only the case of identifying inclusions inside of the domain Ω filled with known conductivity. This inverse problem has been extensively studied both theoretically and numerically. We refer the reader to [10] for related references. Using our method, we can also treat the object identification problem for other systems. We shall report the results elsewhere.

Let D be an open bounded domain with C^1 boundary such that $\bar{D} \subset \Omega$ and $\Omega \setminus \bar{D}$ is connected. Assume $\gamma(x) \in C^2(\bar{\Omega})$ with $\gamma(x) > 0$ for all $x \in \bar{\Omega}$. The conductivity $\tilde{\gamma}(x)$ is a perturbation of γ described by $\tilde{\gamma}(x) = \gamma + \chi_D\gamma_1$, where χ_D is the characteristic function of D and $\gamma_1 \in C(\bar{D})$. We suppose that

$$\gamma_1 \ge 0 \quad \text{in} \quad D \quad \text{and} \quad \gamma_1 > 0 \quad \text{on} \quad \partial D. \tag{5.1}$$

Then we have $\tilde{\gamma}(x) \ge c > 0$ almost everywhere in Ω. Let v be the solution of

$$\begin{cases} \nabla \cdot (\tilde{\gamma}\nabla v) = 0 \quad \text{in} \quad \Omega, \\ v = f \quad \text{on} \quad \partial\Omega. \end{cases} \tag{5.2}$$

The meaning of the solution to (5.2) is understood in the following way. Define

$$[w]_{\partial D} = \mathrm{tr}^+ w - \mathrm{tr}^- w,$$

the jump of the function across ∂D, where tr^+ and tr^- denote, respectively, the trace of w on ∂D from inside and outside of D. For $f \in H^{3/2}(\partial\Omega)$, we define

$$\mathcal{V}_f = \left\{ w \in H^2(D) \oplus H^2(\Omega \setminus \bar{D}) : w|_{\partial\Omega} = f, [w]_{\partial D} = 0, \left[\tilde{\gamma}\frac{\partial w}{\partial \nu}\right]_{\partial D} = 0 \right\}.$$

We say that v is the solution of (5.2) if $v \in \mathcal{V}_f$ and $\nabla \cdot (\tilde{\gamma} v) = 0$ in D and $\Omega \setminus \bar{D}$. The Dirichlet-to-Neumann map is given as

$$\Lambda_D : f \to \tilde{\gamma}\frac{\partial v}{\partial \nu}\Big|_{\partial\Omega},$$

where ν is the unit outer normal of $\partial\Omega$. The inverse problem is to determine the inclusion D from Λ_D. Here we are interested in the reconstruction question.

Since our method shares the same spirit as Ikehata's enclosure method [11], [12], we will briefly describe Ikehata's ideas to motivate our method. Here we take $\gamma \equiv 1$, i.e., $\tilde{\gamma} = 1 + \chi_D \gamma_1$. Denote

$$f_\omega(x, \tau, t) = \exp\{\tau(x \cdot \omega - t) + i\tau x \cdot \omega^\perp\}$$

and

$$I_\omega(\tau, t) = \langle (\Lambda_D - \Lambda_0) f_\omega(\cdot, \tau, t), \overline{f_\omega(\cdot, \tau, t)} \rangle,$$

where Λ_0 is the Dirichlet-to-Neumann map associated with $\Delta u = 0$ in Ω. Let us define

$$h_D(\omega) = \sup_{x \in D} x \cdot \omega.$$

Then the following formulas hold:

$$\left\{ t \in \mathbb{R} : \lim_{\tau \to 0} I_\omega(\tau, t) = 0 \right\} = (h_D(\omega), \infty)$$

and

$$\lim_{\tau \to \infty} \frac{\log |\tau_\omega(\tau, t)|}{2\tau} = h_D(\omega) - t \quad \forall\, t \in \mathbb{R}$$

(see [11], [12]).

To describe our method, we begin with the following integral inequalities given in [19] (also see [10] for a proof).

Lemma 5.1. *Assume that* (5.6) *holds. Let $f \in H^{3/2}(\partial\Omega)$ and u be the unique solution of*

$$\begin{cases} \nabla \cdot (\gamma \nabla u) = 0 & in \quad \Omega, \\ u = f & on \quad \partial\Omega. \end{cases} \tag{5.3}$$

Define $\Lambda_0 : f \to \gamma \frac{\partial u}{\partial \nu}|_{\partial\Omega}$. Then we have

$$\int_{\partial\Omega} (\Lambda_D - \Lambda_0)\bar{f} \cdot f ds \le \int_D \gamma_1 |\nabla u|^2 dx \tag{5.4}$$

and

$$\int_{\partial\Omega}(\Lambda_D - \Lambda_0)\bar{f}\cdot f ds \geq \int_D \frac{\gamma_1\gamma}{\gamma+\gamma_1}|\nabla u|^2 dx. \tag{5.5}$$

It follows from (5.1) that for any $p \in \partial D$, there exists an $\epsilon > 0$ such that

$$\gamma_1 \geq \epsilon \quad \forall\ x \in D \cap B_\epsilon(p). \tag{5.6}$$

Let $x_0 \notin \bar{\Omega}$ and define the open cone Γ_N with $\Gamma_N \cap \Omega \neq \emptyset$ in terms of $\varphi_N = \mathrm{Re}(c_N(x-x_0)^N)$ $(\rho_N = c_N(x-x_0)^N)$ as in Figure 4.1. Likewise, we denote the level curve $\ell_s = \{x \in \Gamma_N : \varphi_N = s^{-1}\}$ for $s > 0$. For $\varepsilon > 0$ and $t > 0$, we take

$$f = f_{N,t,h} = \gamma^{-1/2} w_{N,t,h}|_{\partial\Omega} = \gamma^{-1/2} u_{N,t,h}|_{\partial\Omega}, \tag{5.7}$$

where $w_{N,t,h}$ and $u_{N,t,h}$ are constructed previously. Note that $\gamma^{-1/2}w_{N,t,h}$ is the solution of (5.3). It should be noted that the Dirichlet condition f is localized in $\Gamma_N \cap \partial\Omega$ and supp (f) becomes narrower as N gets bigger. This property is very useful in actual applications.

To construct the inclusion D, we rely on the quantity

$$E(N,t,h) := \int_{\partial\Omega}(\Lambda_D - \Lambda_0)\bar{f}_{N,t,h}\cdot f_{N,t,h} ds. \tag{5.8}$$

Clearly, this quantity is completely determined by the boundary data. From (5.1) and (5.5) we see that

$$E(N,t,h) \geq \int_D \frac{\gamma_1\gamma}{\gamma+\gamma_1}|\nabla(\gamma^{-1/2}w_{N,t,h})|^2 dx \geq 0$$

for all N, t, h. We now prove the following important behavior of $E(N,t,h)$.

THEOREM 5.2. *Let $t > 0$ and $\mathcal{L}_t = \{x \in \Gamma_N : \varphi_N \geq t^{-1}\}$. Then we have the following:*

(i) *if $\mathcal{L}_t \cap \bar{D} = \emptyset$, then there exist $C_1 > 0$, $\varepsilon_1 > 0$, and $h_1 > 0$ such that $E(N,t,h) \leq C_1 e^{-\varepsilon_1/h}$ for all $h \leq h_1$;*

(ii) *if $\mathcal{L}_t \cap D \neq \emptyset$, then there exist $C_2 > 0$, $\varepsilon_2 > 0$, and $h_2 > 0$ such that $E(N,t,h) \geq C_2 e^{\varepsilon_2/h}$ for all $h \leq h_2$.*

Proof. To prove (i), we use the inequality (5.4) to obtain

$$E(N,t,h) \leq \int_D \gamma_1|\nabla(\gamma^{-1/2}w_{N,t,h})|^2 dx \leq C\|w_{N,t,h}\|^2_{H^1(D)}. \tag{5.9}$$

With the help of (4.8), we can replace $w_{N,t,h}$ in (5.9) by $u_{N,t,h}$ with an error $O(e^{-\varepsilon'/h})$. Since $\mathcal{L}_t \cap \bar{D} = \emptyset$, we have $\varphi_N - t^{-1} < 0$ for all $x \in \bar{D} \cap \Gamma_N$. Also, note that $u_{N,t,h} \equiv 0$ in $\Omega \setminus \Gamma_N$. Therefore, by the form of $u_{N,t,h}$ we immediately derive that

$$E(N,t,h) \leq Ce^{-\varepsilon_1/h}$$

for $h \leq h_1$.

To establish (ii), in view of $\mathcal{L}_t \cap D \neq \emptyset$, there exist $z \in \partial D$ and $\epsilon > 0$ such that the jump condition (5.6) holds and

$$\varphi_N - t^{-1} \geq \epsilon \quad \forall\ B_\epsilon(z) \cap D. \tag{5.10}$$

From (5.5) we get

$$\begin{aligned}E(N,t,h) &\geq \int_D \frac{\gamma_1\gamma}{\gamma+\gamma_1}|\nabla(\gamma^{-1/2}w_{N,t,h})|^2dx\\ &\geq C\epsilon\int_{D\cap B_\epsilon(z)}(|\nabla w_{N,t,h}|^2+|w_{N,t,h}|^2)dx\\ &\geq C'\int_{D\cap B_\epsilon(z)}(|\nabla u_{N,t,h}|^2+|u_{N,t,h}|^2)dx - C''e^{-\varepsilon'/h}. \qquad (5.11)\end{aligned}$$

Substituting the form of $u_{N,t,h}$ with the estimate (5.10) into (5.11) implies the statement of (ii). □

THEOREM 5.3. *With the same notation as in Theorem* 5.2*, if* $\ell_t \cap \partial D \neq \emptyset$ *and* $\mathcal{L}_t \cap D = \emptyset$*, then*

$$\liminf_{h\to 0} E(N,t,h) > 0.$$

Recall that $\ell_t = \{x \in \Gamma_N : \varphi_N = t^{-1}\}$.

Proof. In view of (5.6), we pick a sufficiently small $\epsilon > 0$ such that (5.6) is satisfied in $B_\epsilon(p)\cap D$ and $B_\epsilon(p)\cap D \subset (\cup_{s\in(t,t+\varepsilon/2)}\ell_s)\cap D$. So the cut-off function $\phi_{N,t} = 1$ on $B_\epsilon(p)\cap D$. We now introduce a new coordinate system $\Psi(x) = (y_1(x), y_2(x))$ near p with $y_2(x) = \varphi_N - t^{-1}$ such that ℓ_t becomes $y_2 = 0$ near p and $\tilde{D}_\epsilon := \Psi(B_\epsilon(p)\cap D)$ lies in $\{y_2 < 0\}$. We can choose a small cone C_p in $\tilde{D}_\epsilon$ with vertex p and the length of the axis being δ. Denote $J(y)$ the Jacobian of $\Psi^{-1}(y)$. Therefore, using (5.11) we can estimate

$$\begin{aligned}&E(N,t,h)\\ &\geq C'\int_{D\cap B_\epsilon(p)}(|\nabla u_{N,t,h}|^2+|u_{N,t,h}|^2)dx - C''e^{-\varepsilon'/h}\\ &\geq C'\int_{D\cap B_\epsilon(p)}(|\nabla(e^{(\varphi_N-t^{-1}+i\psi_N)/h}(1+r))|^2+|e^{(\varphi_N-t^{-1}+i\psi_N)/h}(1+r)|^2)dx\\ &\quad - C''e^{-\varepsilon'/h}\\ &\geq \frac{\tilde{C}}{h^2}\int_{C_p} e^{2y_2/h}|J|dy_1dy_2 - C''e^{-\varepsilon'/h}\\ &\geq \frac{\tilde{C}'}{h^2}\int_{-\delta}^0 e^{2y_2/h}y_2dy_2 - C''e^{-\varepsilon'/h}\\ &> 0 \quad \text{as} \quad h\to 0. \qquad \square\end{aligned}$$

In view of Theorems 5.2 and 5.3, we are able to reconstruct some part of ∂D by looking into the asymptotic behavior of $E(N,t,h)$ for various t's. More precisely, let

$$t_{D,N} := \sup\left\{t\in(0,\infty) : \lim_{h\to 0}E(N,h,t) = 0\right\};$$

then if $t_{D,N} = \infty$, we have $\Gamma_N \cap D = \emptyset$. On the other hand, if $t_{D,N} < \infty$, then there exists a $p_{D,N} \in \ell_{t_{D,N}} \cap \partial D$.

By taking N arbitrarily large (the opening angle of Γ_N becomes arbitrarily small), we can reconstruct even more information of ∂D. A point p on ∂D is said to be *detectable* if there exists a half-line l starting from p such that l does not intersect ∂D except at p. For example, if D is star-shaped, every point of ∂D is detectable.

COROLLARY 5.4. *Every detectable point of ∂D can be reconstructed from Λ_D.*

Proof. Let p be a detectable point and l be the corresponding half-line. We can choose l which is not tangent to ∂D at p since if the chosen half-line, say l', is tangent to ∂D at p, we can always choose a desired l by perturbing l' a little bit. Assume that $z_0 \in l$ and $z_0 \neq p$. Let L be the straight line containing l. Pick a point $x_0 \in L$ with $\frac{x_0-p}{|x_0-p|} = -\frac{z_0-p}{|z_0-p|}$ and $x_0 \notin \bar{\Omega}$. Let Γ_N be the cone with axis L and vertex x_0 whose opening angle is π/N. For any $N \in \mathbb{N}$, we construct $w_{N,t,h}$, $u_{N,t,h}$, and $f_{N,t,h}$ as above. So we can determine $E(N,t,h)$ from the measurement $\Lambda_D f_{N,t,h}$. Applying Theorems 5.2 and 5.3, we can determine $t_{D,N}$ so that $\ell_{t_{D,N}} \cap \partial D \neq \emptyset$. Then there exists $p_N \in \Gamma_N$, and $\ell_{t_{D,N}} \cap \partial D = p_N$. By taking $N \to \infty$, we can see that $p_N \to p$. □

To end this section, we give an algorithm of our reconstruction method based on Theorem 5.2.

Step 1. Pick a point $x_0 \notin \bar{\Omega}$ (but close to $\bar{\Omega}$). Given $N \in \mathbb{N}$, choose the cone Γ_N which intersects Ω. [Γ_N is defined in section 4]

Step 2. Start with $t > 0$ such that $\ell_t \cap \Omega \neq \emptyset$. Construct $u_{N,t,h}$ and determine the Dirichlet data $f_{N,t,h} = \gamma^{-1/2} u_{N,t,h}|_{\partial\Omega}$. [(5.7)]

Step 3. Compute $E(N,t,h) = \int_{\text{supp }(f_{N,t,h})} (\Lambda_D - \Lambda_0)\bar{f}_{N,t,h} \cdot f_{N,t,h} ds$. [(5.8)]

Step 4. If $E(N,t,h)$ is arbitrarily small, then increase t and repeat Steps 2 and 3; if $E(N,t,h)$ is arbitrarily large, then decrease t and repeat Steps 2 and 3. [Theorem 5.2]

Step 5. Repeat Step 4 to get a good approximation of ∂D in Γ_N. [Theorem 5.2]

Step 6. Move the cone Γ_N around x_0 by taking a different c_N in $\varphi_N = \text{Re}(c_N x^N)$. Repeat Steps 2–5.

Step 7. Choose a larger N and a new cone Γ_N. Repeat Steps 2–6.

Step 8. Pick a different x_0 and repeat Steps 1–7.

6. Numerical results.

We demonstrate some numerical results of our method in this section. Assume that the domain Ω is given by

$$\Omega = \{(x_1, x_2) : -1 < x_1 < 1, -1.01 < x_2 < -0.1\}.$$

We shall use the Dirichlet data localized on $\{(x_1, -1.01) : -1 < x < 1\}$. To set up $\rho_N(x)$, we consider $N = 4$; i.e., the phase function of the CGO solution is $\rho(x) := \rho_4(x)$. In our numerical computations, we use two sweeping schemes. In the first scheme, we fix the reference point x_0 and rotate the "probing cone" (the cone with the vertex at x_0 and the opening angle $\pi/4$). For the second one, we do not rotate the probing cone but move the reference points along the x-axis. More precisely, let the reference point $x_0 = (x_{0,1}, 0)$ for $-1 < x_{0,1} < 1$. In our first scheme, we fix $x_0 = (0,0)$ and rotate the probing cone determined by the shifted angle θ; in the second scheme, we consider different x_0's and choose $\theta = 0$. In other words, for both schemes, we have

$$\rho(x, x_0, \theta) := c(\theta)(x_1 - x_{0,1} + ix_2)^4 = e^{-i4\theta}(x_1 - x_{0,1} + ix_2)^4.$$

Thus, the probing fronts are level curves of $\varphi(x, x_0, \theta) := \text{Re}(\rho(x, x_0, \theta))$. Figure 6.1 shows some probing fronts of $\varphi(x, x_0, \theta)$ with three different θ's and three x_0's, respectively.

We take the background conductivity $\gamma = 1$, and the conductivity inside the inclusion is 4, i.e, $\gamma_1 = 3$. For numerical experiments, we ignore the cut-off function and take

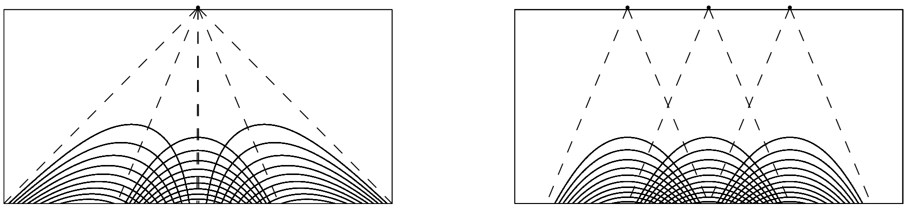


Fig. 6.1. *Probing fronts of our numerical method. In the first column, we consider the probing cone in three different angles. In the second column, we move the probing cone by taking three reference points. In our numerical method, we use* 10 *different probing cones.*

$$g_{x_0,h}|_{\partial\Omega} = \begin{cases} e^{\rho(x,x_0,\theta)/h} & \text{for } (x_1,x_2) \in \partial\Omega_{\text{obs}}, \\ 0, & \partial\Omega \setminus \partial\Omega_{\text{obs}}, \end{cases}$$

where $\partial\Omega_{\text{obs}}$ is determined by x_0 and θ. For example, for $x_0 = (0,0)$ and $\theta = 0$,

$$\partial\Omega_{\text{obs}} = \left\{(x_1,x_2) : -1.01 \times \tan\left(\frac{\pi}{8}\right) < x_1 < 1.01 \times \tan\left(\frac{\pi}{8}\right),\ x_2 = -1.01\right\}.$$

Then for $t > 0$ the required Dirichlet data is given by $f = f_{t,h,x_0} = e^{-t^{-1}/h} g_{x_0,h}$. To get the synthetic data $\Lambda_0 f$ and $\Lambda_D f$, we need to solve the boundary value problems (5.2) and (5.3) with the Dirichlet condition f. To solve these forward problems, we use the PDE Toolbox with the finite element method in MATLAB 7.0. Since we need to collect data on the bottom boundary of Ω, we refine the mesh there; see Figure 6.2.

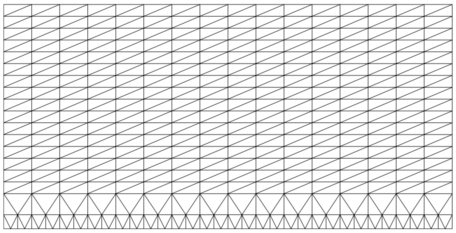


Fig. 6.2. *Example of our finite element method meshes. The mesh has* $2^m + 1$ *nodes on the top boundary and* $2^n + 1$ *nodes on the lower boundary. This example is created with* $m = 4$, $n = 6$. *In solving our forward problems, we choose* $m = 6$, $n = 12$.

To show the effect of noise to our method, we add appropriate noise to the synthetic data. We consider the form of noise given in [10]. To be precise, let $\eta : [-1,1] \mapsto \mathbb{C}$ be a random function defined by

$$\eta(s) = \sum_{k=-32}^{32} (a_k + ib_k) e^{iks\pi/2},$$

where $a_k, b_k \sim \mathcal{N}(0,1)$ are normally distributed random numbers. The number 32 in η is chosen to roughly model a collection of 32 electrodes on the bottom boundary of Ω. Measurement noise is modeled by $\Lambda_D f$ by $\Lambda_D f + c\eta$ with

$$c = \frac{A\|\Lambda_D f\|_\infty}{\|\eta\|_\infty},$$

where $A > 0$.

Our strategy of reconstructing the inclusion is described as follows. We first design M probing cones which are forms by taking either M different vertex points or M different rotating angles. Recall that each cone is congruent to the cone with its vertex at the origin and opening angle $\pi/4$. We then take appropriate h_1 and h_2 with $h_1 > h_2$ and choose a suitable number of probing fronts determined by t_j for $j = 1, \dots, J$ with $t_j < t_{j+1}$. In each probing cone Γ_m $(m = 1, \dots, M)$ given above, we construct the Dirichlet data f supported in the intersection of Γ_m and the bottom boundary of $\partial\Omega$ for every h_k and t_j, $k = 1, 2$, $j = 1, \dots, J$. We now evaluate $E_{j,k} := E(N, t_j, h_k)$ and determine t_n such that

$$E_{n+1,2} > E_{n+1,1}. \tag{6.1}$$

Then the region R_m defined by

$$R_m = \{x \in \Gamma_m : \varphi(x, x_0, \theta) \leq t_n^{-1}\}$$

is the estimated largest region in Γ_m which does not contain the inclusion. So the region $R := \cup_{m=1}^{M} R_m$ is the estimated largest region with the absence of the inclusion with a given sweeping scheme. We would like to point out that condition (6.1) is our rule of thumb in determining whether the level curve $\varphi(x, x_0, \theta) = t^{-1}$ intersects the inclusion in our numerical experiments. It is not equivalent to Theorem 5.2 but is based on the reasoning that $E(N, t, h)$ is *exponentially decaying* when $\varphi(x, x_0, \theta) = t^{-1}$ stays away from the inclusion and *exponentially growing* when $\varphi(x, x_0, \theta) = t^{-1}$ intersects the inclusion. A similar idea was also used in [10].

Our numerical results for each sweeping scheme are shown in Figures 6.3 and 6.4. To save computational time, we show only numerical results obtained from probing

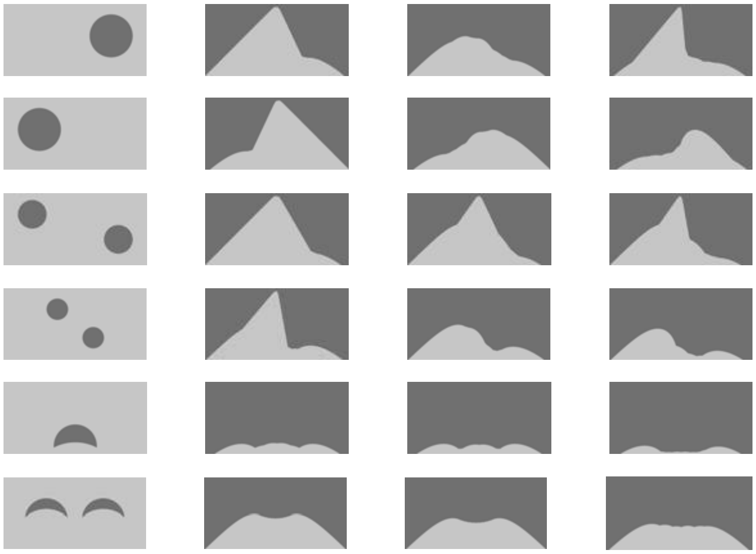


FIG. 6.3. *Numerical results of the first sweeping scheme. All black regions have the conductivity* 4, *and all gray regions have conductivity* 1. *So the gray regions represent the inclusion-free regions. The first column represents the actual location of inclusions. The second column is the theoretical reconstruction when we probe the region only from the bottom. The third column represents the numerical reconstruction from noiseless synthetic data. The fourth column is the numerical reconstruction from data with* 0.01% *noise. To see the effectiveness of our method, we can compare the images in the third column or in the fourth column with those in the second column.*

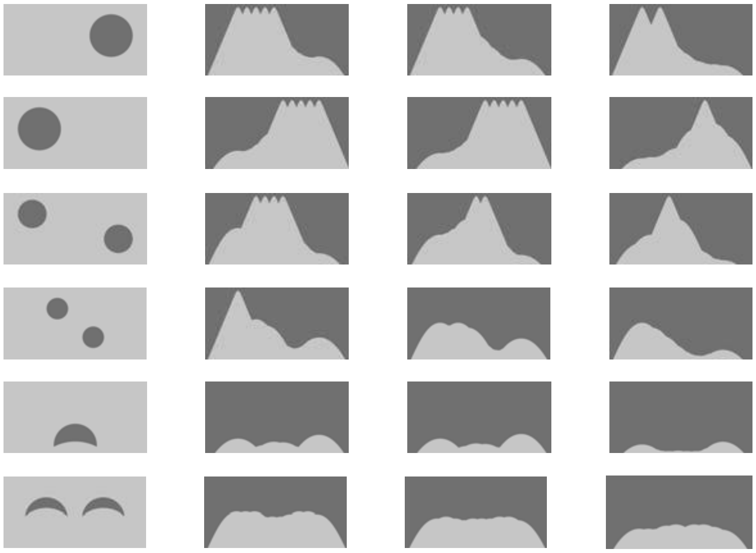


FIG. 6.4. *Numerical results of the second sweeping scheme. All black regions have the conductivity* 4*, and all gray regions have conductivity* 1*. So the gray regions represent the inclusion-free regions. The first column represents the actual location of inclusions. The second column is the theoretical reconstruction when we probe the region only from the bottom. The third column represents the numerical reconstruction from noiseless synthetic data. The fourth column is the numerical reconstruction from data with* 0.01% *noise. To see the effectiveness of our method, we can compare the images in the third column or in the fourth column with those in the second column.*

the region from one side (the bottom part of the boundary). Therefore, the inclusion-free region (with gray color) is near the bottom of the boundary. Since our domain is a rectangle, we can expect to obtain similar results when we probe the region from other sides. We believe that these numerical results are sufficient to demonstrate the applicability of our method.

7. Conclusion.

In this work we present a framework of constructing special complex geometrical optics solutions for several systems of two variables that can be reduced to a system with the Laplacian as the leading term. Here we choose complex polynomials as phase functions. Using these special solutions, we design a novel algorithm to identify embedded objects with boundary measurements. One distinctive feature of our method is that we can probe the region using cones with as small an opening angle as we wish. Theoretically, we are able to reconstruct the exact geometry of the embedded object whose boundary points are all detectable. One typical example is the star-shaped object.

In the numerical experiments, we consider the case of inclusion embedded into a domain with homogeneous conductivity. The numerical results show that our method detects the location of inclusion quite well and is stable under measurements with (small) noise. For computational reasons, we consider only $N = 4$ and use two sweeping schemes separately. It is quite natural to consider higher N's and also combine two sweeping schemes into one. Of course, by doing so, we need to pay the price of increasing computational time.

Our method can be applied to classes of equations or even systems in two dimensions that can be reduced to the Laplacian on the top order part. Its flexibility and effectiveness gives us another technique that can potentially be used in real applications such as medical imaging or nondestructive evaluation.

Acknowledgments. The program codes used in this work are modified from those used in [10]. We would like to thank H. Isozaki, T. Ide, S. Nakata, and S. Siltanen for generously sharing the codes with us. In completing the numerical results of this paper, the second author benefited from conversations with Z. Li, R. LeVeque, G. Nakamura, K. M. Shyue, and C. T. Wu. We also thank the referees and the editor for carefully reading the manuscript and for many valuable suggestions.

REFERENCES

[1] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[2] L. Borcea, *Addendum to: Electrical impedance tomography*, Inverse Problems, 19 (2003), pp. 997–998..

[3] A. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics, Sociedade Brasileira de Matemática, Río de Janeiro, Brazil, 1980, pp. 65–73.

[4] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[5] D. Dos Santos Ferreira, C. E. Kenig, J. Sjöstrand, and G. Uhlmann, *Determining a magnetic Schrödinger operator from partial Cauchy data*, Comm. Math. Phys., 271 (2007), pp. 467–488.

[6] G. Eskin, *Global uniqueness in the inverse scattering problem for the Schrödinger operator with external Yang-Mills potentials*, Comm. Math. Phys., 222 (2001), pp. 503–531.

[7] G. Eskin and J. Ralston, *On the inverse boundary value problem for linear isotropic elasticity*, Inverse Problems, 18 (2002), pp. 907–921.

[8] G. Eskin and J. Ralston, *On the inverse boundary value problem for linear isotropic elasticity and Cauchy-Riemann system*, in Inverse Problems and Spectral Theory, Contemp. Math. 348, AMS, Providence, RI, 2004, pp. 53–69.

[9] H. Heck, G. Uhlmann, and J.-N. Wang, *Reconstruction of obstacles immersed in an incompressible fluid*, Inverse Probl. Imaging, 1 (2007), pp. 63–76.

[10] T. Ide, H. Isozaki, S. Nakata, S. Siltanen, and G. Uhlmann, *Probing for electrical inclusions with complex spherical waves*, Comm. Pure Appl. Math., 60 (2007), pp. 1415–1442.

[11] M. Ikehata, *Reconstruction of the support function for inclusion from boundary measurements*, J. Inverse Ill-Posed Probl., 8 (2000), pp. 367–378.

[12] M. Ikehata and S. Siltanen, *Numerical method for finding the convex hull of an inclusion in conductivity from boundary measurements*, Inverse Problems, 16 (2000), pp. 1043–1052.

[13] M. Ikehata, *A Remark on an Inverse Boundary Value Problem Arising in Elasticity*, preprint.

[14] M. Ikehata, *Mittag-Leffler's function and extracting from Cauchy data*, in Inverse Problems and Spectral Theory, Contemp. Math. 348, AMS, Providence, RI, 2004, pp. 41–52.

[15] M. Ikehata and S. Siltanen, *Electrical impedance tomography and Mittag-Leffler's function*, Inverse Problems, 20 (2004), pp. 1325–1348.

[16] H. Isozaki, *Inverse spectral problems on hyperbolic manifolds and their applications to inverse boundary value problems in Euclidean space*, Amer. J. Math., 126 (2004), pp. 1261–1313.

[17] H. Isozaki and G. Uhlmann, *Hyperbolic geometry and the local Dirichlet-to-Neumann map*, Adv. Math., 188 (2004), pp. 294–314.

[18] J. Jordana, M. Gasulla, and R. Pallas-Areny, *Electrical resistance tomography to detect leaks from buried pipes*, Meas. Sci. Technol., 12 (2001), pp. 1061–1068.

[19] H. Kang, J. K. Seo, and D. Sheen, *The inverse conductivity problem with one measurement: Stability and estimation of size*, SIAM J. Math. Anal., 28 (1997), pp. 1389–1405.

[20] C. E. Kenig, J. Sjöstrand, and G. Uhlmann, *The Calderón problem with partial data*, Ann. of Math. (2), 165 (2007), pp. 567–591.

[21] G. Nakamura and G. Uhlmann, *Global uniqueness for an inverse boundary problem arising in elasticity*, Invent. Math., 118 (1994), pp. 457–474.

[22] G. Nakamura and G. Uhlmann, *Erratum: Global uniqueness for an inverse boundary value problem arising in elasticity*, Invent. Math., 152 (2003), pp. 205–207.

[23] G. Nakamura and G. Uhlmann, *Complex geometric optics solutions and pseudoanalytic matrices*, in Ill-Posed and Inverse Problems, VSP, Zeist, The Netherlands, 2002, pp. 305–338.

[24] A. Ramirez, W. Daily, D. LaBrecque, E. Owen, and D. Chesnut, *Monitoring an underground steam injection process using electrical resistance tomography*, Water Resour. Res., 29 (1993), pp. 73–87.

[25] A. Ramirez, W. Daily, A. Binley, D. LaBrecque, and D. Roelant, *Detection of leaks in underground storage tanks using electrical resistance methods*, J. Envir. Eng. Geophys., 1 (1996), pp. 189–203.

[26] M. Salo and J.-N. Wang, *Complex spherical waves and inverse problems in unbounded domains*, Inverse Problems, 22 (2006), pp. 2299–2309.

[27] L. Slater, A. M. Binley, W. Daily, and R. Johnson, *Cross-hole electrical imaging of a controlled saline tracer injection*, J. Appl. Geophys., 44 (2000), pp. 85–102.

[28] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.

[29] G. Uhlmann, *Developments in inverse problems since Calderón's foundational paper*, in Harmonic Analysis and Partial Differential Equations (Essays in Honor of Alberto P. Calderón), The University of Chicago Press, Chicago, IL, 1999, pp. 295–345.

[30] G. Uhlmann, *Commentary on Calderón's paper: On an inverse boundary value problem*, in Selecta, Papers of Alberto P. Calderón, A. Bellow, C. E. Kenig, and P. Malliavin, eds.

[31] G. Uhlmann and J.-N. Wang, *Complex spherical waves for the elasticity system and probing of inclusions*, SIAM J. Math. Anal., 38 (2007), pp. 1967–1980.

[32] Y. Zou and Z. Guo, *A review of electrical impedance techniques for breast cancer detection*, Med. Eng. Phys., 25 (2003), pp. 79–90.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1045–1079

BIFURCATION ANALYSIS OF A GENERAL CLASS OF NONLINEAR INTEGRATE-AND-FIRE NEURONS*

JONATHAN TOUBOUL†

Abstract. In this paper we define a class of formal neuron models being computationally efficient and biologically plausible, i.e., able to reproduce a wide range of behaviors observed in in vivo or in vitro recordings of cortical neurons. This class includes, for instance, two models widely used in computational neuroscience, the Izhikevich and the Brette–Gerstner models. These models consist of a 4-parameter dynamical system. We provide the full local bifurcation diagram of the members of this class and show that they all present the same bifurcations: an Andronov–Hopf bifurcation manifold, a saddle-node bifurcation manifold, a Bogdanov–Takens bifurcation, and possibly a Bautin bifurcation, i.e., all codimension two local bifurcations in a two-dimensional phase space except the cusp. Among other global bifurcations, this system shows a saddle homoclinic bifurcation curve. We show how this bifurcation diagram generates the most prominent cortical neuron behaviors. This study leads us to introduce a new neuron model, the *quartic model*, able to reproduce among all the behaviors of the Izhikevich and Brette–Gerstner models self-sustained subthreshold oscillations, which are of great interest in neuroscience.

Key words. neuron models, dynamical system analysis, nonlinear dynamics, Hopf bifurcation, saddle-node bifurcation, Bogdanov–Takens bifurcation, Bautin bifurcation, saddle homoclinic bifurcation, subthreshold neuron oscillations

AMS subject classifications. 34C05, 34C23, 34C60, 92B20, 92C20

DOI. 10.1137/070687268

Introduction. During the past few years, in the neurocomputing community, the problem of finding a computationally simple and biologically realistic model of neuron has been widely studied, in order to be able to compare experimental recordings with numerical simulations of large-scale brain models. The key problem is to find a model of neuron realizing a compromise between its simulation efficiency and its ability to reproduce what is observed at the cell level, often considering in-vitro experiments [15, 18, 24].

Among the numerous neuron models, from the detailed Hodgkin–Huxley model [11] still considered as the reference, but unfortunately computationally intractable when considering neuronal networks, down to the simplest integrate-and-fire model [8] very effective computationally, but unrealistically simple and unable to reproduce many behaviors observed, two models seem to stand out [15]: the adaptive quadratic (Izhikevich [14] and related models such as the theta model with adaptation [6, 10]) and exponential (Brette and Gerstner [5]) neuron models. These two models are computationally almost as efficient as the integrate-and-fire model. The Brette–Gerstner model involves an exponential function, which needs to be tabulated if we want the algorithm to be efficient. They are also biologically plausible, and reproduce several important neuronal regimes with a good adequacy with biological data, especially in high-conductance states, typical of cortical in vivo activity. Nevertheless, they fail in reproducing deterministic self-sustained subthreshold oscillations, a behavior of particular interest in cortical neurons for the precision and robustness of spike generation

[*] Received by the editors April 2, 2007; accepted for publication (in revised form) November 12, 2007; published electronically February 29, 2008.
http://www.siam.org/journals/siap/68-4/68726.html

[†] Odyssée Laboratory, INRIA/ENPC/ENS, INRIA, Sophia-Antipolis, 2004 route des Lucioles, BP 93 06902, Sophia-Antipolis Cedex, France (jonathan.touboul@sophia.inria.fr).

patterns, for instance in the inferior olive nucleus [4, 22, 23], in the stellate cells of the entorhinal cortex [1, 2, 17], and in the dorsal root ganglia (DRG) [3, 20, 21]. Some models have been introduced to study from a theoretical point of view the currents involved in the generation of self-sustained subthreshold oscillations [25], but the model failed in reproducing lots of other neuronal behaviors.

The aim of this paper is to define and study a general class of neuron models, containing the Izhikevich and Brette–Gerstner models, from a dynamical systems point of view. We characterize the local bifurcations of these models and show how their bifurcations are linked with different biological behaviors observed in the cortex. This formal study will lead us to define a new model of neuron, whose behaviors include those of the Izhikevich–Brette–Gerstner (IBG) models but also self-sustained subthreshold oscillations.

In the first section of this paper, we introduce a general class of nonlinear neuron models which contains the IBG models. We study the fixed-point bifurcation diagram of the elements of this class, and show that they present the same local bifurcation diagram, with a saddle-node bifurcation curve, an Andronov–Hopf bifurcation curve, a Bogdanov–Takens bifurcation point, and possibly a Bautin bifurcation, i.e., all codimension two bifurcations in dimension two except the cusp. This analysis is applied in the second section to the Izhikevich and the Brette–Gerstner models. We derive their bifurcation diagrams and prove that none of them shows the Bautin bifurcation. In the third section, we introduce a new simple model—the *quartic model*—presenting, in addition to common properties of the dynamical system of this class, a Bautin bifurcation, which can produce self-sustained oscillations. Last, the fourth section is dedicated to numerical experiments. We show that the quartic model is able to reproduce some of the prominent features of biological spiking neurons. We give qualitative interpretations of those different neuronal regimes from the dynamical systems point of view, in order to give a grasp of how the bifurcations generate biologically plausible behaviors. We also show that the new quartic model, presenting supercritical Hopf bifurcations, is able to reproduce the oscillatory/spiking behavior presented, for instance, in the DRG. Finally, we show that numerical simulation results of the quartic model show a good agreement with biological intracellular recordings in the DRG.

1. Bifurcation analysis of a class of nonlinear neuron models.

In this section we introduce a large class of formal neurons which are able to reproduce a wide range of neuronal behaviors observed in cortical neurons. This class of models is inspired by the review made by Izhikevich [15]. He found that the quadratic adaptive integrate-and-fire model was able to simulate efficiently a lot of interesting behaviors. Brette and Gerstner [5] defined a similar model of neuron which presented a good adequacy between simulations and biological recordings.

We generalize these models, and define a new class of neuron models, wide but specific enough to keep the diversity of behaviors of the IBG models.

1.1. The general class of nonlinear models.

In this paper, we are interested in neurons defined by a dynamical system of the type

$$\begin{cases} \frac{\mathrm{d}v}{\mathrm{d}t} = F(v) - w + I, \\ \frac{\mathrm{d}w}{\mathrm{d}t} = a(bv - w), \end{cases}$$

where a, b, and I are real parameters and F is a real function.[1]

In this equation, v represents the membrane potential of the neuron, w is the adaptation variable, I represents the input intensity of the neuron, $1/a$ is the characteristic time of the adaptation variable, and b accounts for the interaction between the membrane potential and the adaptation variable.[2]

This equation is a very general model of neuron. For instance when F is a polynomial of degree three, we obtain a FitzHugh–Nagumo model, when F is a polynomial of degree two the Izhikevich neuron model [14], and when F is an exponential function the Brette–Gerstner model [5]. However, in contrast with continuous models like the FitzHugh–Nagumo model [8], the two latter cases diverge when spiking, and an external reset mechanism is used after a spike is emitted.

In this paper, we want this class of models to have common properties with the IBG neuron models. To this purpose, let us make some assumptions on the function F. The first assumption is a regularity assumption.

Assumption (A1). F is at least three times continuously differentiable.

A second assumption is necessary to ensure us that the system would have the same number of fixed points as the IBG models.

Assumption (A2). The function F is strictly convex.

DEFINITION 1.1 (convex neuron model). *We consider the two-dimensional model defined by the equations*

$$\begin{cases} \frac{dv}{dt} = F(v) - w + I, \\ \frac{dw}{dt} = a(bv - w), \end{cases} \tag{1.1}$$

where F satisfies Assumptions (A1) *and* (A2) *and characterizes the passive properties of the membrane potential.*

Many neurons of this class blow up in finite time. These neurons are the ones we are interested in.

REMARK. *Note that all the neurons of this class do not blow up in finite time. For instance if $F(v) = v\log(v)$, it will not. For F functions such that $F(v) = (v^{1+\alpha})R(v)$ for some $\alpha > 0$, where $\lim_{v\to\infty} R(v) > 0$ (possibly ∞), the dynamical system will possibly blow up in finite time.*

If the solution blows up at time t^*, a spike is emitted, and subsequently we have the following reset process:

$$\begin{cases} v(t^*) = v_r, \\ w(t^*) = w(t^{*-}) + d, \end{cases} \tag{1.2}$$

where v_r is the reset membrane potential and $d > 0$ a real parameter. Equations (1.1) and (1.2), together with initial conditions (v_0, w_0), give us the existence and uniqueness of a solution on $\mathbb{R}^+$.

The two parameters v_r and d are important to understand the repetitive spiking properties of the system. Nevertheless, the bifurcation study with respect to these

[1] The same study can be done for a parameter-dependent function. More precisely, let $E \subset \mathbb{R}^n$ be a parameter space (for a given n) and $F : E \times \mathbb{R} \to \mathbb{R}$ a parameter-dependent real function. All the properties shown in this section are valid for any fixed value of the parameter p. Further p-bifurcations studies can be done for specific $F(p, \cdot)$. The first equation can be derived from the general I-V relation in neuronal models: $C\frac{\mathrm{d}V}{\mathrm{d}t} = I - I_0(V) - g(V - E_K)$, where $I_0(V)$ is the instantaneous I-V curve.

[2] See, for instance, section 2.2, where the parameters of the initial equation (2.2) are related to biological constants and where we proceed to a dimensionless reduction.

parameters is outside the scope of this paper, and we focus here on the bifurcations of the system with respect to (a, b, I), in order to characterize the subthreshold behavior of the neuron.

1.2. Fixed points of the system. To understand the qualitative behavior of the dynamical system defined by (1.1) before the blow up (i.e., between two spikes), we begin by studying the fixed points and analyze their stability. The linear stability of a fixed point is governed by the Jacobian matrix of the system, which we define in the following proposition.

PROPOSITION 1.2. *The Jacobian of the dynamical system* (1.1) *can be written*

$$L := v \mapsto \begin{pmatrix} F'(v) & -1 \\ ab & -a \end{pmatrix}. \tag{1.3}$$

The fixed points of the system satisfy the equations

$$\begin{cases} F(v) - bv + I = 0, \\ bv = w. \end{cases} \tag{1.4}$$

Let $G_b(v) := F(v) - bv$. From (A1) and (A2), we know that the function G_b is strictly convex and has the same regularity as F. To have the same behavior as the IBG models, we want the system to have the same number of fixed points. To this purpose, it is necessary that G_b has a minimum for all $b > 0$. Otherwise, the *convex* function G_b would have no more than one fixed point, since a fixed point of the system is the intersection of an horizontal curve and G_b.

This means for the function F that $\inf_{x\in\mathbb{R}} F'(x) \leq 0$ and $\sup_{x\in\mathbb{R}} F'(x) = +\infty$. Using the monotony property of F', we write Assumption (A3).

Assumption (A3).

$$\begin{cases} \lim\limits_{x\to-\infty} F'(x) \leq 0, \\ \lim\limits_{x\to+\infty} F'(x) = +\infty. \end{cases}$$

Assumptions (A1), (A2), and (A3) ensure us that for all $b \in \mathbb{R}_+^*$, G_b has a unique minimum, denoted $m(b)$, which is reached. Let $v^*(b)$ be the point where this minimum is reached.

This point is the solution of the equation

$$F'(v^*(b)) = b. \tag{1.5}$$

PROPOSITION 1.3. *The point $v^*(b)$ and the value $m(b)$ are continuously differentiable with respect to b.*

Proof. We know that F' is a bijection. The point $v^*(b)$ is defined implicitly by the equation $H(b, v) = 0$, where $H(b, v) = F'(v) - b$. H is a C^1-diffeomorphism with respect to b, and the differential with respect to b never vanishes. The implicit function theorem (see, for instance, [7, Annex C.6]) ensures us that $v^*(b)$ solution of $H(b, v^*(b)) = 0$ is continuously differentiable with respect to b, and so does $m(b) = G(v^*(b)) - bv^*(b)$. □

THEOREM 1.4. *The parameter curve defined by $\{(I, b);\ I = -m(b)\}$ separates three behaviors of the system (see Figure* 1.1*):*

(i) *If $I > -m(b)$, then the system has no fixed point.*

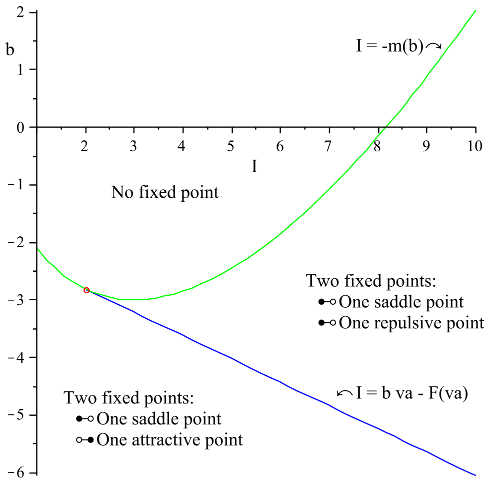


FIG. 1.1. *Number of fixed points and their stability in the plane (I, b) for the exponential adaptive model.*

(ii) *If $I = -m(b)$, then the system has a unique fixed point, $(v^*(b), w^*(b))$, which is nonhyperbolic. It is unstable if $b > a$.*

(iii) *If $I < -m(b)$, then the dynamical system has two fixed points $(v_-(I, b)$, $v_+(I, b))$ such that*

$$v_-(I, b) < v^*(b) < v_+(I, b).$$

The fixed point $v_+(I, b)$ is a saddle fixed point, and the stability of the fixed point $v_-(I, b)$ depends on I and on the sign of $(b - a)$:

(a) *If $b < a$, the fixed point $v_-(I, b)$ is attractive.*

(b) *If $b > a$, there is a unique smooth curve $I^*(a, b)$ defined by the implicit equation $F'(v_-(I^*(a, b), b)) = a$. This curve reads $I^*(a, b) = bv_a - F(v_a)$, where v_a is the unique solution of $F'(v_a) = a$.*

(b.1) *If $I < I^*(a, b)$, the fixed point is attractive.*

(b.2) *If $I > I^*(a, b)$, the fixed point is repulsive.*

Proof.

(i) We have $F(v) - bv \geq m(b)$ by definition of $m(b)$. If $I > -m(b)$, then for all $v \in \mathbb{R}$ we have $F(v) - bv + I > 0$ and the system has no fixed point.

(ii) Let $I = -m(b)$. We have already seen that G_b is strictly convex and continuously differentiable and for $b > 0$ reaches its unique minimum at the point $v^*(b)$. This point is such that $G_b(v^*(b)) = m(b)$, and so it is the only point satisfying $F(v^*(b)) - bv^*(b) - m(b) = 0$.

Furthermore, this point satisfies $F'(v^*(b)) = b$. The Jacobian of the system at this point reads

$$L(v^*(b)) = \begin{pmatrix} b & -1 \\ ab & -a \end{pmatrix}.$$

Its determinant is 0, and so the fixed point is nonhyperbolic (0 is eigenvalue of the Jacobian matrix). The trace of this matrix is $b - a$. So the fixed point $v^*(b)$ is attractive when $b > a$ and repulsive when $b > a$. The case $a = b$,

$I = -m(b)$ is a degenerate case which we will study more precisely in section 1.3.3.

(iii) Let $I < -m(b)$. By the strict convexity assumption, Assumption (A2), of the function G together with Assumption (A3), we know that there are only two intersections of the curve G to a level $-I$ higher than its minimum. These two intersections define our two fixed points. At the point v^* the function is strictly lower than $-I$, and so the two solutions satisfy $v_-(I,b) < v^*(b) < v_+(I,b)$.

Let us now study the stability of these two fixed points. To this end, we have to characterize the eigenvalues of the Jacobian matrix of the system at these points.

We can see from formula (1.3) and the convexity assumption, Assumption (A2), that the Jacobian determinant, equal to $-aF'(v) + ab$, is a decreasing function of v and vanishes at $v^*(b)$, and so $\det(L(v_+(I,b))) < 0$ and the fixed point is a saddle point (the Jacobian matrix has a positive and a negative eigenvalue).

For the other fixed point $v_-(I,b)$, the determinant of the Jacobian matrix is strictly positive. So the stability of the fixed point depends on the trace of the Jacobian. This trace reads $F'\big(v_-(I,b)\big) - a$.

(a) When $b < a$, we have a stable fixed point. Indeed, the function F' is an increasing function equal to b at $v^*(b)$, and so $\text{Trace}\big(L(v_-(I,b))\big) \leq F'(v^*(b)) - a = b - a < 0$ and the fixed point is attractive.

(b) If $b > a$, then the type of dynamics around the fixed point v_- depends on the input current (parameter I). Indeed, the trace reads

$$T(I,b,a) := F'\big(v_-(I,b)\big) - a,$$

which is continuous and continuously differentiable with respect to I and b, and which is defined for $I < -m(b)$. We have

$$\begin{cases} \lim\limits_{I \to -m(b)} T(I,b,a) = b - a > 0, \\ \lim\limits_{I \to -\infty} T(I,b,a) = \lim\limits_{x \to -\infty} F'(x) - a < 0. \end{cases}$$

So there exists a curve $I^*(a,b)$ defined by $T(I,b,a) = 0$ and such that

- for $I^*(b) < I < -m(b)$, the fixed point $v_-(I,b)$ is repulsive;
- for $I < I^*(b)$, the fixed point v_- is attractive.

To compute the equation of this curve, we use the fact that point $v_-(I^*(b),b)$ is such that $F'(v_-(I^*(b),b)) = a$. We know from the properties of F that there is a unique point v_a satisfying this equation. Since $F'(v^*(b)) = b$, $a < b$, and F' is increasing, the condition $a < b$ implies that $v_a < v^*(b)$.

The associated input current satisfies fixed points equation $F(v_a) - bv_a + I^*(a,b) = 0$, or equivalently

$$I^*(a,b) = bv_a - F(v_a).$$

The point $I = I^*(a,b)$ will be studied in detail in the next section, since it is a bifurcation point of the system. □

Figure 1.1 represents the different zones enumerated in Theorem 1.4 and their stability in the parameter plane (I,b).

Remark. *In this proof, we used the fact that F' is invertible on* $[0,\infty)$. *Assumption* (A3) *ensures us that it will be the case and that F has a unique minimum. Assumption* (A3) *is the weakest possible to have this property.*

1.3. Bifurcations of the system. In the study of the fixed points and their stability, we identified two bifurcation curves where the stability of the fixed points changes. The first curve $I = -m(b)$ corresponds to a saddle-node bifurcation and the curve $I = I^*(a,b)$ to an Andronov–Hopf bifurcation. These two curves meet in a specific point, $b = a$ and $I = -m(a)$. This point has a double 0 eigenvalue, and we show that it is a Bogdanov–Takens bifurcation point.

Let us show that the system undergoes these bifurcations with no other assumption than (A1), (A2), and (A3) on F. We also prove that the system can undergo only one other codimension two bifurcation, a Bautin bifurcation.

1.3.1. Saddle-node bifurcation curve. In this section we characterize the behavior of the dynamical system along the curve of equation $I = -m(b)$, and we prove the following theorem.

Theorem 1.5. *The dynamical system* (1.1) *undergoes a saddle-node bifurcation along the parameter curve:*

$$(SN) : \{(b, I) \; ; \; I = -m(b)\}, \tag{1.6}$$

when $F''(v^*(b)) \neq 0$.

Proof. We derive the normal form of the system at this bifurcation point. Following the works of Guckenheimer and Holmes [9] and Kuznetsov [19], we check only the transversality conditions to be sure that the normal form at the bifurcation point will have the expected form.

Let $b \in \mathbb{R}^+$ and $I = -m(b)$. Let $v^*(b)$ be the unique fixed point of the system for these parameters. The point $v^*(b)$ is the unique solution of $F'(v^*(b)) = b$. At this point, the Jacobian matrix (1.3) reads

$$L(v^*(b)) = \begin{pmatrix} b & -1 \\ ab & -a \end{pmatrix}.$$

This matrix has two eigenvalues 0 and $b - a$. The pairs of right eigenvalues and right eigenvectors are

$$0, U := \begin{pmatrix} 1/b \\ 1 \end{pmatrix} \quad \text{and} \quad b - a, \begin{pmatrix} 1/a \\ 1 \end{pmatrix}.$$

Its pairs of left eigenvalues and left eigenvectors are

$$0, V := (-a, 1) \quad \text{and} \quad b - a, (-b, 1).$$

Let $f_{b,I}$ be the vector field

$$f_{b,I}(v, w) = \begin{pmatrix} F(v) - w + I \\ a(bv - w) \end{pmatrix}.$$

The vector field satisfies

$$\begin{aligned} V\left(\frac{\partial}{\partial I} f_{b,I}(v^*(b), w^*(b))\right) &= (-a, 1) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= -a < 0. \end{aligned}$$

So the coefficient of the normal form corresponding to the Taylor expansion along the parameter I does not vanish.

Finally, let us show that the quadratic terms of the Taylor expansion in the normal form does not vanish. With our notations, this condition reads

$$V\Big(D_x^2 f_{b,-m(b)}(v^*(b), w^*(b))(U,U)\Big) \neq 0.$$

This property is satisfied in our framework. Indeed,

$$\begin{aligned}
V\Big(D_x^2 f_{b,-m(b)}(v^*(b), w^*(b))(U,U)\Big) &= V\left(\begin{pmatrix} U_1^2 \dfrac{\partial^2 f_1}{\partial v^2} + 2U_1U_2 \dfrac{\partial^2 f_1}{\partial v \partial w} + U_2^2 \dfrac{\partial^2 f_1}{\partial w^2} \\ U_1^2 \dfrac{\partial^2 f_2}{\partial v^2} + 2U_1U_2 \dfrac{\partial^2 f_2}{\partial v \partial w} + U_2^2 \dfrac{\partial^2 f_2}{\partial w^2} \end{pmatrix}\right)\\
&= V\left(\begin{pmatrix} \frac{1}{b^2} F''(v^*) \\ 0 \end{pmatrix}\right)\\
&= (-a, 1) \cdot \begin{pmatrix} \frac{1}{b^2} F''(v^*) \\ 0 \end{pmatrix}\\
&= -\frac{a}{b^2} F''(v^*) < 0.
\end{aligned}$$

So the system undergoes a saddle-node bifurcation along the manifold $I = -m(b)$. □

Remark. *Note that $F''(v^*(b))$ can vanish only countably many times since F is strictly convex.*

1.3.2. Andronov–Hopf bifurcation curve. In this section we consider the behavior of the dynamical system along the parameter curve $I = I^*(b)$, and we consider the fixed point v_-.

Theorem 1.6. *Let $b > a$, v_a be the unique point such that $F'(v_a) = a$ and $A(a,b)$ be defined by the formula*

$$A(a,b) := F'''(v_a) + \frac{1}{b-a}\left(F''(v_a)\right)^2. \tag{1.7}$$

If $F''(v_a) \neq 0$ and $A(a,b) \neq 0$, then the system undergoes an Andronov–Hopf bifurcation at the point v_a, along the parameter line

$$(AH) := \Big\{(b, I)\ ;\ b > a \ \text{ and } \ I = bv_a - F(v_a)\Big\}. \tag{1.8}$$

This bifurcation is subcritical if $A(a,b) > 0$ and supercritical if $A(a,b) < 0$.

Proof. The Jacobian matrix at the point v_a reads

$$L(v_a) = \begin{pmatrix} a & -1 \\ ab & -a \end{pmatrix}.$$

Its trace is 0 and its determinant is $a(b-a) > 0$, and so the matrix at this point has a pair of pure imaginary eigenvalues $(i\omega, -i\omega)$, where $\omega = \sqrt{a(b-a)}$. Along the curve of equilibria when I varies, the eigenvalues are complex conjugates with real part $\mu(I) = \frac{1}{2}\operatorname{Tr}\big(L(v_-(I,b))\big)$ which vanishes at $I = I^*(a,b)$.

We recall that from Proposition 1.3, this trace varies smoothly with I. Indeed, $v_-(b,I)$ satisfies $F(v_-(I,b)) - bv_-(I,b) + I = 0$ and is differentiable with respect to I. We have

$$\frac{\partial v_-(I,b)}{\partial I}(F'(v_-(I,b)) - b) = -1.$$

At the point $v_-(I^*(b), b) = v_a$, we have $F'(v_a) = a < b$, and so for I close to this equilibrium point, we have

$$\frac{\partial v_-(I,b)}{\partial I} > 0.$$

Now let us check that the transversality condition of an Andronov–Hopf bifurcation is satisfied (see [9, Theorem 3.4.2]). There are two conditions to be satisfied: the transversality condition $\frac{\mathrm{d}\mu(I)}{\mathrm{d}I} \neq 0$ and the nondegeneracy condition $l_1 \neq 0$, where l_1 is the first Lyapunov coefficient at the bifurcation point.

First of all, we prove that the transversality condition is satisfied:

$$\begin{aligned}\mu(I) &= \frac{1}{2}\mathrm{Tr}(L(v_-(I,b)))\\ &= \frac{1}{2}(F'(v_-(I,b)) - a),\\ \frac{\mathrm{d}\mu(I)}{\mathrm{d}I} &= \frac{1}{2}F''(v_-(I,b))\frac{\mathrm{d}v_-(I,b)}{\mathrm{d}I}\\ &> 0.\end{aligned}$$

Let us now write the normal form at this point. To this purpose, we change variables:

$$\begin{cases} v - v_a = x,\\ w - w_a = ax + \omega y.\end{cases}$$

The (x, y) equation reads

$$\text{(1.9)} \qquad \begin{cases} \dot{x} = -\omega y + (F(x + v_a) - ax - w_a) =: -\omega y + f(x),\\ \dot{y} = \omega x + \frac{a}{\omega}(ax - F(x + v_a) + w_a - I) =: \omega x + g(x).\end{cases}$$

According to Guckenheimer in [9], we state that the Lyapunov coefficient of the system at this point has the same sign as B, where B is defined by

$$B := \frac{1}{16}[f_{xxx}+f_{xyy}+g_{xxy}+g_{yyy}]+\frac{1}{16\omega}[f_{xy}(f_{xx}+f_{yy})-g_{xy}(g_{xx}+g_{yy})-f_{xx}g_{xx}+f_{yy}g_{yy}].$$

Replacing f and g by the expressions found in (1.9), we obtain the expression of A:

$$\begin{aligned}B &= \frac{1}{16}F'''(v_a) + \frac{a}{16\omega^2}(F''(v_a))^2\\ &= \frac{1}{16}F'''(v_a) + \frac{1}{16(b-a)}(F''(v_a))^2\\ &= \frac{1}{16}A(a,b).\end{aligned}$$

Hence when $A(a, b) \neq 0$, the system undergoes an Andronov–Hopf bifurcation. When $A(a, b) > 0$, the bifurcation is subcritical and the periodic orbits generated by the Hopf bifurcation are repelling, and when $A(a, b) < 0$, the bifurcation is supercritical and the periodic orbits are attractive (the formula of A has also been introduced by Izhikevich in [16, eq. (15), p. 213]). □

Remark. *The case $A(a,b) = 0$ is not treated in the theorem and is a little bit more intricate. We fully treat it in section* 1.3.4 *and show that a Bautin (generalized Hopf) bifurcation can occur if the A-coefficient vanishes. Since the third derivative is a priori unconstrained, this case can occur, and we prove in section* 3 *that this is the case for a simple (quartic) model.*

1.3.3. Bogdanov–Takens bifurcation.

We have seen in the study that this formal model presents an interesting point in the parameter space, corresponding to the intersection of the saddle-node bifurcation curve and the Andronov–Hopf bifurcation curve. At this point, we show that the system undergoes a Bogdanov–Takens bifurcation.

Theorem 1.7. *Let F be a real function satisfying Assumptions* (A1), (A2), *and* (A3). *Let $a \in \mathbb{R}_+^*$ and $b = a$, and let v_a be the only point such that $F'(v_a) = a$. Assume again that $F''(v_a) \neq 0$.*

Then at this point and with these parameters, the dynamical system (1.1) *undergoes a subcritical Bogdanov–Takens bifurcation of normal form:*

$$\begin{cases} \dot\eta_1 = \eta_2, \\ \dot\eta_2 = \left(\frac{8F''(v_a)\,a\,I_1}{(a+b_1)^3}\right) - \left(\frac{2(2\,b_1\,a+I_1\,F''(v_a))}{(a+b_1)^2}\right)\eta_1 + \eta_1^2 + \eta_1\eta_2 + \mathcal{O}(\|\eta\|^3), \end{cases} \tag{1.10}$$

where $b_1 := b - a$ and $I_1 = I + m(a)$.

Proof. The Jacobian matrix (1.3) at this point reads

$$L(v_a) = \begin{pmatrix} a & -1 \\ a^2 & -a \end{pmatrix}.$$

This matrix is nonzero and has two 0 eigenvalues (its determinant and trace are 0). The matrix $Q := \left(\begin{smallmatrix} a & 1 \\ a^2 & -a \end{smallmatrix}\right)$ is the passage matrix to the Jordan form of the Jacobian matrix:

$$Q^{-1} \cdot L(v_a) \cdot Q = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

To prove that the system undergoes a Bogdanov–Takens bifurcation, we show that the normal form reads

$$\begin{cases} \dot\eta_1 = \eta_2, \\ \dot\eta_2 = \beta_1 + \beta_2\eta_1 + \eta_1^2 + \sigma\eta_1\eta_2 + \mathcal{O}(\|\eta\|^3) \end{cases} \tag{1.11}$$

with $\sigma = \pm 1$. The proof of this theorem consists of (i) proving that the system undergoes a Bogdanov–Takens bifurcation, (ii) finding a closed-form expression for the variables β_1 and β_2, and (iii) proving that $\sigma = 1$.

First of all, let us prove that the normal form can be written in the form of (1.11). This is equivalent to showing some transversality conditions on the system (see, for instance, [19, Theorem 8.4]).

To this end, we center the equation at this point and write the system in the coordinates given by the Jordan form of the matrix. Let $\binom{y_1}{y_2} = Q^{-1}\binom{v-v_a}{w-w_a}$ at the point $b = a + b_1$, $I = -m(a) + I_1$. We get

$$\begin{cases} \dot y_1 = y_2 + \frac{b_1}{a}(ay_1 + y_2), \\ \dot y_2 = F(ay_1 + y_2 + v_a) - w_a - m(a) + I_1 - a^2 y_1 - ay_2 - b_1(ay_1 + y_2). \end{cases} \tag{1.12}$$

Let us denote $v_1 = ay_1 + y_2$. The Taylor expansion on the second equation gives us

$$\begin{aligned}
\dot{y}_2 &= F(v_1 + v_a) - w_a - m(a) + I_1 - a^2 y_1 - a y_2 - b_1(ay_1 + y_2)\\
&= F(v_a) + F'(v_a)v_1 + \frac{1}{2}F''(v_a)v_1^2 - w_a - m(a)\\
&+ I_1 - a^2 y_1 - a y_2 - b_1(ay_1 + y_2) + \mathcal{O}(\|v_1\|^3)\\
&= (F(v_a) - w_a - m(a)) + I_1 + (F'(v_a) - a)v_1 - b_1 v_1 + \frac{1}{2}F''(v_a)v_1^2\\
&+ \mathcal{O}(\|v_1\|^3)\\
&= I_1 - b_1(ay_1 + y_2) + \frac{1}{2}F''(v_a)(ay_1 + y_2)^2 + \mathcal{O}(\|y\|^3). \qquad (1.13)
\end{aligned}$$

Let us denote for the sake of clarity $\alpha = (b_1, I_1)$ and write (1.12) as

$$\begin{cases} \dot{y}_1 = y_2 + a_{00}(\alpha) + a_{10}(\alpha)y_1 + a_{01}(\alpha)y_2,\\ \dot{y}_2 = b_{00}(\alpha) + b_{10}(\alpha)y_1 + b_{01}(\alpha)y_2 + \frac{1}{2}b_{20}(\alpha)y_1^2 + b_{11}(\alpha)y_1y_2 + \frac{1}{2}b_{02}(\alpha)y_2^2 + \mathcal{O}(\|y\|^3). \end{cases} \qquad (1.14)$$

From (1.12) and (1.13), it is straightforward to identify the expressions for the coefficients $a_{ij}(\alpha)$ and $b_{ij}(\alpha)$.

Let us now use the change of variables:

$$\begin{cases} u_1 = y_1,\\ u_2 = y_2 + \frac{b_1}{a}(ay_1 + y_2). \end{cases}$$

The dynamical system governing (u_1, u_2) reads

$$\begin{cases} \dot{u}_1 = u_2,\\ \dot{u}_2 = (1 + \frac{b_1}{a}) - b_1\, a\, u_1 + \frac{1}{2}\frac{a^3F''(v_a)}{a+b_1}u_1^2 + \frac{a^2F''(v_a)}{a+b_1}u_1\, u_2 + \frac{1}{2}\frac{aF''(v_a)}{a+b_1}u_2^2. \end{cases}$$

The transversality conditions of a Bogdanov–Takens bifurcation [9, 19] can easily be verified from this expression:

(BT.1) The Jacobian matrix is not 0.

(BT.2) With the notations of (1.14), we have $a_{20} = 0$ and $b_{11}(0) = aF''(v_a) > 0$, and so $a_{20}(0) + b_{11}(0) = aF''(v_a) > 0$.

(BT.3) $b_{20} = a^2F''(v_a) > 0$.

(BT.4) We show that the map

$$\left(x := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \alpha := \begin{pmatrix} I_1 \\ b_1 \end{pmatrix}\right) \mapsto \Big[f(x, \alpha), \mathrm{Tr}\big(D_x f(x, \alpha)\big), \mathrm{Det}\big(Dxf(x, \alpha)\big)\Big]$$

is regular at the point of interest.

From the two first assumptions, we know that the system can be put in the form of (1.11). Guckenheimer in [9] proves that this condition can be reduced to the nondegeneracy of the differential with respect to (I_1, b_1) of the vector $\binom{\beta_1}{\beta_2}$ of (1.11).

In our case, we can compute these variables β_1 and β_2 following the calculation steps of [19], and we get

$$\begin{cases} \beta_1 = \frac{8F''(v_a)\, a\, I_1}{(a+b_1)^3},\\ \beta_2 = -\frac{2(2\, b_1\, a + I_1\, F''(v_a))}{(a+b_1)^2}. \end{cases} \qquad (1.15)$$

Hence the differential of the vector $\binom{\beta_1}{\beta_2}$ with respect to the parameters (I_1, b_1) at the point $(0,0)$ reads

$$D_\alpha \beta|_{(0,0)} = \begin{pmatrix} \frac{8F''(v_a)}{a^2} & 0 \\ -2\frac{F''(v_a)}{a^2} & -4/a \end{pmatrix}.$$

This matrix has a nonzero determinant if and only if $F''(v_a) \neq 0$.

Therefore we have proved the existence of a Bogdanov–Takens bifurcation under the condition $F''(v_a) \neq 0$.

Let us now show that $\sigma = 1$. Indeed, this coefficient is given by the sign of $b_{20}(0)\big(a_{20}(0) + b_{11}(0)\big)$ which in our case is equal to $a^3 F''(v_a)^2 > 0$, and so the bifurcation is always of the type (1.10) (generation of an unstable limit cycle) for all the members of our class of models. □

The existence of a Bogdanov–Takens bifurcation point implies the existence of a smooth curve corresponding to a saddle homoclinic bifurcation in the system (see [19, Lemma 8.7]).

Corollary 1.8. *There is a unique smooth curve (P) corresponding to a saddle homoclinic bifurcation in the system (1.1) originating at the parameter point $b = a$ and $I = -m(a)$ defined by the implicit equation:*

$$(P) := \Bigg\{ (I = -m(a) + I_1,\, b = a + b_1)\ ;$$

$$I_1 = \frac{\left(-\frac{25}{6}\, a - \frac{37}{6}\, b_1 + \frac{5}{6}\sqrt{25\, a^2 + 74\, b_1\, a + 49\, {b_1}^2}\right) a}{F''(v_a)} + o(|\, b_1\,| + |\, I_1\,|) \tag{1.16}$$

$$\textit{and } b_1 > -\frac{I_1 F''(v_a)}{2a} \Bigg\}.$$

Moreover, for (b, I) in a neighborhood of $(a, -m(a))$, the system has a unique and hyperbolic unstable cycle for parameter values inside the region bounded by the Hopf bifurcation curve and the homoclinic bifurcation curve (P), and it has no cycle outside this region.

Proof. As noticed, from the Bogdanov–Takens bifurcation point, we have the existence of this saddle homoclinic bifurcation curve. Let us now compute the equation of this curve in the neighborhood of the Bogdanov–Takens point. To this purpose we use the normal form we derived in Theorem 1.7 and use the local characterization given, for instance, in [19, Lemma 8.7] for the saddle homoclinic curve:

$$(P) := \left\{ (\beta_1, \beta_2)\ ;\ \beta_1 = -\frac{6}{25}\beta_2^2 + o(\beta_2^2),\, \beta_2 < 0 \right\}.$$

Using the expressions (1.15) yields

$$(P) := \Bigg\{ (I = -m(a) + I_1,\, b = a + b_1)\ ;$$

$$\frac{8F''(v_a) a I_1}{(a+b_1)^3} = \frac{24}{25}\frac{(2\, b_1\, a + I_1\, F''(v_a))^2}{(a+b_1)^4} + o(|\, b_1\,| + |\, I_1\,|)$$

$$\text{and } b_1 > -\frac{I_1 F''(v_a)}{2a} \Bigg\}.$$

We can solve this equation. There are two solutions but only one satisfying $I_1 = 0$ when $b_1 = 0$. This solution is the curve of saddle homoclinic bifurcations. $\square$

1.3.4. Formal conditions for a Bautin bifurcation. In the study of the Andronov–Hopf bifurcation, we showed that the sub- or supercritical type of bifurcation depended on the variable $A(a, b)$ defined by (1.7). If this variable changes sign when b varies, then the stability of the limit cycle along Hopf bifurcation changes stability. This can occur if the point v_a satisfies the following condition.

Assumption (A4). For v_a such that $F'(v_a) = a$, we have

$$F'''(v_a) < 0.$$

Indeed, if this happens, the type of Andronov–Hopf bifurcation changes, since we have

$$\begin{cases} \lim\limits_{b \to a^-} A(a, b) = +\infty, \\ \lim\limits_{b \to +\infty} A(a, b) = F'''(v_a) < 0. \end{cases}$$

In this case the first Lyapunov exponent vanishes for

$$b = a - \frac{(F''(v_a))^2}{F'''(v_a)}.$$

At this point, the system has the characteristics of a Bautin (generalized Hopf) bifurcation. Nevertheless, we still have to check two nondegeneracy conditions to ensure that the system actually undergoes a Bautin bifurcation:

(BGH.1) The second Lyapunov coefficient of the dynamical system l_2 does not vanish at this equilibrium point.

(BGH.2) Let $l_1(I, b)$ be the first Lyapunov exponent of this system and $\mu(I, b)$ the real part of the eigenvalues of the Jacobian matrix. The map

$$(I, b) \mapsto (\mu(I, b), l_1(I, b))$$

is regular at this point.

In this case the system would be locally topologically equivalent to the normal form:

$$\begin{cases} \dot{y}_1 = \beta_1 y_1 - y_2 + \beta_2 y_1(y_1^2 + y_2^2) + \sigma y_1(y_1^2 + y_2^2)^2, \\ \dot{y}_2 = \beta_1 y_2 - y_1 + \beta_2 y_2(y_1^2 + y_2^2) + \sigma y_2(y_1^2 + y_2^2)^2. \end{cases}$$

We reduce the problem to the point that checking the two conditions of a BGH bifurcation becomes straightforward.

Let (v_a, w_a) be the point where the system undergoes the Bautin bifurcation (when it exists). Since we already computed the eigenvalues and eigenvectors of the Jacobian matrix along the Andronov–Hopf bifurcation curve, we can use it to reduce the problem. The basis where we express the system is given by

$$\begin{cases} Q := \begin{pmatrix} \frac{1}{b} & \frac{\omega}{ab} \\ 1 & 0 \end{pmatrix}, \\ \binom{x}{y} := Q^{-1} \binom{v - v_a}{w - w_a}. \end{cases}$$

Let us write the dynamical equations satisfied by (x, y):

$$\begin{cases} \dot{x} = \omega y, \\ \dot{y} = \frac{ab}{\omega}\left(F\big(v_a + \frac{1}{b}x + \frac{\omega}{ab}y\big) - w_a - x + I_a - ay\right). \end{cases}$$

To ensure that we have a Bautin bifurcation at this point we will need to perform a Taylor expansion up to the fifth order, and so we need to make the following assumption.

Assumption (A5). The function F is six times continuously differentiable at (v_a, w_a).

First, let us denote $v_1(x, y) = \frac{1}{b}x + \frac{\omega}{ab}y$; the Taylor expansion reads

$$\begin{aligned} \dot{y} = {} & \frac{ab}{\omega}\big(F(v_a) - w_a + I\big) + \frac{ab}{\omega}\left[F'(v_a)v_1(x, y) - ay\right] + \frac{1}{2}\frac{ab}{\omega}\left[F''(v_a)v_1(x, y)^2\right] \\ & + \frac{1}{6}\frac{ab}{\omega}F'''(v_a)v_1(x, y)^3 + \frac{1}{4!}\frac{ab}{\omega}F^{(4)}(v_a)v_1(x, y)^4 \\ & + \frac{1}{5!}\frac{ab}{\omega}F^{(5)}(v_a)v_1(x, y)^5 + \mathcal{O}\left(\left\|\begin{pmatrix} x \\ y \end{pmatrix}\right\|^6\right). \end{aligned}$$

This expression, together with the complex left and right eigenvectors of the Jacobian matrix, allows us to compute the first and second Lyapunov coefficients and to check the existence of a Bautin bifurcation.

Nevertheless, we cannot push the computation any further at this level of generality, but, for a given function F presenting a change in the sign of $A(a, b)$, this can easily be done through the use of a symbolic computation package. The interested reader is referred to Appendix A for checking the Bautin bifurcation transversality conditions, where calculations are given for the quartic neuron model.

1.4. Conclusion: The full bifurcation diagram. We now summarize the results obtained in this section in the two following theorems.

THEOREM 1.9. *Let us consider the formal dynamical system*

$$\begin{cases} \dot{v} = F(v) - w + I, \\ \dot{w} = a(bv - w), \end{cases} \tag{1.17}$$

where a is a fixed real, b and I bifurcation parameters, and $F : \mathbb{R} \mapsto \mathbb{R}$ a real function. If the function F satisfies the assumptions that

(A.1) *the function F is three times continuously differentiable,*

(A.2) *F is strictly convex, and*

(A.3) *F' satisfies the conditions*

$$\begin{cases} \lim\limits_{x \to -\infty} F'(x) \leq 0, \\ \lim\limits_{x \to \infty} F'(x) = \infty, \end{cases}$$

then the dynamical system (1.17) *shows the following bifurcations:*

(B1) *A saddle-node bifurcation curve:*

$$(SN) : \{(b, I)\ ;\ I = -m(b)\},$$

where $m(b)$ is the minimum of the function $F(v) - bv$ (if the second derivative of F does not vanish at this point).

(B2) *An Andronov–Hopf bifurcation line:*

$$(AH) := \Big\{(b, I)\ ;\ b > a \ \ and\ \ I = bv_a - F(v_a)\Big\},$$

where v_a is the unique solution of $F'(v_a) = a$ (if $F''(v_a) \neq 0$). This type of Andronov–Hopf bifurcation is given by the sign of the variable

$$A(a, b) = F'''(v_a) + \frac{1}{b-a} F''(v_a)^2.$$

If $A(a, b) > 0$, then the bifurcation is subcritical, and if $A(a, b) < 0$, then the bifurcation is supercritical.

(B3) *A Bogdanov–Takens bifurcation point at the point $b = a$ and $I = -m(a)$ if $F''(v_a) \neq 0$.*

(B4) *A saddle homoclinic bifurcation curve characterized in the neighborhood of the Bogdanov–Takens point by*

$$(P) := \Bigg\{(I = -m(a) + I_1,\ b = a + b_1)\ ;$$
$$I_1 = \frac{\left(-\frac{25}{6}\,a - \frac{37}{6}\,b_1 + \frac{5}{6}\sqrt{25\,a^2 + 74\,b_1\,a + 49\,{b_1}^2}\right) a}{F''(v_a)} + o(|\,b_1\,| + |\,I_1\,|)$$
$$and\ b_1 > -\frac{I_1 F''(v_a)}{2a}\Bigg\}.$$

Theorem 1.10. *Consider the system* (1.1)*, where a is a given real number, b and I are real bifurcation parameters, and $F : E \times \mathbb{R} \mapsto \mathbb{R}$ is a function satisfying the following assumptions:*

(A.5) *The function F is six times continuously differentiable.*

(A.2) *F is strictly convex.*

(A.3) *F' satisfies the conditions*

$$\begin{cases} \lim_{x\to-\infty} F'(x) \leq 0, \\ \lim_{x\to\infty} F'(x) = \infty. \end{cases}$$

(A.4) *Let v_a be the unique real such that $F'(v_a) = a$. We have*

$$F'''(v_a) < 0.$$

Furthermore, consider the following conditions:

(BGH.1) *The second Lyapunov coefficient of the dynamical system $l_2(v_a) \neq 0$.*

(BGH.2) *Let $l_1(v)$ denote the first Lyapunov exponent and $\lambda(I, b) = \mu(I, b) \pm i\omega(I, b)$ the eigenvalues of the Jacobian matrix in the neighborhood of the point of interest. The map $(I, b) \to (\mu(I, b), l_1(I, b))$ is regular at this point.*

Having these, the system undergoes a Bautin bifurcation at the point v_a for the parameters $b = a - \frac{F''(v_a)^2}{F'''(v_a)}$ and $I = bv_a - F(v_a)$.

Remark. *Theorem* 1.9 *enumerates some of the bifurcations that any dynamical system of the class* (1.1) *will always undergo. Together with Theorem* 1.10*, they summarize all the local bifurcations the system can undergo, and no other fixed-point bifurcation is possible. In section* 3 *we introduce a model actually showing all these local bifurcations.*

2. Applications: Izhikevich and Brette–Gerstner models. In this section we show that the neuron models proposed by Izhikevich in [14] and Brette and Gerstner in [5] are part of the class studied in section 1. Using the results of the latter section, we derive their bifurcation diagram and obtain that they show exactly the same types of bifurcations.

2.1. Izhikevich quadratic adaptive model. We produce here a complete description of the bifurcation diagram of the adaptive quadratic integrate-and-fire model proposed by Izhikevich in [14] and [16, Chapter 8]. We use here the dimensionless equivalent version of this model with the fewest parameters:

$$\begin{cases} \dot{v} = v^2 - w + I, \\ \dot{w} = a(bv - w). \end{cases} \tag{2.1}$$

Equation (2.1) is clearly a particular case of (1.1) with

$$F(v) = v^2.$$

F is clearly strictly convex and C^∞. $F'(v) = 2v$, and so it also satisfies Assumption (A3). Furthermore, the second derivative never vanishes, and so the system undergoes the three bifurcations stated in Theorem 1.9.

(Izh.B1) A saddle-node bifurcation curve defined by

$$\left\{ (b, I) \; ; \; I = \frac{b^2}{4} \right\}.$$

For $(I, b) \in \mathbb{R}^2$, the fixed point is given by $(v^*(b) = \frac{1}{2}b, \; w^*(b) = \frac{1}{2}b^2)$.
For $I < \frac{b^2}{4}$, the fixed point(s) are

$$v_\pm(b, I) = \frac{1}{2}\big(b \pm \sqrt{b^2 - 4I}\big).$$

(Izh.B2) An Andronov–Hopf bifurcation line:

$$\left\{ (I, b) \; ; \; b > a \;\text{ and }\; I = \frac{a}{2}\left(b - \frac{a}{2}\right) \right\},$$

whose type is given by the sign of the variable

$$A(a, b) = \frac{4}{b - a}.$$

This value is always strictly positive, and so the bifurcation is always subcritical.

(Izh.B3) A Bogdanov–Takens bifurcation point for $b = a$ and $I = \frac{a^2}{4}$, $v_a = \frac{a}{2}$.

(Izh.B4) A saddle homoclinic bifurcation curve satisfying the quadratic equation near the Bogdanov–Takens point:

$$\begin{aligned} (P) := \Bigg\{ & \left(I = \frac{a^2}{2} + I_1, \, b = a + b_1 \right) \; ; \\ & I_1 = \frac{a}{2}\left(-\frac{25}{6}\, a - \frac{37}{6}\, b_1 + \frac{5}{6} \sqrt{25\, a^2 + 74\, b_1\, a + 49\, {b_1}^2} \right) + o(|\, b_1 \,| + |\, I_1 \,|) \\ & \text{and } b_1 > -\frac{I_1}{a} \Bigg\}. \end{aligned}$$

Figure 2.1 represents the fixed points of this dynamical system, and their stability, together with the bifurcation curves.

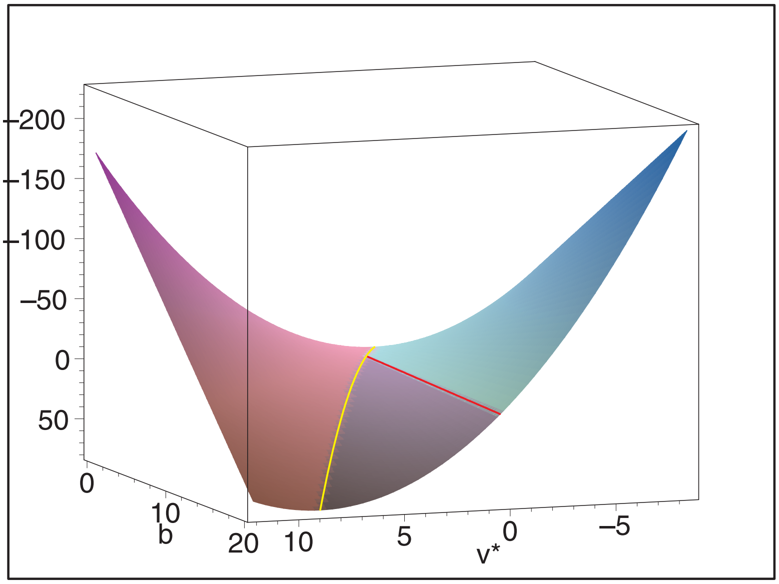


Fig. 2.1. *Representation of the v fixed point with respect to the parameters I and b in the Izhikevich model. The reddish component is the surface of saddle fixed points, the purplish one corresponds to the repulsive fixed points, and the greenish/bluish one corresponds to the attractive fixed points The yellow curve corresponds to a saddle-node bifurcation and the red one to an Andronov–Hopf bifurcation.*

2.2. Brette–Gerstner exponential adaptative integrate-and-fire neuron. In this section we study the bifurcation diagram of the adaptive exponential neuron. This model has been introduced by Brette and Gerstner in [5]. This model, inspired by the Izhikevich adaptive quadratic model, can be fitted to biological values, takes into account the adaptation phenomenon, and is able to reproduce many behaviors observed in cortical neurons. The bifurcation analysis we derived in section 1 allows us to understand how the parameters of the model can affect the behavior of this neuron. We show that this model is part of the general class studied in section 1, and we obtain the fixed-point bifurcation diagram of the model.

2.2.1. Reduction of the original model. This original model is based on biological constants and is expressed with a lot of parameters. We first reduce this model to a simpler form with the fewest number of parameters.

The basic equations proposed in the original paper [5] read

$$\begin{cases} C\frac{\mathrm{d}V}{\mathrm{d}t} = -g_L(V-E_L) + g_L\Delta_T \exp\left(\frac{V-V_T}{\Delta_T}\right) \\ \qquad\qquad -g_e(t)(V-E_e) - g_i(t)(V-E_i) - W + I_m, \\ \tau_W\frac{\mathrm{d}W}{\mathrm{d}t} = \kappa(V-E_L) - W. \end{cases} \tag{2.2}$$

First, we do not assume that the reversal potential of the w equation is the same as the leakage potential E_L, and we write the equation for the adaptation variable by

$$\tau_W\frac{\mathrm{d}W}{\mathrm{d}t} = a(V-\bar{V}) - W.$$

Next we assume that $g_e(\cdot)$ and $g_i(\cdot)$ are constant (in the original paper it was assumed that the two conductances were null).

After some straightforward algebra, we eventually get the following dimensionless equation equivalent to (2.2):

$$\begin{cases} \dot{v} = -v + e^v - w + I, \\ \dot{w} = a(bv - w), \end{cases} \tag{2.3}$$

where we denoted

$$\begin{cases} \tilde{g} := g_L + g_e + g_i, \\ \tau_m := \frac{C}{\tilde{g}}, \\ B := \frac{\kappa}{\tilde{g}} \left(\frac{E_L}{\Delta_T} + \log(\frac{g_L}{\tilde{g}} e^{-V_T/\Delta_T}) \right), \\ v(\tau) := \frac{V(\tau\tau_m)}{\Delta_T} + \log\left(\frac{g_L}{\tilde{g}} e^{-V_T/\Delta_T} \right), \\ w(\tau) := \frac{W(\tau\tau_m)}{\tilde{g}\Delta_T} + B, \\ a := \frac{\tau_m}{\tau_W}, \\ b := \frac{\kappa}{\tilde{g}}, \\ I := \frac{I_m + g_L E_L + g_e E_e + g_i E_i}{\tilde{g}\Delta_T} + \log(\frac{g_l}{\tilde{g}} e^{-V_T/\Delta_T}) + B \end{cases} \tag{2.4}$$

and where the dot denotes the derivative with respect to τ.

Remark. *These expressions confirm the qualitative interpretation of the parameters a, b, and I of the model* (1.1). *Indeed,* $a = \frac{\tau_m}{\tau_w}$ *accounts for the time scale of the adaptation (with the membrane time scale as reference), and the parameter* $b = \frac{\kappa}{\tilde{g}}$ *is proportional to the interaction between the membrane potential and the adaptation variable and inversely proportional to the total conductivity of the membrane potential. Eventually, I is an affine function of the input current* I_m *and models the input current of the neurons.*

2.2.2. Bifurcation diagram. From (2.3) we can clearly see that the Brette–Gerstner model is included in the formal class studied in the paper with

$$F(v) = e^v - v.$$

This function satisfies Assumptions (A1), (A2), and (A3). Furthermore, its second order derivative never vanishes.

Theorem 1.9 shows that the system undergoes the following bifurcations:

(BG.B1) A saddle-node bifurcation curve defined by

$$\{(b, I) \; ; \; I = (1+b)(1 - \log(1+b))\}.$$

So $v^*(b) = \log(1+b)$. For $I \leq (1+b)(1-\log(1+b))$, the system has the fixed points

$$\begin{cases} v_-(I,b) := -W_0\left(-\frac{1}{1+b} e^{\frac{I}{1+b}}\right) + \frac{I}{1+b}, \\ v_+(I,b) := -W_{-1}\left(-\frac{1}{1+b} e^{\frac{I}{1+b}}\right) + \frac{I}{1+b}, \end{cases} \tag{2.5}$$

where W_0 is the principal branch of Lambert's W function[3] and W_{-1} the real branch of Lambert's W function such that $W_{-1}(x) \leq -1$, defined for $-e^{-1} \leq x < 1$.

[3] The Lambert W function is the inverse function of $x \mapsto xe^x$.

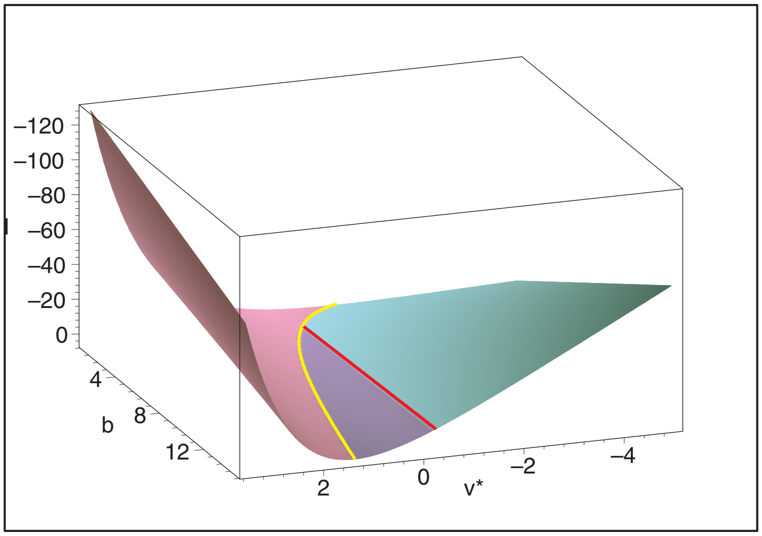


FIG. 2.2. *Representation of the v fixed point of the Brette–Gerstner model with respect to the parameters I and b. The reddish/pinkish component is the surface of saddle fixed points, the purplish one corresponds to the repulsive fixed points, and the bluish/greenish one corresponds to the attractive fixed points The yellow curve corresponds to a saddle-node bifurcation and the red one to an Andronov–Hopf bifurcation.*

(BG.B2) An Andronov–Hopf bifurcation line for

$$\{(b, I)\ ;\ b > a \text{ and } I = I^*(a, b) = (1 + b)\log(1 + a) - (1 + a)\}$$

at the equilibrium point ($v_a = \log(1+a)$, $w_a = bv_a$). This type of Andronov–Hopf bifurcation is given by the sign of the variable

$$A(a, b) = F'''(v_a) + \frac{1}{b - a}F''(v_a)^2 = (1 + a) + \frac{4}{b - a}(1 + a)^2 > 0.$$

So the bifurcation is always subcritical, and there is not any Bautin bifurcation.

(BG.B3) A Bogdanov–Takens bifurcation point at the point $b = a$ and $I = \log(1 + a)$.

(BG.B4) A saddle homoclinic bifurcation curve satisfying, near the Bogdanov–Takens point, the equation

$$(P) := \Bigg\{ (I = (1 + a)(\log(1 + a) - 1) + I_1,\ b = a + b_1)\ ;$$

$$I_1 = \frac{\left(-\frac{25}{6}\,a - \frac{37}{6}\,b_1 + \frac{5}{6}\,\sqrt{25\,a^2 + 74\,b_1\,a + 49\,{b_1}^2}\right) a}{(1 + a)} + o(|\,b_1\,| + |\,I_1\,|)$$

$$\text{and } b_1 > -\left(1 + \frac{1}{a}\right) I_1 \Bigg\}.$$

In Figure 2.2 we represent the fixed points of the exponential model and their stability, together with the bifurcation curves, in the space (I, b, v).

3. The richer quartic model.

In this section, we introduce a new specific model having a richer bifurcation diagram than the two models studied in section 2. It is as simple as the two previous models from the mathematical and computational points of view. To this end, we define a model which is part of the class studied in section 1 by specifying the function F.

3.1. The quartic model: Definition and bifurcation map.

Let $a > 0$ be a fixed real and $\alpha > a$. We instantiate the model (1.1) with the function F a quartic polynomial:

$$F(v) = v^4 + 2av.$$

REMARK. *The choice of the function F here is just an example where all the formulas are rather simple. Exactly the same analysis can be done with any F function satisfying $F'''(v_a) < 0$ and the transversality conditions given in Theorem* 1.10. *This would be the case, for instance, for any quartic polynomial $F(v) = v^4 + \alpha v$ for $\alpha > a$.*

The function F satisfies Assumptions (A1), (A2), and (A5). $F'(v) = 4v^3 + 2a$ satisfies Assumption (A3).

Nevertheless, we have to bear in mind that the second order derivative vanishes at $v = 0$:

$$\begin{cases} \dot{v} = v^4 + 2av - w + I, \\ \dot{w} = a(bv - w). \end{cases} \tag{3.1}$$

Theorem 1.9 shows that the quartic model undergoes the following bifurcations:

(B1) A saddle-node bifurcation curve defined by

$$(SN) := \left\{ (b, I) \ ; \ I = 3\left(\frac{b-2a}{4}\right)^{(4/3)} \right\}.$$

Proof. Indeed, the function G reads $G(v) = v^4 + (2a - b)v$ and reaches its minimum at the point $v = (\frac{b-2a}{4})^{(1/3)}$. So the minimum of G is $m(b) = -3\,(\frac{b-2a}{4})^{(4/3)}$. □

The point $v^*(b)$ is $(\frac{b-2a}{4})^{(1/3)}$, and we have closed-form expressions (but rather complicated) for the two fixed points for $I < 3\,(\frac{b-2a}{4})^{(4/3)}$ since the quartic equation is solvable in radicals. The closed form expression can be obtained using a symbolic computation package like Maple using the command

```
S:=allvalues( solve( x^4 + (2*a - b) * x + I0 = 0,x));
```

(B2) An Andronov–Hopf bifurcation curve for $b > a$ along the straight line

$$(AH) := \left\{ (I, b) \ ; \ b > a \text{ and } I = -\left(\frac{a}{4}\right)^{1/3} b - \left(\frac{a}{4}\right)^{4/3} \right\}.$$

The fixed point where the system undergoes this bifurcation is $v_a = -(\frac{a}{4})^{1/3}$. The kind of Andronov–Hopf bifurcation we have is governed by the sign of

$$\alpha = -24\left(\frac{a}{4}\right)^{1/3} + \frac{144}{b-a}\left(\frac{a}{4}\right)^{4/3}.$$

Finally, the type of bifurcation changes when b varies.

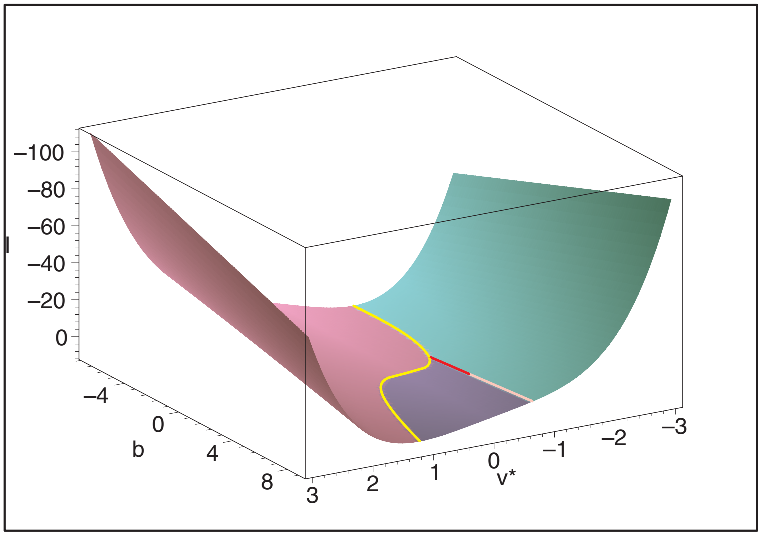


FIG. 3.1. *v fixed points and their stability in function of I and b. The reddish/pinkish component is the surface of saddle fixed points, the purplish one corresponds to the repulsive fixed points, and the bluish/greenish one corresponds to the attractive fixed points. The yellow curve corresponds to a saddle-node bifurcation, the red curve to a subcritical Andronov–Hopf bifurcation, and the greyish one to the supercritical Andronov–Hopf bifurcation. The intersection point between the yellow and the red curve is the Bogdanov–Takens bifurcation point, and the intersection point of the red and greyish curves is the Bautin bifurcation point.*

- When $b < \frac{5}{2}a$, then $\alpha > 0$, hence $l_1 > 0$, and the Andronov–Hopf bifurcation is subcritical.
- When $b > \frac{5}{2}a$, then $\alpha < 0$, hence $l_1 < 0$, and the Andronov–Hopf bifurcation is supercritical.

We prove below that the change in the type of Hopf bifurcation is obtained via a Bautin bifurcation.

(B3) A Bogdanov–Takens bifurcation point is located at $b = a$ and $I = -3\,(\frac{a}{4})^{(4/3)}$.

(B4) A saddle homoclinic bifurcation curve satisfying, near the Bogdanov–Takens point, the equation

$$\begin{aligned}(P) := \Bigg\{ &\left(I = -3\left(\frac{a}{4}\right)^{(4/3)} + I_1,\, b = a + b_1\right)\ ;\\ &I_1 = \frac{1}{12}\left(-\frac{25}{6}\,a - \frac{37}{6}\,b_1 + \frac{5}{6}\sqrt{25\,a^2 + 74\,b_1\,a + 49\,{b_1}^2}\right)a^{1/3}\\ &+ o(|\,b_1\,| + |\,I_1\,|)\\ &\text{and } b_1 > -6 I_1 a^{-1/3} \Bigg\}.\end{aligned}$$

(B5) A Bautin bifurcation at the point $\big(b = \frac{5}{2}a,\ I = -3(\frac{a}{4})^{4/3}\,(2\,a - 1)\big)$ and a saddle node bifurcation of periodic orbits coming along (see section 3.2).

Figure 3.1 represents the bifurcation curves and the fixed point of the quartic model in the space (I, b, v).

3.2. The Bautin bifurcation. As we have seen in the last section, at the point

$$\begin{cases} v_a = -\left(\frac{a}{4}\right)^{1/3}, \\ I = -3\left(\frac{a}{4}\right)^{4/3}(2\,a-1), \\ b = \frac{5}{2}a \end{cases} \tag{3.2}$$

the Jacobian matrix of the system has a pair of purely conjugate imaginary eigenvalues and a vanishing first Lyapunov exponent.

To prove the existence of a Bautin bifurcation, we start our computations from the point of section 1.3.4. In this case the calculations can be led until the end, but the expressions are very intricate, and we do not reproduce it here. In Appendix A we show the calculations to perform. We prove that the system actually undergoes a Bautin bifurcation except for two particular values of the parameter a.[4]

With this method we obtain a closed-form expression for the second Lyapunov exponent. We show that this second Lyapunov exponent vanishes for two values of a, whose expressions are complicated. These calculations are rigorous, but nevertheless, the interested reader can find numerical expressions of this exponent to get a grasp on its behavior in the appendix (see (A.7)) and of the two numerical values of a such that $l_2(a)$ vanishes.

Things are even more involved when we are interested in the regularity of the map $(I, b) \mapsto (\mu(I, b), l_1(I, b))$. Nevertheless, we obtain that this determinant never vanishes.

Eventually, for all a different from the critical values where the second Lyapunov exponent vanishes, the system undergoes a Bautin bifurcation.

Note finally that the Bautin bifurcation point separates two branches of sub- and supercritical Hopf bifurcations. For nearby parameter values, the system has two coexisting limit cycles, an attractive one and a repelling one, which collide and disappear via a saddle-node bifurcation of periodic orbits.

4. Numerical simulations.

In the previous sections we emphasized the fact that the class of models we defined in section 1 was able to reproduce the behaviors observed by Izhikevich in [15]. In this section, first we show that the quartic model indeed reproduces the behaviors observed by Izhikevich and which correspond to cortical neuron behaviors observed experimentally. We also produce some simulations of self-sustained subthreshold oscillations which occur only when the dynamical system has attracting periodic orbits, which is not the case in the IBG models.

Izhikevich in [15] explains the main features we obtain in numerical simulations from the neurocomputational point of view. In this paper, we comment on these same features from the dynamical systems point of view. This analysis also gives us a systematic way of finding the parameters associated with one of the possible behaviors.

4.1. Simulation results. We now provide simulation results of the quartic model introduced in section 3. In the simulated model, the spike is not represented by the blow up of the potential membrane v, but we consider that the neuron emits a spike when its membrane potential crosses a constant threshold.[5]

[4] All the computations have been performed using Maple, but the expressions are very involved and are not reproduced here.

[5] Note that the numerical simulations are very robust with respect to the choice of the threshold, if taken large enough, since the underlying equation blows up in finite time.

Let θ be our threshold. The simulated model considered in this section is the solution of the equations

$$\begin{cases} \dot{v} = v^4 + 2av - w + I, \\ \dot{w} = a(bv - w) \end{cases} \tag{4.1}$$

together with the spike-and-reset condition

$$\text{If } v(t^-) > \theta \Rightarrow \begin{cases} v(t) = v_r, \\ w(t) = w(t^-) + d. \end{cases} \tag{4.2}$$

Simulations have been done using an Euler numerical scheme, with a time step ranging from 10^{-1} to 10^{-2} depending on the precision needed, and with time intervals ranging from 10 to 500. This method is very efficient numerically and remains precise. Other integration methods could be used, and the qualitative results we obtained do not depend on the integration scheme, as soon as the time step is small enough.

Remark (on Figure 4.1). *Note that we did not reproduce the last three behaviors presented by Izhikevich in* [15, *Figs.* 1.(R), 1.(S), *and* 1.(T)]. *Indeed, these behaviors are not in the scope of the present paper and do not correspond to the model we studied.*

More precisely, in the study of the general model (1.1), *we considered for phenomenological reasons $a > 0$, modelling the leak of the adaptation variable: the adaptation would converge to its rest value if it was not influenced by the membrane potential v. If we considered $a < 0$, this adaptation variable would diverge exponentially from this rest value if it was not controlled by the membrane potential v. The inhibition-induced behaviors* [15, *Figs.* 1.(S) *and* 1.(T)] *require a to be strictly negative, and so we will not comment on these behaviors any further.*

Similarly, the accommodation behavior presented by Izhikevich in [15, *Fig.* 1.(R)] *is a limit case when w is very slow and the adaptation efficiency b very high. Mathematically speaking, it corresponds to a case where $a \to 0$ and $ab \to \lambda \neq 0$. This case is not taken into account in our study and amounts to replacing* (1.1) *by an equation of the type*

$$\begin{cases} \frac{dv}{dt} = F(v) - w + I, \\ \frac{dw}{dt} = ab(v - v_0), \end{cases} \tag{4.3}$$

and the study of this equation is not in the scope of the present paper.

The simulated behaviors we obtained in Figure 4.1 have been obtained playing with the bifurcation parameters in the phase plane. The way the parameters were set was based on a qualitative reasoning on the phase plane and the bifurcation diagram in a way we now describe.

4.2. Bifurcations and neuronal dynamics. In this section we link the neuronal behaviors shown in Figure 4.1 with the bifurcations of the system.

- (i) *Tonic spiking*: This behavior corresponds to the saddle-node bifurcation. The system starts from a (stable) equilibrium point near the saddle-node bifurcation curve (see Figure 4.2). Then we apply a greater constant current I, and the new dynamical system has no fixed point (we "cross" the saddle-node bifurcation curve). So the neuron begins spiking. The stabilization of the spiking frequency is linked with the existence of what we will call a *limit spiking cycle*. Indeed, we can see that the phase plane trajectory

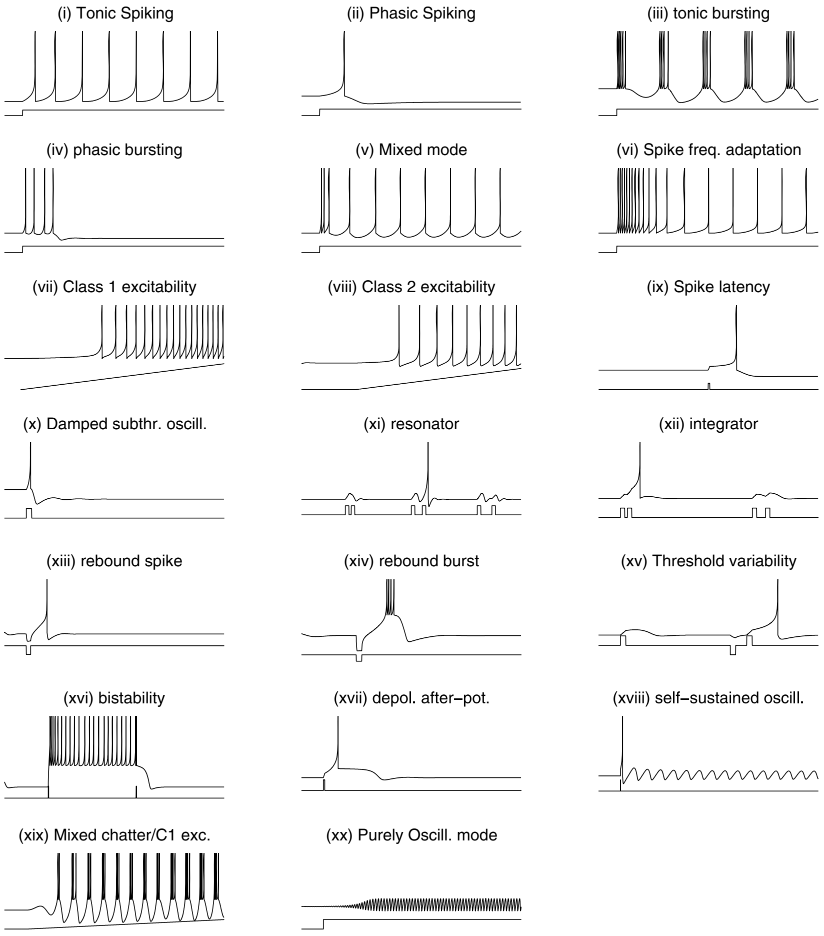


FIG. 4.1. *Different remarkable neurocomputational interesting behaviors of the neuron model* (4.1) *with the reset condition* (4.2) *for different choices of the parameters* (a, b, I, v_r, d). *The higher curve represents the membrane potential* v *and the lower one the input current* I *(see Appendix* B *for the numerical values of each simulations).*

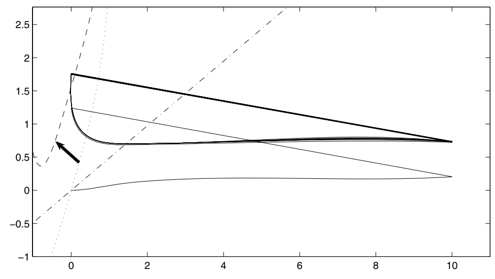
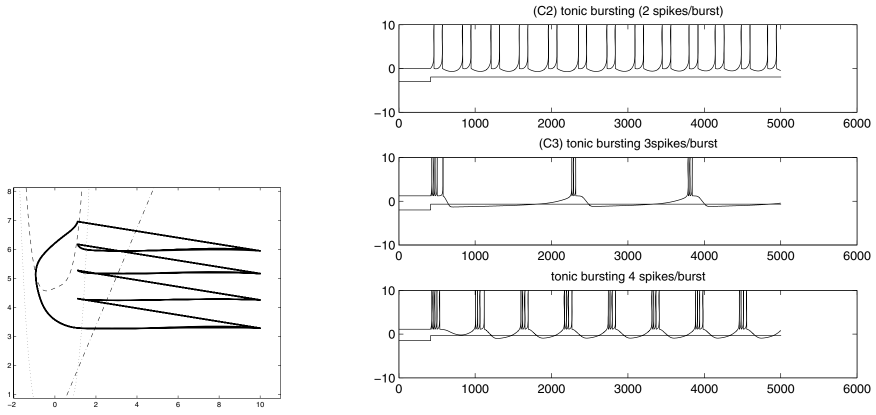


FIG. 4.2. *Tonic spiking: phase plane trajectory. The dotted curve is the* v *nullcline at the initial time. It is shifted to the dashed one when applying a constant input current. The new dynamical system has no fixed point and spikes regularly. We can see the* spiking cycle *appearing.*



(a) Phase plane of the tonic spiking (without the transient phase)

(b) Controlling the number of spikes per burst

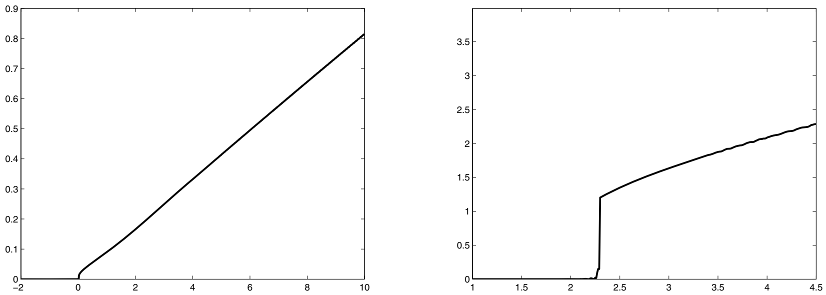
FIG. 4.3. *Tonic bursting: phase plane trajectory. The dotted curve is the v nullcline at the initial time. It is shifted to the dashed one when applying a constant input current. The new dynamical system has no fixed point. We can see the* multiple spike limit cycle *here.*

converges to a kind of cycle. This cycle includes a spike point ($v = \infty$, or $v =$ threshold in the numerical case), and so it is not a classical limit cycle. The v is always reset to the same value, and we can see that the adaptation variable w converges to an attracting stable value w_{spike}. This value satisfies $w_s(t_{\text{spike}}) + b = w_{\text{spike}}$, where $w_s(\cdot)$ is solution of (4.1) with the initial conditions

$$\begin{cases} v(0) = v_r, \\ w(0) = w_{\text{spike}} \end{cases}$$

and where t_{spike} denotes the time of the spike.

- (ii) *Phasic spiking*: This behavior occurs on the stable fixed point portion of the phase plane. The system starts at a fixed point. Then we apply a constant current to the neuron greater than the initial current but lower than the current associated with the saddle-node bifurcation. This stimulation forces the neuron to spike. Nevertheless, the reset point falls in the attraction basin of the new fixed point, and the trajectory converges to this point.
- (iii) *Tonic bursting*: This behavior is also linked to the saddle-node bifurcation. The system starts at a (stable) fixed point, and when we apply a constant current, we cross this bifurcation. The new dynamical system has no fixed point and is in a spiking behavior. The only difference with the tonic spiking behavior is that the point (v_r, w_{spike}) is in the zone $\{(v, w);\ \dot{v} < 0\}$. So the system emits quickly a precise number of spikes and then crosses the v nullcline. At this point, the membrane potential decays before spiking. We can see numerically that the system converges to a stable *spiking cycle* (see Figure 4.3(a)) containing a given number of spikes, a decay, and then the same sequence of spikes again. So the two-dimensional system is able to reproduce the diagrams presented by Izhikevich in [13] in an (at least) three-dimensional space. This is possible in two dimensions because of the



(a) Class 1 excitability

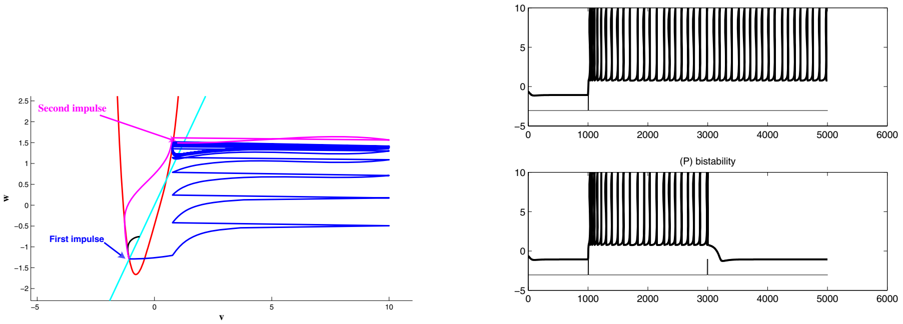
(b) Class 2 excitability

Fig. 4.4. *Spiking frequency vs. input current I for different choices of b. These curves have been obtained running simulations for different values of the input current, computing the frequency of the emitted spikes in a time range $T = 10000$.*

singularity of the model (explosion or threshold/reinitialization). If the system was regular, this behavior would not have been possible because it would have contradicted the Cauchy–Lipschitz theorem of existence and uniqueness of a solution.

Note that we can choose exactly the number of spikes per burst by changing the adaptation parameter d and that the bursting can be of parabolic or square-wave type as defined in Hoppensteadt and Izhikevich [12] (see Figure 4.3(b)).

- (iv) *Phasic bursting*: This behavior is linked with what we discussed in (ii) and (iii): the system starts at a stable fixed point. When the input current is turned on, the nullcline is shifted and the initial point is now in the spiking zone, and so a spike is emitted. Nevertheless, in contrast with (ii), the reset does not fall in the attraction basin of the new stable fixed point, but the point (v_0, w_{spike}) is inside this attraction basin. So a certain number of spikes are emitted before returning to the new fixed point. Here again we are able to control the number of spikes in the initial burst.
- (v) *Mixed mode*: The dynamical system interpretation is mixed between the phasic bursting and the tonic spiking. A certain number of spikes are necessary to converge to the spiking cycle.
- (vi) *Spike frequency adaptation*: This behavior is a particular case of tonic bursting where the convergence to the stable spiking cycle is slow.
- (vii)/(viii) *Class one/two excitability*: Figure 4.4(a) and (b) represents the spiking frequency of the neurons as a function of the input current. We can see that for the first choice of parameter, the frequency can be very small and increases regularly, and for the second choice of parameter, we can see that the system cannot spike in a given range of frequency (this frequency cannot be lower than 1.2Hz). Those simulations show that, depending on the chosen parameters, the system can be class 1 or class 2 excitable.
- (ix)/(xvii) *Spike latency/DAP*: It is a particular case of phasic spiking when the equilibrium v^* or the reset point v_r is near a point such that $F(v) = F'(v) = 0$. The membrane potential dynamics is very slow around this point. In the spike latency behavior, the initial point is close to this point, which generates the observed latency. In our case, it is around the minimum of



(a) Bistability: return to equilibrium via the same impulse

(b) Bistability

FIG. 4.5. *Bistability phenomenon: The first impulse induces a self-sustained tonic spiking behavior while the system has a stable fixed point. The second impulse perturbs this regular spiking behavior, and the system falls in the attraction basin of the stable fixed point.*

the function F (see Figure 4.6(ix)). In the depolarized after-potential (DAP) case, the reset occurs near this point, which is also in the attraction basin of the stable fixed point.

- (x) *Damped subthreshold oscillations*: This behavior occurs in the neighborhood of the stable fixed point: the stimulation evokes a spike, and the reset falls in the attraction basin of the stable fixed point, which has complex eigenvalues with negative real parts. This generates damped subthreshold oscillations.
- (xi) *Resonator*: This behavior occurs at the stable fixed point when the Jacobian matrix has complex eigenvalues. The first spike induces damped subthreshold oscillations. The spike is emitted if the second spike is given at the period of those oscillations, which is given by the argument of the complex eigenvalue. If it occurs before or after, then no spike is emitted.
- (xii) *Integrator*: This behavior occurs when we stimulate the system from the stable fixed point when the Jacobian matrix has real (negative) eigenvalues. If the first stimulation is not sufficient to make the neuron spike, then the stimulation is damped. Nevertheless, the membrane potential returns to equilibrium slowly, and if the same stimulation arrives to the "destabilized" neuron, it can generate a spike. The closer the second stimulation is to the first one, the more probable the omission of the spike.
- (xiii)/(xiv) *Rebound spike or burst*: The input impulse makes the neuron spike, and the reset (or the second, third, nth reset) falls in the attraction basin of the stable fixed point.
- (xv) *Threshold variability*: This phenomenon is exactly the same as the integrator, but instead of destabilizing the variable v we play on the adaptation variable.
- (xvi) *Bistability*: This behavior starts from the stable fixed point. The *attracting reset* (v_r, w_{spike}) is outside the attraction basin of the fixed point but still close to this zone. The first impulse generates a spike and initiates a tonic spiking mode. Nevertheless, it is possible via a small perturbation of the trajectory to fall into the attraction basin of the fixed point (see Figure 4.5).

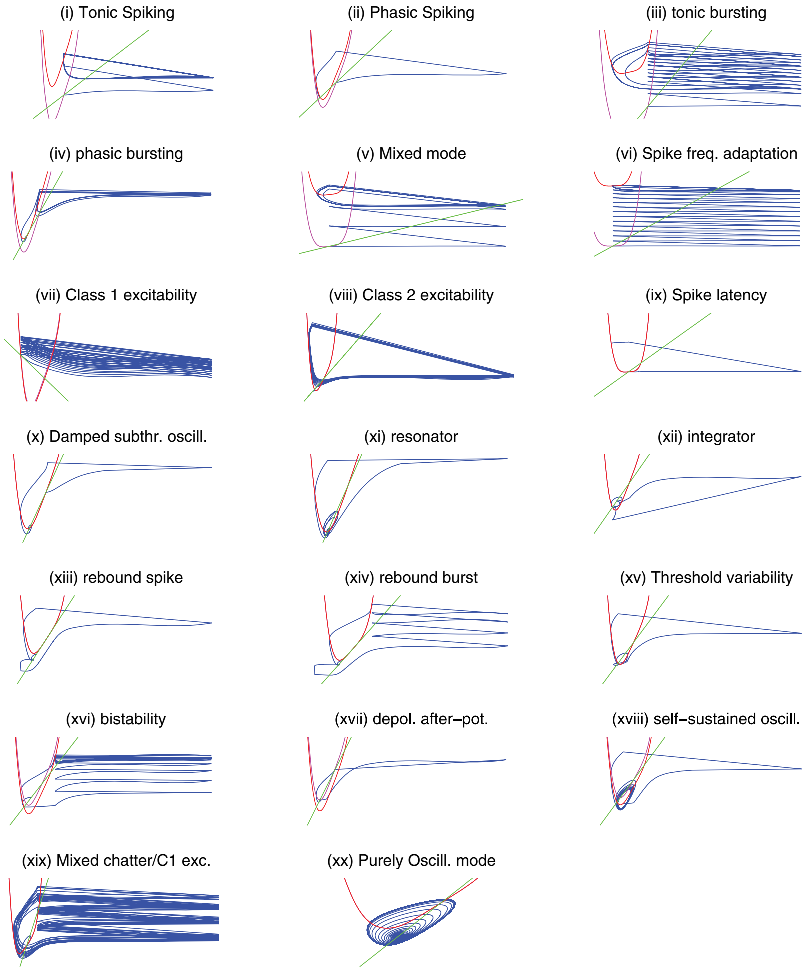
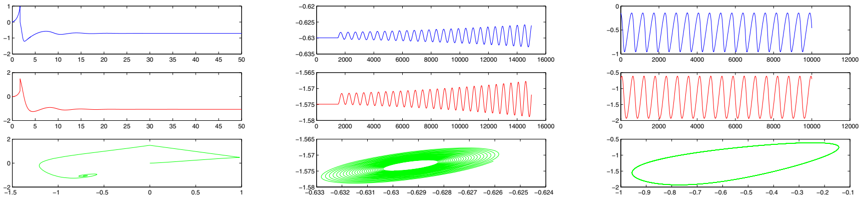


FIG. 4.6. *Phase diagrams corresponding to the behaviors presented in Figure* 4.1.

- (xviii)/(xx) *Self-sustained subthreshold oscillations and purely oscillating mode*: They are linked with the supercritical Hopf bifurcation and its stable periodic orbit. These two behaviors cannot be obtained in the IBG models since the Hopf bifurcations are always subcritical.

4.3. Self-sustained subthreshold oscillations in cortical neurons. In this study we gave a set of sufficient conditions to obtain an IBG-like model of neuron. In this framework we proposed a model that displays a Bautin bifurcation the IBG neurons lack; as a consequence our model can produce subthreshold oscillations. In this section, we explain from a biological point of view the origin and the role of those oscillations and reproduce in vivo recordings.

In the IBG models, the Andronov–Hopf bifurcation is always subcritical. The only oscillations created in these models are damped (see Figure 4.7(a)) and correspond in the phase plane to the convergence to a fixed point where the Jacobian matrix has complex eigenvalues. Our quartic model undergoes supercritical Andronov–Hopf bifurcations, and so there are attracting periodic solutions. This means that the



(a) Damped oscillations

(b) Transient phase towards the self-sustained oscillations

(c) Self-sustained oscillations (stationary state)

FIG. 4.7. *The quartic model shows damped subthreshold oscillations like the IBG models (Figure* 4.7(a)*): the trajectory collapses to a fixed point (parameters: $a = 1$, $b = 1.5$, $I = 0.1$, $T_{max} = 100$, $dt = 0.01$). The upper (blue) curve represents the solution in v, the middle (red) one w, and the lower one (green) the trajectory in the plane (v, w). Self-sustained subthreshold oscillations of the quartic model (Figures* 4.7(b) *and* 4.7(c)*): the trajectory is attracted towards a limit cycle (parameters: $a = 1$, $b = 5/2$, $I = -3(a/4)^{4/3}(2a-1)$, $T_{max} = 150000$, $dt = 0.01$, $I = (-3(a/4)^{4/3}(2a-1)+0.001)$.*

neurons can show self-sustained subthreshold oscillations (Figures 4.7(b) and 4.7(c)), which is of particular importance in neuroscience.

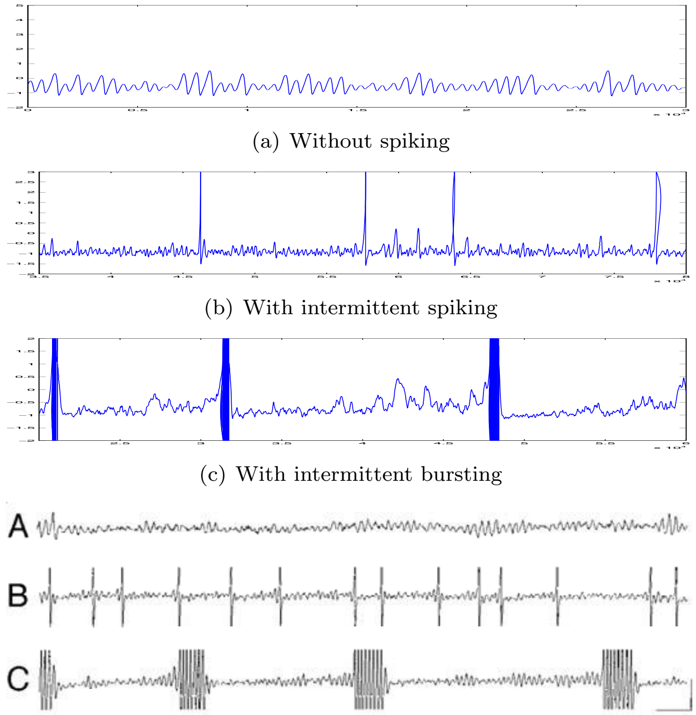
Most biological neurons show a sharp transition from silence to a spiking behavior, which is reproduced in all the models of class (1.1). However, experimental studies suggest that some neurons may experience a regime of small oscillations [21]. These subthreshold oscillations can facilitate the generation of spike oscillations when the membrane gets depolarized or hyperpolarized [22, 23]. They also play an important role in shaping specific forms of rhythmic activity that are vulnerable to the noise in the network dynamics.

For instance, the inferior olive nucleus, a part of the brain that sends sensory information to the cerebellum, is composed of neurons able to support oscillations around the rest potential. It has been shown by Llinás and Yarom [22, 23] that the precision and robustness of these oscillations are important for the precision and the robustness of spike generation patterns. The quartic model is able to reproduce the main features of the inferior olive neuron dynamics:

i. autonomous subthreshold periodic and regular oscillations (see intracellular recordings of inferior olive neurons in brain stem slices in [23]),
ii. rhythmic generation of action potentials.

The robust subthreshold oscillations shown by in vivo recordings [4, 21, 23] correspond in our quartic model to the stable limit cycle coming from the supercritical Hopf bifurcation. The oscillations generated by this cycle are stable, and they have a definite amplitude and frequency. This oscillation occurs at the same time as the rhythmic spike generation in the presence of noisy or varying input. Note that other neuron models such as those studied above, even if they do not undergo a supercritical Hopf bifurcation, can also exhibit oscillations in the presence of noise, for instance near a subcritical Hopf bifurcation. Nevertheless, these oscillations do not have the regularity in the amplitude and the frequency linked with the presence of an attracting limit cycle. The results we obtain simulating the quartic model are very similar to those obtained by in vivo recordings (see Figure 4.8).

But the inferior olive neurons are not the only neurons to present subthreshold membrane potential oscillations. For instance, stellate cells in the enthorhinal cortex demonstrate theta frequency subthreshold oscillations [1, 2, 17], linked with the persistent Na$^+$ current I_{NaP}.



(a) Without spiking

(b) With intermittent spiking

(c) With intermittent bursting

(d) Biological recordings

FIG. 4.8. *Subthreshold membrane oscillations, qualitatively reproducing the recordings from* [20] *in DRG neurons. Traces illustrate (*4.8(a)*) oscillations without spiking, (*4.8(b)*) oscillations with intermittent spiking, and (*4.8(c)*) oscillations with intermittent bursting (in the figures, spikes are truncated). The noisy input is an Ornstein–Ulhenbeck process. The biological recordings* 4.8(d) *are reproduced from* [20, *Fig.* 1] *and used with permission.*

We now conclude this section on the specific example of subthreshold self-sustained oscillations given by the dorsal root ganglia (DRG) neuron. This neuron presents subthreshold membrane potential oscillations coupled with repetitive spike discharge or burst, for instance in the case of a nerve injury [20, 3]. Figure 4.8(d) shows biological in vivo intracellular recordings performed by Liu et al. [20] from a DRG neuron of an adult male rat. The recorded membrane potentials exhibit high frequency subthreshold oscillation in the presence of noise, combined with a repetitive spiking or bursting. These behaviors can be reproduced by the quartic model, as we can see in Figure 4.8, around a point where the system undergoes a supercritical Hopf bifurcation.[6]

Conclusion. In this paper we defined a general class of neuron models able to reproduce a wide range of neuronal behaviors observed in experiments on cortical neurons. This class includes the Izhikevich and the Brette–Gerstner models, which are widely used. We derived the bifurcation diagram of the neurons of this class and proved that they all undergo the same types of bifurcations: a saddle-node bifurcation curve, an Andronov–Hopf bifurcation curve, and a codimension two Bogdanov–Takens

[6] The amplitude and frequency of the subthreshold oscillations can be controlled choosing a point on the supercritical Hopf bifurcation curve.

bifurcation. We proved that there was only one other possible fixed-point bifurcation, a Bautin bifurcation. Then using those theoretical results we proved that the Izhikevich and the Brette–Gerstner models had the same bifurcation diagram.

This theoretical study allows us to search for interesting models in this class of neurons. Indeed, Theorem 1.9 ensures us that the bifurcation diagram will present at least the bifurcations stated. This information is of great interest if we want to control the subthreshold behavior of the neuron of interest.

Following these ideas, we introduced a new neuron model of our global class undergoing the Bautin bifurcation. This model, called the *quartic model*, is computationally and mathematically as simple as the IBG models and able to reproduce some cortical neuron behaviors which the IBG models cannot reproduce.

This study focused on the subthreshold properties of this class of neurons. The adaptative reset of the model is of great interest and is a key parameter in the repetitive spiking properties of the neuron. Its mathematical study is very rich and is still an ongoing work.

Appendix A. Bautin bifurcation. In this appendix we prove that the quartic model undergoes a Bautin bifurcation at the point

$$\text{(A.1)} \qquad \begin{cases} b = \frac{5}{2}\,a, \\ I = -3\left(\frac{a}{4}\right)^{4/3}(2\,a-1)\,, \\ v_a = -\left(\frac{a}{4}\right)^{1/3}. \end{cases}$$

A.1. The first Lyapunov exponent. Indeed, using a suitable affine change of coordinates, the system at this point reads

$$\text{(A.2)} \qquad \begin{cases} \dot{x} = \omega y, \\ \dot{y} = \frac{ab}{\omega}\left(6v_a^2 v_1(x,y)^2 + 4v_a v_1(x,y)^3 + v_1(x,y)^4\right) \\ \quad = \frac{1}{2}F_2\left(\binom{x}{y},\binom{x}{y}\right) + \frac{1}{6}F_3\left(\binom{x}{y},\binom{x}{y},\binom{x}{y}\right) + \frac{1}{24}F_3\left(\binom{x}{y},\binom{x}{y},\binom{x}{y},\binom{x}{y}\right), \end{cases}$$

where $v_1(x,y) = \frac{1}{b}x + \frac{\omega}{ab}y$. We also denote $F_2(X,Y)$, $F_3(X,Y,Z)$, and $F_4(X,Y,Z,T)$ the multilinear symmetric vector functions of (A.2) ($X,Y,Z,T \in \mathbb{R}^2$):

$$\begin{cases} F_2\left(\binom{x}{y},\binom{z}{t}\right) = \binom{0}{12\frac{ab}{\omega}v_a^2 v_1(x,y) v_1(z,t)}, \\ \dots \end{cases}$$

To compute the two first Lyapunov exponents of the system, we follow Kuznetsov's method [19]. In this method we need to compute some specific right and left complex eigenvectors, which can be chosen in our case to be

$$\text{(A.3)} \qquad \begin{cases} p = \begin{pmatrix} \frac{1}{-i\sqrt{a\,b-a^2}+a} \\ 1 \end{pmatrix}, \\ q = \begin{pmatrix} \frac{1}{2}\,\frac{(i\sqrt{a(b-a)}+a)b}{b-a-i\sqrt{a(b-a)}} \\ 1/2\,\frac{(i\sqrt{a(b-a)}+a)^2}{a\,(b-a-i\sqrt{a(b-a)})} \end{pmatrix}. \end{cases}$$

We now put the system in a complex form letting $z = x + iy$.

We can now compute the complex Taylor coefficients g_{ij}:

$$\tag{A.4} \begin{cases} g_{20} = \langle p, F_2(q,q)\rangle, \\ g_{11} = \langle p, F_2(q,\bar{q})\rangle, \\ g_{02} = \langle p, F_2(\bar{q},\bar{q})\rangle, \\ \\ g_{30} = \langle p, F_3(q,q,q)\rangle, \\ g_{21} = \langle p, F_3(q,q,\bar{q})\rangle, \\ g_{12} = \langle p, F_3(\bar{q},\bar{q},\bar{q})\rangle, \\ g_{03} = \langle p, F_3(\bar{q},\bar{q},\bar{q})\rangle, \\ \dots \end{cases}$$

So the Taylor coefficients (A.4) read

$$\tag{A.5} \begin{cases} g_{20} = 12\frac{ab}{\omega} v_a^2 v_1 \left(\frac{1}{2}\frac{(i\sqrt{a(b-a)}+a)b}{b-a-i\sqrt{a(b-a)}}, \frac{1}{2}\frac{(i\sqrt{a(b-a)}+a)^2}{a\,(b-a-i\sqrt{a(b-a)})} \right)^2, \\ g_{11} = 12\frac{ab}{\omega} v_a^2 v_1(q) v_1(\bar{q}), \\ g_{02} = 12\frac{ab}{\omega} v_a^2 v_1(\bar{q}) v_1(\bar{q}), \\ \\ \dots \end{cases}$$

Now let $S(I,b) := F'(v_-(I,b))$ be the value of the derivative of the function F, defined around the bifurcation point we are interested in.

The Jacobian matrix in the neighborhood of the point (A.1) reads

$$L(v) = \begin{pmatrix} S(I,b) & 1 \\ ab & -a \end{pmatrix}.$$

Let us denote $\alpha = \binom{I}{b}$ the parameter vector and $\lambda(\alpha) = \mu(\alpha) \pm i\omega(\alpha)$ the eigenvalues of the Jacobian matrix. We have

$$\begin{cases} \mu(\alpha) = \frac{1}{2}\left(S(\alpha) - a\right), \\ \omega(\alpha) = \frac{1}{2}\sqrt{-(S(\alpha)-a)^2 + 4ab}. \end{cases}$$

With these notations, let $c_1(\alpha)$ be the complex defined by

$$c_1(\alpha) = \frac{g_{20}g_{11}(2\lambda+\bar{\lambda})}{2|\lambda|^2} + \frac{|g_{11}|^2}{\lambda} + \frac{|g_{02}|^2}{2(2\lambda-\bar{\lambda})} + \frac{g_{21}}{2}$$

(in this formula we omit the dependence in α of λ for the sake of clarity).

The first Lyapunov exponent $l_1(\alpha)$ eventually reads

$$\tag{A.6} \boxed{l_1(\alpha) = \frac{Re(c_1(\alpha))}{\omega(\alpha)} - \frac{\mu(\alpha)}{\omega(\alpha)^2}\mathrm{Im}(c_1(\alpha))}$$

A.2. The second Lyapunov exponent. The method to compute the second Lyapunov exponent is the same as the one we described in the previous section. The expression is given by the following formula:

$$\begin{aligned} 2l_2(0) = {}& \frac{1}{\omega(0)}\,\mathrm{Re}[g_{32}] \\ & + \frac{1}{\omega(0)^2}\,\mathrm{Im}\left[g_{20}\,\bar{g_{31}} - g_{11}\,(4\,g_{31} + 3\,\bar{g_{22}}) - \frac{1}{3}g_{02}\,(g_{40} + \bar{g_{13}}) - g_{30}\,g_{12}\right] \\ & + \frac{1}{\omega(0)^3}\left\{\mathrm{Re}\left[g_{20}\left(\bar{g_{11}}(3\,g_{12} - \bar{g_{30}}) + g_{02}\,(\bar{g_{12}} - 1/3\,g_{30}) + \frac{1}{3}\bar{g_{02}}g_{03}\right)\right.\right. \\ & \left. + g_{11}\left(\bar{g_{02}}\left(\frac{5}{3}\bar{g_{30}} + 3\,g_{12}\right) + \frac{1}{3}g_{02}\,\bar{g_{03}} - 4\,g_{11}\,g_{30}\right)\right] \\ & \left. + 3\,\mathrm{Im}[g_{20}\,g_{11}]\,\mathrm{Im}[g_{21}]\right\} \\ & + \frac{1}{\omega(0)^4}\left\{\mathrm{Im}\left[g_{11}\bar{g_{02}}\left(\bar{g_{20}}^2 - 3\,\bar{g_{20}}g_{11} - 4\,g_{11}^2\right)\right]\right. \\ & \left. + \mathrm{Im}[g_{20}\,g_{11}]\left(3\,\mathrm{Re}(g_{20}\,g_{11}) - 2\,|g_{02}|^2\right)\right\}. \end{aligned}$$

This expression is quite intricate in our case. Nevertheless, we have a closed-form expression depending on the parameter a, vanishing for two values of the parameter a. We evaluate numerically this second Lyapunov exponent. We get the following expression:

$$\begin{aligned} l_2(a) \approx {}& -0.003165\,a^{-\frac{28}{3}} - 0.1898\,a^{-\frac{22}{3}} + 0.3194\,a^{-16/3} \\ & - 0.05392\,a^{-\frac{25}{3}} + 0.1400\,a^{-\frac{19}{3}} - 0.3880\,a^{-7/3} + 0.5530\,a^{-10/3} \\ & + 0.7450\,a^{-13/3}. \end{aligned} \tag{A.7}$$

We can see that this numerical exponent vanishes only for two values of the parameter a which are

$$\{0.5304, 2.385\}.$$

The expression of the determinant of the matrix $D_{I,b}\,(\mu(I,b), l_1(I,b))$ is even more involved, and so we do not reproduce it here (it would take pages to write down its numerical expression!). Nevertheless, we proceed exactly as we did for the second Lyapunov exponent and obtain again the rigorous result that this determinant never vanishes for all $a > 0$.

Appendix B. Numerical values for the simulations. In this annex we give the numerical values used to generate Figure 4.1.

(i) Tonic Spiking $a = 1$; $b = 0.49$; $v_r = 0$; $I(t) = 1.56\mathbb{1}_{t>1}(t)$; $d = 1$; $T = 10$; $dt = 0.01$; $\theta = 10$;	(ii) Phasic Spiking $a = 1$; $b = 0.76$; $v_r = 0.2$; $I = 0.37\mathbb{1}_{t>1}(t)$; $d = 1$; $T = 10$; $dt = 0.01$; $\theta = 10$;	(iii) Tonic Bursting $a = 0.15$; $b = 1.68$; $v_r = (-2a + b)^{\frac{1}{3}}$; $I = 4.67\mathbb{1}_{t>1}(t)$; $d = 1$; $T = 30$; $dt = 0.01$; $\theta = 10$;
(iv) Phasic Bursting $a =1.58$; $b =1.70$; $v_r =-\frac{a}{4}^{\frac{1}{3}}$; $I(t) =0.73\mathbb{1}_{t>1}(t)$; $d = 0.01$; $T =50$; $dt =0.01$; $\theta =10$.	(v) Mixed Mode $a =0.07$; $b =0.32$; $v_r =0$; $I(t) =3.84\mathbb{1}_{t>1}(t)$; $d = 1.50$; $T =50$; $dt =0.01$; $\theta =10$.	(vi) Spike Freq. Adaptation $a =0.02$; $b =0.74$; $v_r =0$; $I(t) =4.33\mathbb{1}_{t>1}(t)$; $d = 0.36$; $T =50$; $dt =0.01$; $\theta =10$.
(vii) Class 1 Excitability $a =4$; $b =0.67$; $v_r =-1.3$; $I(t) =-0.1 + 0.23t$; $d = 1$; $T =30$; $dt =0.01$; $\theta =10$.	(viii) Class 2 Excitability $a =1$; $b =1.09$; $v_r =-1.2$; $I(t) =0.06t$; $d = 5$; $T =50$; $dt =0.01$; $\theta =20$.	(ix) Spike Latency $a =0.02$; $b =0.42$; $v_r =0$; $I(t) =5\delta_{7.5}(t)$; $d = 1$; $T =15$; $dt =0.01$; $\theta =10$.
(x) Damped Subthr. Oscill. $a =2.58$; $b =4.16$; $v_r =0.1$; $I(t) =2\delta_2(t)$; $d = 0.05$; $T =20$; $dt =0.01$; $\theta =10$.	(xi) Resonator $a =5.00$; $b =7.88$; $v_r =-1.28$; $I(t) =\delta_{6,6.8,15,16.5,24,26}(t)$; $d = 0.5$; $T =30$; $dt =0.01$; $\theta =10$.	(xii) Integrator $a =1.00$; $b =1.10$; $v_r =-0.97$; $I(t) =\delta_{2.5,3.3,17.5,19}(t)$; $d = 0.5$; $T =25$; $dt =0.01$; $\theta =10$.
(xiii) Rebound Spike $a =1$; $b =2$; $v_r =-0.63$; $I(t) =-0.48 - 5\delta_{2.5}(t)$; $d = 1$; $T =50$; $dt =0.1$; $\theta =10$.	(xiv) Rebound Burst $a =1$; $b =2$; $v_r =1.3$; $I(t) =-0.48 - 30\delta_{6.5}(t)$; $d = 1$; $T =20$; $dt =0.01$; $\theta =10$.	(xv) Threshold variability $a =1$; $b =1.23$; $v_r =-0.91$; $I(t) =\delta_{2,16.5} - \delta_{15}$; $d = 1$; $T =20$; $dt =0.01$; $\theta =10$.
(xvi) Bistability $a =1$; $b =1.2$; $v_r =0.8$; $I(t) =-0.47 + 20 * (\delta_{10} - \delta_{30})$; $d = 0.5$; $T =50$; $dt =0.01$; $\theta =10$.	(xvii) Depol. after-pot $a =1$; $b =1.5$; $v_r =0.06$; $I(t) =2\delta_3$; $d = 0.01$; $T =30$; $dt =0.01$; $\theta =10$.	(xviii) Self-sustained oscill. $a =1$; $b =2.5$; $v_r =-0.63$; $I(t) =-0.475 + 10 * \delta_{10}$; $d = 1$; $T =100$; $dt =0.01$; $\theta =10$.
(xix) Mixed Chatter/ C^1 exc. $a =0.89$; $b =3.65$; $v_r =1.12$; $I(t) =0.07t$; $d = 1$; $T =50$; $dt =0.01$; $\theta =10$.	(xx) Purely oscill. $a =1$; $b =2.6$; $v_r =-0.63$; $I(t) =-0.47\mathbb{1}_{t>1}$; $d = 1$; $T =500$; $dt =0.01$; $\theta =10$.	

Remark. *The $\delta_u(t)$ function is defined by*

$$\delta_{u_1,\dots,u_N}(t) = \begin{cases} 1 & \textit{if } t \in \displaystyle\bigcup_{k\in\{1,\dots,N\}} [u_k, u_k + 0.3], \\ 0 & \textit{else.} \end{cases}$$

Acknowledgments. The author warmly acknowledges T. Viéville, B. Cessac, and Olivier Faugeras for fruitful discussions and suggestions.

REFERENCES

[1] A. Alonso and R. Klink, *Differential electroresponsiveness of stellate and pyramidal-like cells of medial entorhinal cortex layer* II, J. Neurophysiol., 70 (1993), pp. 128–143.

[2] A. Alonso and R. Llinás, *Subthreshold Na+-dependent theta-like rhythmicity in stellate cells of entorhinal cortex layer* II, Nature, 342 (1989), pp. 175–177.

[3] R. Amir, M. Michaelis, and M. Devor, *Membrane potential oscillations in dorsal root ganglion neurons: Role in normal electrogenesis and neuropathic pain*, J. Neurosci., 19 (1999), pp. 8589–8596.

[4] L. S. Bernardo and R. E. Foster, *Oscillatory behavior in inferior olive neurons: Mechanism, modulation, cell agregates*, Brain Res. Bull., 17 (1986), pp. 773–784.

[5] R. Brette and W. Gerstner, *Adaptive exponential integrate-and-fire model as an effective description of neuronal activity*, J. Neurophysiol., 94 (2005), pp. 3637–3642.

[6] B. Ermentrout, M. Pascal, and B. Gutkin, *The effects of spike frequency adaptation and negative feedback on the synchronization of neural oscillators*, Neural Comput., 13 (2001), pp. 1285–1310.

[7] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[8] W. Gerstner and W. M. Kistler, *Spiking Neuron Models*, Cambridge University Press, Cambridge, UK, 2002.

[9] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.

[10] B. Gutkin, B. Ermentrout, and A. Reyes, *Phase-response curves give the responses of neurons to transient inputs*, J. Neurophysiol., 94 (2005), pp. 1623–1635.

[11] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiol., 117 (1952), pp. 500–544.

[12] F. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Springer-Verlag, New York, Secaucus, NJ, 1997.

[13] E. M. Izhikevich, *Neural excitability, spiking, and bursting*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 1171–1266.

[14] E. M. Izhikevich, *Simple model of spiking neurons*, IEEE Transactions on Neural Networks, 14 (2003), pp. 1569–1572.

[15] E. M. Izhikevich, *Which model to use for cortical spiking neurons?*, IEEE Trans. Neural Netw., 15 (2004), pp. 1063–1070.

[16] E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge, MA, 2007.

[17] R. S. G. Jones, *Synaptic and intrinsic properties of neurones of origin of the perforant path in layer* II *of the rat entorhinal cortex in vitro*, Hippocampus, 4 (1994), pp. 335–353.

[18] C. Koch and I. Segev, eds., *Methods in Neuronal Modeling: From Ions to Networks*, MIT Press, Cambridge, MA, 1998.

[19] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[20] C. Liu, M. Michaelis, R. Amir, and M. Devor, *Spinal nerve injury enhances subthreshold membrane potential oscillations in DRG neurons: Relation to neuropathic pain*, J. Neurophysiol., 84 (2000), pp. 205–215.

[21] R. Llinás, *The intrinsic electrophysiological properties of mammalian neurons: Insights into central nervous system function*, Science, 242 (1988), pp. 1654–1664.

[22] R. Llinás and Y. Yarom, *Electrophysiology of mammalian inferior olivary neurones in vitro. Different types of voltage-dependent ionic conductances*, J. Physiol., 315 (1981), pp. 549–567.

[23] R. Llinás and Y. Yarom, *Oscillatory properties of guinea-pig inferior olivary neurones and their pharmacological modulation: An in vitro study*, J. Physiol., 376 (1986), pp. 163–182.

[24] J. Rinzel and B. Ermentrout, *Analysis of Neural Excitability and Oscillations*, MIT Press, Cambridge, MA, 1989.

[25] J. A. White, T. Budde, and A. R. Kay, *A bifurcation analysis of neuronal subthreshold oscillations*, Biophys. J., 69 (1995), pp. 1203–1217.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1080–1096

METHODS FOR SOLVING ELLIPTIC PDEs IN SPHERICAL COORDINATES*

G. DASSIOS† AND A. S. FOKAS‡

Abstract. A new method for investigating boundary value problems in two dimensions has recently been introduced by one of the authors. The main achievement of this method is that it yields explicit *integral* (as oppose to series) representations for a variety of boundary value problems. In addition, this method also provides an alternative, apparently simpler, approach for deriving those solution representations that are traditionally constructed by the method of images and of classical integral transforms. Here, we implement this latter approach to boundary value problems formulated in spherical coordinates. In particular, we do the following: (a) We derive the classical Poisson integral formula for the solutions of the Dirichlet problem for the Poisson equation in the interior of a sphere, the analogous formula for the Neumann problem, and the generalizations of these formulae in n dimensions. (b) We derive the solutions of various boundary value problems for the inhomogeneous Helmholtz equation in the interior of a sphere. (c) We solve the Dirichlet problem for the Laplace equation in the interior of a spherical sector.

Key words. elliptic partial differential equations, boundary value problems, spectral methods, Fourier expansions

AMS subject classifications. 35C15, 35J05, 35J55

DOI. 10.1137/070679223

1. Introduction.

By using Green's identity and the associated fundamental solutions, it is straightforward to obtain integral representations for the solution of the Poisson and the inhomogeneous Helmholtz equations in an arbitrary three dimensional domain. These representations involve a volume integral of the forcing term as well as surface integrals of the Dirichlet and of the Neumann boundary values. This classical formulation provides the starting point for proving certain rigorous results and for constructing efficient numerical approximations. However, it does *not* provide an analytic representation of the solution, since depending on the given problem, either the Dirichlet or the Neumann boundary values are unknown. For the Poisson equation with simple boundary conditions (such as Dirichlet) formulated on simple domains (such as spheres), this difficulty can be bypassed by employing the *method of images*. More complicated boundary value problems can be solved by the *method of integral transforms*.

A new method for solving boundary value problems for linear and for integrable nonlinear PDEs in two dimensions has been introduced by one of the authors [7, 8] and implemented for a large class of problems; see [9]. For linear PDEs this method does the following: (i) It yields analytic solutions to a variety of boundary value problems, for which classical techniques are apparently ineffective. (ii) It expresses the solutions of classical problems in terms of *integrals* as opposed to the traditional *series*

*Received by the editors January 4, 2007; accepted for publication (in revised form) November 19, 2007; published electronically March 19, 2008. This work was performed under the *Marie Curie Chair of Excellence Project BRAIN*, granted to the authors by the European Commission under code EXCL 023928.

http://www.siam.org/journals/siap/68-4/67922.html

†University of Patras and ICE-HT/FORTH, Patras, Greece. Current address: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK (G.Dassios@damtp.cam.ac.uk).

‡Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK (T.Fokas@damtp.cam.ac.uk).

representations; these novel representations have analytical and numerical advantages. (iii) It provides an alternative, apparently simpler, approach for solving those problems that can be solved by the methods of images and of integral transforms.

The aim of this paper is to implement the approach (iii) above to boundary value problems formulated in spherical coordinates. Although the approach (iii) is the least novel among the approaches (i)–(iii), the implementation of the classical method of integral transforms increases in complexity as the number of dimensions increases; thus perhaps the approach (iii) becomes more useful as the number of dimensions increases.

Notation.

- The symbol "ˆ" on the top of a vector indicates that this vector has unit length.
- $\mathrm{ds}(\boldsymbol{\rho})$ and $\mathrm{dv}(\boldsymbol{\rho})$ denote surface and volume elements, respectively, at the point $\boldsymbol{\rho}$.
- $\hat{\boldsymbol{n}}$ denotes the outward unit normal on the boundary $\partial\Omega$ of a smooth domain Ω.
- $\frac{\partial}{\partial n} = \hat{\boldsymbol{n}} \cdot \nabla$ denotes the outward normal differentiation on $\partial\Omega$.

1.1. An alternative to the method of images. Let the scalar valued function $u(\boldsymbol{r})$ satisfy a linear PDE in a smooth domain Ω. The alternative to the method of images proposed here consists of the following three steps:

(a) Supplement the classical representation of the solution obtained by Green's identity and the relevant fundamental solution with an equation which is valid in the complement of the domain Ω, which will be denoted by Ω^c.

(b) Use an appropriate transformation to map the equation valid in Ω^c to an equation valid in Ω.

(c) Manipulate the *two* above equations valid in Ω to eliminate either the Neumann or the Dirichlet boundary values.

Example 1.1. Let $u(\boldsymbol{r})$ satisfy the Poisson equation in the interior of a sphere of radius α,

$$\Delta u(\boldsymbol{r}) = f(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \tag{1}$$

where the function f has sufficient smoothness. Then

$$u(\boldsymbol{r}) = u_p(\boldsymbol{r}) + u_0(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \tag{2}$$

where u_p is defined by

$$u_p(\boldsymbol{r}) = -\frac{1}{4\pi}\int_{|\boldsymbol{\rho}|<\alpha} \frac{f(\boldsymbol{\rho})}{|\boldsymbol{\rho}-\boldsymbol{r}|}\mathrm{dv}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{3}$$

and u_0 is the following harmonic function:

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi}\oint_{|\boldsymbol{\rho}|=\alpha}\left[\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|}\frac{\partial}{\partial\rho}u_0(\boldsymbol{\rho}) - u_0(\boldsymbol{\rho})\frac{\partial}{\partial\rho}\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|}\right]\mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha. \tag{4}$$

We will now use steps (a)–(c) above to determine u_0.

(a) Using

$$\frac{\partial}{\partial\rho}\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|} = -\frac{\hat{\boldsymbol{\rho}}\cdot(\boldsymbol{\rho}-\boldsymbol{r})}{|\boldsymbol{\rho}-\boldsymbol{r}|^3}$$

(4) becomes

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi} \oint_{|\boldsymbol{\rho}|=\alpha} \left[\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|} \frac{\partial}{\partial \rho} u_0(\boldsymbol{\rho}) + \frac{\alpha^2 - \boldsymbol{\rho}\cdot\boldsymbol{r}}{\alpha|\boldsymbol{\rho}-\boldsymbol{r}|^3} u_0(\boldsymbol{\rho}) \right] \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha. \tag{5}$$

The left-hand side of (5) vanishes if $|\boldsymbol{r}| > \alpha$; thus we supplement (5) with the equation

$$0 = \frac{1}{4\pi} \oint_{|\boldsymbol{\rho}|=\alpha} \left[\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|} \frac{\partial}{\partial \rho} u_0(\boldsymbol{\rho}) + \frac{\alpha^2 - \boldsymbol{\rho}\cdot\boldsymbol{r}}{\alpha|\boldsymbol{\rho}-\boldsymbol{r}|^3} u_0(\boldsymbol{\rho}) \right] \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| > \alpha. \tag{6}$$

(b) Kelvin's inversion, i.e., the transformation

$$\boldsymbol{r} \to \frac{\alpha^2}{r^2}\boldsymbol{r}, \tag{7}$$

maps the exterior to the sphere of radius α to its interior, and vice versa [11]. Furthermore, under this transformation, if $|\boldsymbol{\rho}| = \alpha$, then

$$|\boldsymbol{\rho}-\boldsymbol{r}| = \sqrt{\alpha^2 - 2\alpha r \hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}} + r^2} \to \sqrt{\alpha^2 - 2\alpha\frac{\alpha^2}{r}\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}} + \frac{\alpha^4}{r^2}} = \frac{\alpha}{r}|\boldsymbol{\rho}-\boldsymbol{r}|. \tag{8}$$

Consequently, using the transformation (7) and dropping the multiplicative factor r/α, (6) becomes

$$0 = \frac{1}{4\pi} \oint_{|\boldsymbol{\rho}|=\alpha} \left[\frac{1}{|\boldsymbol{\rho}-\boldsymbol{r}|} \frac{\partial}{\partial \rho} u_0(\boldsymbol{\rho}) + \frac{r^2 - \boldsymbol{\rho}\cdot\boldsymbol{r}}{\alpha|\boldsymbol{\rho}-\boldsymbol{r}|^3} u_0(\boldsymbol{\rho}) \right] \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{9}$$

which now holds in the interior of the sphere.

(c) Equations (5) and (9) are both valid for $|\boldsymbol{r}| < \alpha$; thus by subtracting these equations we can eliminate the Neumann boundary values $\partial_\rho u_0(\boldsymbol{\rho})$ and obtain

$$u_0(\boldsymbol{r}) = \frac{\alpha^2 - r^2}{4\pi\alpha} \oint_{|\boldsymbol{\rho}|=\alpha} \frac{u_0(\boldsymbol{\rho})}{|\boldsymbol{\rho}-\boldsymbol{r}|^3} \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha. \tag{10}$$

Equation (10) is the well-known Poisson integral formula for the solution of the Dirichlet problem [1, 2, 11, 16].

It is shown in section 2 that by further manipulating (5) and (9) we can eliminate the Dirichlet boundary value $u_0(\boldsymbol{\rho})$ and obtain

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi\alpha} \oint_{|\boldsymbol{\rho}|=\alpha} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \left[\frac{2\alpha}{|\boldsymbol{\rho}-\boldsymbol{r}|} - \ln\left(|\boldsymbol{\rho}-\boldsymbol{r}| + \frac{\boldsymbol{\rho}\cdot(\boldsymbol{\rho}-\boldsymbol{r})}{\alpha} \right) \right] \mathrm{ds}(\boldsymbol{\rho}). \tag{11}$$

Equation (2), with $u_p(\boldsymbol{r})$ and $u_0(\boldsymbol{r})$ given by (3) and (9), respectively, provides the analogue of the Poisson integral formula for the solution of the Neumann problem [1, 2, 11, 16].

The extension of these formulae to n dimensions is given in section 2.

1.2. An alternative to the method of integral transforms. If u satisfies either the Poisson or the inhomogeneous Helmholtz equations with a forcing term f in a smooth domain Ω, then Green's third identity implies the relation

$$\int_D w(\boldsymbol{\rho}) f(\boldsymbol{\rho}) \mathrm{dv}(\boldsymbol{\rho}) = \oint_{\partial D} \left[w(\boldsymbol{\rho}) \frac{\partial}{\partial n} u(\boldsymbol{\rho}) - u(\boldsymbol{\rho}) \frac{\partial}{\partial n} w(\boldsymbol{\rho}) \right] \mathrm{ds}(\boldsymbol{\rho}), \tag{12}$$

where $D \subset \Omega$ is *any* smooth subdomain of Ω and w is *any* solution of the Laplace or the Helmholtz equation, respectively. We will refer to this well-known equation as the *global relation* [9], since it connects the Dirichlet and the Neumann values on the boundary in a global sense.

1.2.1. The Dirichlet-to-Neumann map. The use of the global relation apparently provides the most effective approach for constructing the Dirichlet-to-Neumann map, i.e., computing the Neumann boundary values in terms of the Dirichlet data *directly*, without constructing the solution in the interior of the domain. This approach consists of the following steps:

(a) Apply the global relation in the domain Ω for a suitable function w.

(b) Eliminate the unknown boundary values by choosing appropriately the parameters occurring in w. This procedure yields an integral transform for the Neumann boundary values in terms of the Dirichlet data.

(c) Invert the above integral transform by using standard Sturm–Liouville techniques.

Example 1.2. Let u satisfy the Laplace equation in the spherical sector Ω:

$$\Omega = \{(\rho, \theta, \varphi) \mid 0 < \rho < \alpha,\ 0 \leqslant \theta < \theta_0,\ 0 \leqslant \varphi < 2\pi\}, \quad 0 < \theta_0 < \pi. \tag{13}$$

It is shown in section 4 that steps (a) and (b) above yield the following integral transform for the Neumann boundary values on the spherical cap $\{\rho = \alpha,\ 0 \leqslant \theta < \theta_0,\ 0 \leqslant \varphi < 2\pi\}$:

$$\int_0^{2\pi}\int_0^{\theta_0} \frac{\partial u(\alpha,\theta,\varphi)}{\partial r} P_{l_n}^m(\cos\theta) e^{im\varphi} \sin\theta \,\mathrm{d}\theta\,\mathrm{d}\varphi = M_n^m(\alpha,\theta_0), \tag{14}$$

where the constants $M_n^m(\alpha,\theta_0)$ are given in terms of the Dirichlet data in (66), $P_{l_n}^m$ denote the usual Legendre functions, and $\{l_n\}_{n=0}^\infty$ is a sequence of nonnegative real numbers defined by

$$P_{l_n}^m(\cos\theta_0) = 0, \quad n = 0, 1, 2, \dots, \quad m \in \mathbb{Z}. \tag{15}$$

1.2.2. A representation of the solution. The appropriate use of the global relation also yields the solutions in the interior. However, step (a) above is now replaced by the following:

(a′) By choosing appropriate subdomains D and appropriate solutions w, obtain a *set* of global relations.

Example 1.3. For Example 1.2, it is shown in section 4 that steps (a′) and (b) above yield the following integral transform for the solution of the Dirichlet problem in Ω:

$$\int_0^{2\pi}\int_0^{\theta_0} u(r,\theta,\varphi) P_{l_n}^m(\cos\theta) e^{im\varphi} \sin\theta \,\mathrm{d}\theta\,\mathrm{d}\varphi = K_n^m(r,\theta_0), \quad n = 0, 1, 2, \dots, \quad m \in \mathbb{Z}, \tag{16}$$

where the functions $K_n^m(r,\theta_0)$ are defined in terms of the known Dirichlet data in (70).

It is emphasized that the derivation of (14) and (16) involves only *algebraic* manipulations. The only analysis needed is the inversion of the integrals on the left-hand side of (14) and (16); see section 4.

1.2.3. A hybrid method using the fundamental solution and the global relation. For very simple boundary value problems, it is possible to obtain the solution in the interior of the domain by applying the global relation to the domain Ω instead of using a set of subdomains. For such problems our approach consists of the following three steps:

(a) Formulate the global relation in Ω.

(b) Expand the fundamental solution in terms of the functions w appearing in the global relation (12).

(c) Use the algebraic manipulation of the global relation and of the equation obtain by replacing in the classical integral representation of u the fundamental solution in terms of the expansion obtained in (b) to eliminate the unknown boundary values.

This approach has the disadvantage that it requires knowledge of the expansion described in (b), but it has the advantage that the remaining steps are entirely algebraic.

Example 1.4. Let u satisfy the inhomogeneous Helmholtz equation in the interior of a sphere of radius α,

$$(\Delta + \lambda^2)u(\boldsymbol{r}) = f(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \tag{17}$$

where the function f has sufficient smoothness and λ is a positive constant. Then u is given by (2), where u_p and u_0 are defined by

$$u_p(\boldsymbol{r}) = -\frac{1}{4\pi}\int_{|\boldsymbol{\rho}|<\alpha} f(\boldsymbol{\rho})\frac{e^{i\lambda|\boldsymbol{\rho}-\boldsymbol{r}|}}{|\boldsymbol{\rho}-\boldsymbol{r}|}\mathrm{dv}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{18}$$

and

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi}\oint_{|\boldsymbol{\rho}|=\alpha}\left[\frac{e^{i\lambda|\boldsymbol{\rho}-\boldsymbol{r}|}}{|\boldsymbol{\rho}-\boldsymbol{r}|}\frac{\partial}{\partial\rho}u_0(\boldsymbol{\rho}) - u_0(\boldsymbol{\rho})\frac{\partial}{\partial\rho}\frac{e^{i\lambda|\boldsymbol{\rho}-\boldsymbol{r}|}}{|\boldsymbol{\rho}-\boldsymbol{r}|}\right]\mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha. \tag{19}$$

We will now solve the Dirichlet boundary value problem using steps (a)–(c) above.

(a) Letting $w(\boldsymbol{\rho}) = j_n(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})$ in the global relation (12), where j_n is the spherical Bessel function of the first kind and P_n is the Legendre polynomial, we find the following equation, which is valid for any unit vector $\hat{\boldsymbol{k}}$ and any $n \in \mathbb{Z}$:

$$\begin{aligned}&\int_{|\boldsymbol{\rho}|<\alpha} f(\boldsymbol{\rho})j_n(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})\mathrm{dv}(\boldsymbol{\rho})\\ &= \oint_{|\boldsymbol{\rho}|=\alpha}\left[j_n(\lambda\rho)\frac{\partial}{\partial\rho}u(\boldsymbol{\rho}) - u(\boldsymbol{\rho})\frac{\partial}{\partial\rho}j_n(\lambda\rho)\right]P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})\mathrm{ds}(\boldsymbol{\rho})\\ &= j_n(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}\frac{\partial u(\boldsymbol{\rho})}{\partial\rho}P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})\mathrm{ds}(\boldsymbol{\rho}) - \lambda j_n'(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}u(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})\mathrm{ds}(\boldsymbol{\rho}),\end{aligned} \tag{20}$$

where for the second equality we have used that on the boundary $\rho = \alpha$.

(b) The fundamental solution for the Helmholtz equation admits the following expansion [14, 15]:

$$\frac{1}{4\pi}\frac{e^{i\lambda|\boldsymbol{\rho}-\boldsymbol{r}|}}{|\boldsymbol{\rho}-\boldsymbol{r}|} = \frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)h_n^{(1)}(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}}), \quad r < \rho, \tag{21}$$

where $h_n^{(1)}$ is the spherical Hankel function of the first kind.

(c) Using the representation (21) and (2) in (19) we find that for $r < \alpha$

$$\begin{aligned}u_0(\boldsymbol{r}) = &-\frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)\\ &\times\left[\lambda h_n^{(1)'}(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}(u(\boldsymbol{\rho}) - u_p(\boldsymbol{\rho}))P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho})\right.\\ &\left.- h_n^{(1)}(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}\frac{\partial(u(\boldsymbol{\rho}) - u_p(\boldsymbol{\rho}))}{\partial\rho}P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho})\right],\end{aligned} \tag{22}$$

where u_p is given by (18).

In order to eliminate the Neumann boundary values $\partial_\rho u(\boldsymbol{\rho})$ from (22) we let $\hat{\boldsymbol{k}} = \hat{\boldsymbol{r}}$ in (20), we multiply the resulting equation by the expression

$$-\frac{i\lambda}{4\pi}(2n+1)\frac{h_n^{(1)}(\lambda\alpha)}{j_n(\lambda\alpha)}j_n(\lambda r),$$

we use the Wronskian relation

$$n_n'(x)j_n(x) - n_n(x)j_n'(x) = \frac{1}{x^2}, \tag{23}$$

we sum over n, and we add the resulting equation to (22). This yields

$$\begin{aligned} u_0(\boldsymbol{r}) = \frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)\Bigg[&\frac{-i}{\lambda\alpha^2 j_n(\lambda\alpha)}\oint_{|\boldsymbol{\rho}|=\alpha} u(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho}) \\ &+ \lambda h_n^{(1)\prime}(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha} u_p(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho}) \\ &- h_n^{(1)}(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}\frac{\partial u_p(\boldsymbol{\rho})}{\partial\rho}P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho}) \\ &+ \frac{h_n^{(1)}(\lambda\alpha)}{j_n(\lambda\alpha)}\int_{|\boldsymbol{\rho}|<\alpha} f(\boldsymbol{\rho})j_n(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{dv}(\boldsymbol{\rho})\Bigg], \end{aligned} \tag{24}$$

where the right-hand side of (24) involves only known quantities. Equation (2), with u_p and u_0 given by (18) and (24), respectively, provides the solution of the Dirichlet boundary value problem. The Neumann and the Robin problems are solved in section 3.

2. The Poisson integral formula and its analogue for the Neumann problem in *n* dimensions.

PROPOSITION 2.1. *Let u satisfy the Poisson equation in the interior of an n dimensional sphere of radius α,*

$$\Delta u(\boldsymbol{r}) = f(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \quad \boldsymbol{r} \in \mathbb{R}^n, \quad n \geqslant 3, \tag{25}$$

with either Dirichlet

$$u(\boldsymbol{r}) = D(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \tag{26}$$

or Neumann

$$\frac{\partial u(\boldsymbol{r})}{\partial r} = N(\boldsymbol{r}), \quad |\boldsymbol{r}| < \alpha, \tag{27}$$

boundary conditions, where the functions f, D, N have sufficient smoothness. Then

$$\begin{aligned} u(\boldsymbol{r}) &= u_0(\boldsymbol{r}) - \frac{1}{(n-2)\omega_n}\int_{|\boldsymbol{\rho}|<\alpha}\frac{f(\boldsymbol{\rho})}{|\boldsymbol{\rho}-\boldsymbol{r}|^{n-2}}\mathrm{dv}(\boldsymbol{\rho}) \\ &= u_0(\boldsymbol{r}) - u_p(\boldsymbol{r}), \end{aligned} \tag{28}$$

where

$$\omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)}, \tag{29}$$

Γ *denotes the Euler Gamma function, and* u_0 *is given by the following expressions:*

For the Dirichlet problem,

$$u_0(\boldsymbol{r}) = \frac{\alpha^2 - r^2}{\alpha\omega_n} \oint_{|\boldsymbol{\rho}|=\alpha} \frac{D_0(\boldsymbol{\rho})}{|\boldsymbol{\rho} - \boldsymbol{r}|^n} \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{30}$$

where $D_0 = D - u_p$ *on the boundary.*

For the Neumann problem,

$$r\frac{\partial u_0(\boldsymbol{r})}{\partial r} = \frac{\alpha^2 - r^2}{\omega_n} \oint_{|\boldsymbol{\rho}|=\alpha} \frac{N_0(\boldsymbol{\rho})}{|\boldsymbol{\rho} - \boldsymbol{r}|^n} \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{31}$$

where $N_0 = N - u_p$ *on the boundary, or equivalently,*

$$u_0(\boldsymbol{r}) = \frac{1}{\alpha\omega_n} \oint_{|\boldsymbol{\rho}|=\alpha} I(\boldsymbol{\rho}, \boldsymbol{r}) N_0(\boldsymbol{\rho}) \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{32}$$

with

$$\begin{aligned} I(\boldsymbol{\rho}, \boldsymbol{r}) &= \int_\alpha^r \frac{\alpha(\alpha^2 - t^2)}{t|\boldsymbol{\rho} - t\hat{\boldsymbol{r}}|^n} dt \\ &= \frac{\ln(ra)}{\alpha^{n-3}} + \frac{1}{\alpha^{n-3}} \sum_{\kappa=0}^{\infty} \left[\frac{1}{\kappa} \left(\frac{r}{\alpha}\right)^{\kappa} - \frac{1}{\kappa+2} \left(\frac{r}{\alpha}\right)^{\kappa+2} \right] C_\kappa^{n/2}(\hat{\boldsymbol{\rho}} \cdot \hat{\boldsymbol{r}}), \end{aligned} \tag{33}$$

where $C_\kappa^{n/2}$ *is the Gegenbauer polynomial of degree* κ *and order* $n/2$.

For $n = 3$, (28), (30), and (32) reduce to (2), (10), and (11), respectively.

Proof. We first concentrate on the particular case of $n = 3$. It is straightforward to show that $r\partial_r u_0$ is a harmonic function. Then (10) implies

$$r\frac{\partial u_0(\boldsymbol{r})}{\partial r} = \frac{\alpha^2 - r^2}{4\pi} \oint_{|\boldsymbol{\rho}|=\alpha} \frac{1}{|\boldsymbol{\rho} - \boldsymbol{r}|^3} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \mathrm{ds}(\boldsymbol{\rho}). \tag{34}$$

Alternatively, this equation can be obtained as follows: Adding (5) and (9) we find that

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi\alpha} \oint_{|\boldsymbol{\rho}|=\alpha} \left[2\alpha \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} + u_0(\boldsymbol{\rho}) \right] \frac{1}{|\boldsymbol{\rho} - \boldsymbol{r}|} \mathrm{ds}(\boldsymbol{\rho}). \tag{35}$$

Differentiating this equation and using (9) we find (34). Integrating (34) with respect to r we obtain

$$u_0(\boldsymbol{r}) = \frac{1}{4\pi\alpha} \oint_{|\boldsymbol{\rho}|=\alpha} I(\boldsymbol{\rho}, \boldsymbol{r}) \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{36}$$

where the kernel I is given by

$$\begin{aligned} I(\boldsymbol{\rho}, \boldsymbol{r}) &= \int \frac{\alpha(\alpha^2 - r^2)}{r|\boldsymbol{\rho} - \boldsymbol{r}|^3} \mathrm{d}r \\ &= \frac{2\alpha}{|\boldsymbol{\rho} - \boldsymbol{r}|} - \ln \frac{\alpha - \boldsymbol{\rho} \cdot \hat{\boldsymbol{r}} + |\boldsymbol{\rho} - \boldsymbol{r}|}{2r}. \end{aligned} \tag{37}$$

The Gauss theorem implies

$$\oint_{|\boldsymbol{\rho}|=\alpha} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \mathrm{ds}(\boldsymbol{\rho}) = 0; \tag{38}$$

hence, we actually define the *Neumann kernel* as

$$N(\boldsymbol{\rho}, \boldsymbol{r}) = \frac{2\rho}{|\boldsymbol{\rho} - \boldsymbol{r}|} - \ln\left(|\boldsymbol{\rho} - \boldsymbol{r}| + \frac{\boldsymbol{\rho} \cdot (\boldsymbol{\rho} - \boldsymbol{r})}{\alpha}\right). \tag{39}$$

Therefore, the analogue of formula (10) for the Neumann problem is recovered.

The fundamental solution of the Laplace equation in n dimensions is given by

$$G(\boldsymbol{\rho}, \boldsymbol{r}) = -\frac{1}{(n-2)\omega_n} \frac{1}{|\boldsymbol{\rho} - \boldsymbol{r}|^{n-2}}, \quad n \geqslant 3, \tag{40}$$

where ω_n is given by (29). Using Green's third identity,

$$\begin{aligned} &\int_{|\boldsymbol{\rho}|<\alpha} [u(\boldsymbol{\rho}) \Delta_{\boldsymbol{\rho}} G(\boldsymbol{\rho}, \boldsymbol{r}) - G(\boldsymbol{\rho}, \boldsymbol{r}) \Delta_{\boldsymbol{\rho}} u(\boldsymbol{\rho})] \mathrm{dv}(\boldsymbol{\rho}) \\ &= \oint_{|\boldsymbol{\rho}|=\alpha} [u(\boldsymbol{\rho}) \partial_\rho G(\boldsymbol{\rho}, \boldsymbol{r}) - G(\boldsymbol{\rho}, \boldsymbol{r}) \partial_\rho u(\boldsymbol{\rho})] \mathrm{ds}(\boldsymbol{\rho}), \end{aligned} \tag{41}$$

we find (28), where u_0 satisfies Laplace's equation. In view of the identity

$$\frac{\partial}{\partial \rho} \frac{1}{|\boldsymbol{\rho} - \boldsymbol{r}|^{n-2}} = (2-n) \frac{\boldsymbol{\rho} \cdot (\boldsymbol{\rho} - \boldsymbol{r})}{\rho |\boldsymbol{\rho} - \boldsymbol{r}|^n}, \tag{42}$$

(40) implies

$$\oint_{|\boldsymbol{\rho}|=\alpha} \left[u_0(\boldsymbol{\rho}) \frac{\boldsymbol{\rho} \cdot (\boldsymbol{\rho} - \boldsymbol{r})}{\alpha |\boldsymbol{\rho} - \boldsymbol{r}|^n} + \frac{1}{(n-2)|\boldsymbol{\rho} - \boldsymbol{r}|^{n-2}} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \right] \mathrm{ds}(\boldsymbol{\rho}) = \begin{cases} \omega_n u_0(\boldsymbol{r}), & |\boldsymbol{r}| < \alpha, \\ 0, & |\boldsymbol{r}| > \alpha. \end{cases} \tag{43}$$

Applying Kelvin's inversion (7), (8) to (43) we obtain

$$\oint_{|\boldsymbol{\rho}|=\alpha} \left[u_0(\boldsymbol{\rho}) \frac{r^2 - \boldsymbol{\rho} \cdot \boldsymbol{r}}{\alpha |\boldsymbol{\rho} - \boldsymbol{r}|^n} + \frac{1}{(n-2)|\boldsymbol{\rho} - \boldsymbol{r}|^{n-2}} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} \right] \mathrm{ds}(\boldsymbol{\rho}) = 0, \quad |\boldsymbol{r}| < \alpha. \tag{44}$$

Subtracting (44) from the first of (43) we obtain Poisson's integral formula (30).

On the other hand, adding the first of (43) and (44) we obtain

$$\omega_n u_0(\boldsymbol{r}) = \oint_{|\boldsymbol{\rho}|=\alpha} \left[\frac{2}{n-2} \frac{\partial u_0(\boldsymbol{\rho})}{\partial \rho} + \frac{1}{\alpha} u_0(\boldsymbol{\rho}) \right] \frac{1}{|\boldsymbol{\rho} - \boldsymbol{r}|^{n-2}} \mathrm{ds}(\boldsymbol{\rho}). \tag{45}$$

Dividing (31) by r and integrating with respect to r we obtain

$$u_0(\boldsymbol{r}) = \frac{1}{\alpha \omega_n} \oint_{|\boldsymbol{\rho}|=\alpha} I(\boldsymbol{\rho}, \boldsymbol{r}) N_0(\boldsymbol{\rho}) \mathrm{ds}(\boldsymbol{\rho}), \quad |\boldsymbol{r}| < \alpha, \tag{46}$$

where

$$\begin{aligned} I(\boldsymbol{\rho}, \boldsymbol{r})|_{|\boldsymbol{\rho}|=\alpha} &= \int \frac{\alpha(\alpha^2 - r^2)}{r |\boldsymbol{\rho} - \boldsymbol{r}|^n} \mathrm{d}r \\ &= \int \frac{\alpha(\alpha^2 - r^2)}{r(\alpha^2 - 2\alpha r \hat{\boldsymbol{\rho}} \cdot \hat{\boldsymbol{r}} + r^2)^{n/2}} \mathrm{d}r. \end{aligned} \tag{47}$$

The Gauss theorem

$$\oint_{|\boldsymbol{\rho}|=\alpha} N_0(\boldsymbol{\rho}) \mathrm{ds}(\boldsymbol{\rho}) = 0 \tag{48}$$

implies that $I(\boldsymbol{\rho}, \boldsymbol{r})$ is well defined independent of the constant of integration. For any specific value of n the integral in (47) can be evaluated explicitly. Alternatively, it can be expanded in terms of Gegenbauer polynomials of order $n/2$ [12]. Indeed, the Gegenbauer polynomials $C_\kappa^{n/2}(\gamma)$, $\gamma = \hat{\boldsymbol{\rho}} \cdot \hat{\boldsymbol{r}}$, $\kappa = 0, 1, 2, \ldots$, are generated as follows:

$$(1 - 2\gamma z + z^2)^{-n/2} = \sum_{\kappa=0}^{\infty} C_\kappa^{n/2}(\gamma) z^\kappa, \qquad |z| < 1. \tag{49}$$

Hence

$$|\boldsymbol{\rho} - \boldsymbol{r}|^{-n}|_{|\boldsymbol{\rho}|=\alpha} = \frac{1}{\alpha^n} \sum_{\kappa=0}^{\infty} C_\kappa^{n/2}(\gamma) \left(\frac{r}{\alpha}\right)^\kappa, \qquad |z| < \alpha, \tag{50}$$

and

$$\begin{aligned} I(\boldsymbol{\rho}, \boldsymbol{r})|_{|\boldsymbol{\rho}|=\alpha} &= \frac{1}{\alpha^{n-1}} \sum_{\kappa=0}^{\infty} C_\kappa^{n/2}(\gamma) \frac{1}{\alpha^\kappa} \int (\alpha^2 - r^2) r^{\kappa-1} \mathrm{d}r \\ &= \frac{\ln r}{\alpha^{n-3}} + \frac{1}{\alpha^{n-3}} \sum_{\kappa=1}^{\infty} C_\kappa^{n/2}(\gamma) \frac{1}{\kappa} \left(\frac{r}{\alpha}\right)^\kappa - \frac{1}{\alpha^{n-3}} \sum_{\kappa=0}^{\infty} C_\kappa^{n/2}(\gamma) \frac{1}{\kappa+2} \left(\frac{r}{\alpha}\right)^{\kappa+2}. \end{aligned} \tag{51}$$

This completes the proof of Proposition 2.1. □

Remark 1. Equation (45) with $n = 3$ provides a new integral representation for the solution of the Dirichlet problem involving only $|\boldsymbol{\rho} - \boldsymbol{r}|$, as opposed to $|\boldsymbol{\rho} - \boldsymbol{r}|^3$.

Remark 2. For the corresponding exterior problems [5], one needs to impose the far-field behavior

$$u_0(\boldsymbol{r}) = O\left(\frac{1}{r^{n-2}}\right), \quad r \to \infty, \tag{52}$$

and to change the sign of every surface integral.

3. The inhomogeneous Helmholtz equation. We first recall the derivation of the well-known global relation (12). Let u satisfy the inhomogeneous Helmholtz equation (17) and let w satisfy the Helmholtz equation. Manipulating the two equations we find that

$$\nabla \cdot [w(\boldsymbol{r}) \nabla u(\boldsymbol{r}) - u(\boldsymbol{r}) \nabla w(\boldsymbol{r})] = w(\boldsymbol{r}) f(\boldsymbol{r}) \tag{53}$$

and the Gauss theorem yields (12).

PROPOSITION 3.1. *Suppose that the function u satisfies the inhomogeneous equation* (17) *in $\mathbb{R}^3$ together with any of the three boundary conditions*

$$u(\boldsymbol{r}) = D(\boldsymbol{r}), \quad |\boldsymbol{r}| = \alpha, \tag{54}$$

$$\frac{\partial u(\boldsymbol{r})}{\partial r} = N(\boldsymbol{r}), \quad |\boldsymbol{r}| = \alpha, \tag{55}$$

$$\frac{\partial u(\boldsymbol{r})}{\partial r} + \nu u(\boldsymbol{r}) = R(\boldsymbol{r}), \quad |\boldsymbol{r}| = \alpha, \tag{56}$$

where the given functions D, N, R have sufficient smoothness and ν is a constant such that $n+\alpha\nu \neq 0$, $n = 0, 1, 2, \ldots$. Then u is given by (2)*, where u_p is defined by* (18)*, and u_0 for the Dirichlet, the Neumann, and the Robin conditions is given by the following expressions:*

$$\begin{aligned} u_0^D(\boldsymbol{r}) = \frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)\Bigg[&T(\hat{\boldsymbol{r}}) + \frac{h_n^{(1)}(\lambda\alpha)}{j_n(\lambda\alpha)}S(\hat{\boldsymbol{r}}) \\ &- \frac{i}{\lambda\alpha^2 j_n(\lambda\alpha)}\oint_{|\boldsymbol{\rho}|=\alpha} D(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho})\Bigg], \end{aligned} \tag{57}$$

$$\begin{aligned} u_0^N(\boldsymbol{r}) = \frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)\Bigg[&T(\hat{\boldsymbol{r}}) - \frac{h_n^{(1)'}(\lambda\alpha)}{j_n'(\lambda\alpha)}S(\hat{\boldsymbol{r}}) \\ &+ \frac{i}{\lambda^2\alpha^2 j_n'(\lambda\alpha)}\oint_{|\boldsymbol{\rho}|=\alpha} N(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho})\Bigg], \end{aligned} \tag{58}$$

$$\begin{aligned} u_0^R(\boldsymbol{r}) = \frac{i\lambda}{4\pi}\sum_{n=0}^{\infty}(2n+1)j_n(\lambda r)\Bigg[&T(\hat{\boldsymbol{r}}) + \frac{\lambda h_n^{(1)'}(\lambda\alpha) + \nu h_n^{(1)}(\lambda\alpha)}{\lambda j_n'(\lambda\alpha) + j_n(\lambda\alpha)}S(\hat{\boldsymbol{r}}) \\ &- \frac{i}{\lambda\alpha^2(\lambda j_n'(\lambda\alpha) + j_n(\lambda\alpha))}\oint_{|\boldsymbol{\rho}|=\alpha} R(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho})\Bigg], \end{aligned} \tag{59}$$

where the direction dependent functions T and S are defined by

$$\begin{aligned} T(\hat{\boldsymbol{r}}) = \lambda h_n^{(1)'}(\lambda\alpha)&\oint_{|\boldsymbol{\rho}|=\alpha} u_p(\boldsymbol{\rho})P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho}) \\ &- h_n^{(1)}(\lambda\alpha)\oint_{|\boldsymbol{\rho}|=\alpha}(\partial_\rho u_p(\boldsymbol{\rho}))P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{ds}(\boldsymbol{\rho}) \end{aligned} \tag{60}$$

and

$$S(\hat{\boldsymbol{r}}) = \int_{|\boldsymbol{\rho}|<\alpha} f(\boldsymbol{\rho})j_n(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{r}})\mathrm{dv}(\boldsymbol{\rho}). \tag{61}$$

In the formulae above we have used the notation

$$\left[\frac{\partial}{\partial\rho}j_n(\lambda\rho)\right]_{\rho=\alpha} = \lambda j_n'(\lambda\alpha) \tag{62}$$

and similarly for $h_n^{(1)}$.

Proof. The derivation of u_0^D was demonstrated in the introduction (Example 1.4). For the derivation of u_0^N and u_0^R we follow similar arguments. □

Remark 3. It is straightforward to solve the analogous problem in the exterior of the sphere. In fact, for exterior problems the following modifications are required: First, one needs to specify the appropriate asymptotic behavior of the solution at infinity. For a compactly supported source function f, the solution of the inhomogeneous Helmholtz equation has to satisfy the radiation condition [5]

$$u(\boldsymbol{r}) = g(\hat{\boldsymbol{r}})\frac{\mathrm{e}^{i\lambda r}}{r} + O(r^{-2}), \quad r \to \infty, \tag{63}$$

where g is the far-field pattern.

Second, the terms in the surface integral representation involving the normal differentiation on the boundary must change sign.

Third, one must choose an appropriate set of eigenfunctions. In particular, for the sphere the interior eigenfunctions have to be replaced by the radiative functions $w(\boldsymbol{\rho}) = h_n^{(1)}(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})$, where $n = 0, 1, 2, \dots$ and $\hat{\boldsymbol{k}}$ is a unit vector. We note that the above choice of exterior eigenfunctions is dictated by the radiation condition (63). If the radiation condition is taken in the form

$$u(\boldsymbol{r}) = g(\hat{\boldsymbol{r}})\frac{\mathrm{e}^{-i\lambda r}}{r} + O(r^{-2}), \quad r \to \infty, \tag{64}$$

then the proper choice of exterior eigenfunctions is $w(\boldsymbol{\rho}) = h_n^{(2)}(\lambda\rho)P_n(\hat{\boldsymbol{\rho}}\cdot\hat{\boldsymbol{k}})$. For scattering problems these choices will ensure that as we approach infinity, the solutions tend to outgoing spherical waves having the appropriate geometrical attenuation [5].

4. Laplace's equation in a spherical sector.

Separation of variables, for the Laplace equation in spherical coordinates, yields the following solutions:

$$w_1^{\pm}(\boldsymbol{r}) = r^l P_l^k(\cos\theta)e^{\pm ik\varphi} \quad \text{and} \quad w_2^{\pm}(\boldsymbol{r}) = r^{-(l+1)}P_l^k(\cos\theta)e^{\pm ik\varphi}, \tag{65}$$

where l and k are arbitrary complex constants.

4.1. The Dirichlet-to-Neumann map.

PROPOSITION 4.1. *Let u satisfy Laplace's equation in the spherical sector Ω defined in* (13). *Then the Neumann boundary values on the spherical cap $\{\rho = \alpha, 0 \leqslant \theta < \theta_0, 0 \leqslant \varphi < 2\pi\}$ can be expressed in terms of the Dirichlet boundary values by* (14)*, where l_n is defined by* (15) *and M_n^m are given by*

$$\begin{aligned} M_n^m(\alpha,\theta_0) = {} & \frac{\sin\theta_0}{\alpha^{l_n+2}}\frac{\partial P_{l_n}^m(\cos\theta_0)}{\partial\theta}\int_0^{2\pi}\int_0^{\alpha} u(r',\theta_0,\varphi)(r')^{l_n+1}e^{im\varphi}\,\mathrm{d}r'\,\mathrm{d}\varphi \\ & - \frac{l_n}{\alpha}\int_0^{2\pi}\int_0^{\theta_0} u(\alpha,\theta,\varphi)P_{l_n}^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi. \end{aligned} \tag{66}$$

Proof. We will derive (14) by applying steps (a) and (b) of section 1.2.1. Applying the global relation (12) in Ω (see Figure 1) with $w = w_1^+$ defined in (65), we find that

$$\begin{aligned} & \int_0^{2\pi}\int_0^{\alpha}\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}\rho^{l+1}P_l^k(\cos\theta_0)e^{ik\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ & + \int_0^{2\pi}\int_0^{\theta_0}\frac{\partial u(\alpha,\theta,\varphi)}{\partial r}\alpha^{l+2}P_l^k(\cos\theta)e^{ik\varphi}\sin\theta\mathrm{d}\theta\,\mathrm{d}\varphi \\ & + \int_0^{\theta_0}\int_0^{\alpha}\left(iku(\rho,\theta,0) - \frac{\partial u(\rho,\theta,0)}{\partial\varphi}\right)\rho^{l-1}P_l^k(\cos\theta)\,\mathrm{d}\rho\,\mathrm{d}\theta \\ & = \int_0^{2\pi}\int_0^{\alpha} u(\rho,\theta_0,\varphi)\rho^{l+1}\frac{\partial P_l^k(\cos\theta_0)}{\partial\theta}e^{ik\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ & + \int_0^{2\pi}\int_0^{\theta_0} u(\alpha,\theta,\varphi)l\alpha^{l+1}\sin\theta\,P_l^k(\cos\theta)e^{ik\varphi}\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ & + \int_0^{\theta_0}\int_0^{\alpha}\left(iku(\rho,\theta,2\pi) - \frac{\partial u(\rho,\theta,2\pi)}{\partial\varphi}\right)\rho^{l-1}P_l^k(\cos\theta)e^{i2k\pi}\,\mathrm{d}\rho\,\mathrm{d}\theta, \end{aligned} \tag{67}$$

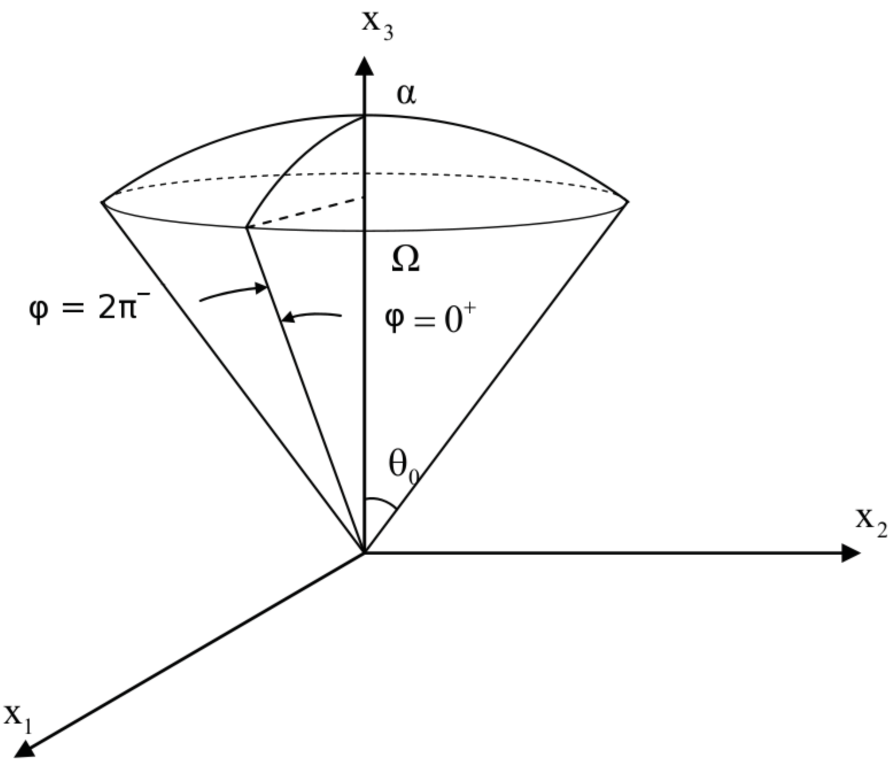


FIG. 1.

where the angular normal derivatives are given by

$$\hat{\boldsymbol{\theta}}\cdot\nabla = \frac{1}{r}\frac{\partial}{\partial\theta} \qquad \text{and} \qquad \hat{\boldsymbol{\varphi}}\cdot\nabla = \frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi}. \tag{68}$$

The third integral on the left-hand side and the third integral on the right-hand side of (67) cancel each other if we choose $k = m \in \mathbb{Z}$. Then (67) becomes

$$\begin{aligned} &\int_0^{2\pi}\int_0^{\alpha} \frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}\rho^{l+1}P_l^m(\cos\theta_0)e^{im\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ &+\int_0^{2\pi}\int_0^{\theta_0} \frac{\partial u(\alpha,\theta,\varphi)}{\partial r}\alpha^{l+2}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &=\int_0^{2\pi}\int_0^{\alpha} u(\rho,\theta_0,\varphi)\rho^{l+1}\frac{\partial P_l^m(\cos\theta_0)}{\partial\theta}e^{im\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ &+\int_0^{2\pi}\int_0^{\theta_0} u(\alpha,\theta,\varphi)l\alpha^{l+1}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi. \end{aligned} \tag{69}$$

By choosing $l = l_n$, as defined by (15), (69) becomes (15) with M_n^m given by (66). ▯

4.2. The spherical sector problem. We will now derive (16) by applying step (a′) of section 1.2.2 and step (b) of section 1.2.1.

PROPOSITION 4.2. *Let u be as in Proposition* 4.1. *Then the solution u of the Dirichlet problem satisfies* (16), *where l_n are defined by* (15) *and K_n^m are given in terms of the known Dirichlet values by*

$$\begin{aligned} K_n^m(r,\theta_0) &= \left(\frac{r}{\alpha}\right)^{l_n+1}\int_0^{2\pi}\int_0^{\theta_0} u(\alpha,\theta,\varphi)P_{l_n}^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &\quad -\left(\frac{r}{\alpha}\right)^{l_n+1}\frac{\sin\theta_0}{2l_n+1}\frac{\partial P_{l_n}^m(\cos\theta_0)}{\partial\theta} \\ &\quad\times\left\{\int_0^{2\pi}\int_0^{r} u(\rho,\theta_0,\varphi)\left[\left(\frac{\alpha}{r}\right)^{2l_n+1}-1\right]\left(\frac{\rho}{\alpha}\right)^{l_n+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi\right. \\ &\quad\left.+\int_0^{2\pi}\int_r^{\alpha} u(\rho,\theta_0,\varphi)\left[\left(\frac{\alpha}{\rho}\right)^{2l_n+1}-1\right]\left(\frac{\rho}{\alpha}\right)^{l_n+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi\right\}. \end{aligned} \tag{70}$$

Proof. We apply the global relation (12) (with $f=0$) in the subdomain Ω_1 defined by

$$\Omega_1 = \{(\rho,\theta,\varphi)\mid 0<\rho<r,\ 0\leqslant\theta<\theta_0,\ 0\leqslant\varphi<2\pi\},\quad 0<\theta_0<\pi, \tag{71}$$

and depicted in Figure 2, with $w=w_1^+$ defined in (65). This yields the following equation, which is valid for $k\in\mathbb{C}$ and $\operatorname{Re} l\geqslant 0$:

$$\begin{aligned} &\int_0^{2\pi}\int_0^{r}\left[\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}P_l^k(\cos\theta_0)-u(\rho,\theta_0,\varphi)\frac{\partial P_l^k(\cos\theta_0)}{\partial\theta}\right]\rho^{l+1}e^{ik\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ &+\int_0^{2\pi}\int_0^{\theta_0}\left[r\frac{\partial u(r,\theta,\varphi)}{\partial r}-lu(r,\theta,\varphi)\right]r^{l+1}P_l^k(\cos\theta)e^{ik\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &+\int_0^{\theta_0}\int_0^{r}\left(iku(\rho,\theta,0)-\frac{\partial u(\rho,\theta,0)}{\partial\varphi}\right)\rho^{l-1}P_l^k(\cos\theta)\,\mathrm{d}\rho\,\mathrm{d}\theta \\ &-\int_0^{\theta_0}\int_0^{r}\left(iku(\rho,\theta,2\pi)-\frac{\partial u(\rho,\theta,2\pi)}{\partial\varphi}\right)\rho^{l-1}P_l^k(\cos\theta)e^{i2k\pi}\,\mathrm{d}\rho\,\mathrm{d}\theta \\ &=0. \end{aligned} \tag{72}$$

The requirement $\operatorname{Re} l\geqslant 0$ is needed in order for the integrals to make sense near $r=0$. The last two integrals in (72) involve the Neumann boundary values $\partial_\varphi u(\rho,\theta,0)$ and $\partial_\varphi u(\rho,\theta,2\pi)$. However, these unknown functions can be eliminated by choosing $k=m\in\mathbb{Z}$.

We next apply the global relation (12) (with $f=0$) in the domain $\Omega_2=\Omega-\Omega_1$ with $w=w_1^+$ and with $w=w_2^+$ defined in (65), where $\operatorname{Re} l\geqslant 0$ and $k=m\in\mathbb{Z}$. This yields the following two equations, which are valid for $\operatorname{Re} l\geqslant 0$ and $m\in\mathbb{Z}$:

$$\begin{aligned} &\int_0^{2\pi}\int_r^{\alpha}\left[\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}P_l^m(\cos\theta_0)-u(\rho,\theta_0,\varphi)\frac{\partial P_l^m(\cos\theta_0)}{\partial\theta}\right]\rho^{l+1}e^{im\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi \\ &+\int_0^{2\pi}\int_0^{\theta_0}\left[\alpha\frac{\partial u(\alpha,\theta,\varphi)}{\partial r}+lu(\alpha,\theta,\varphi)\right]\alpha^{l+1}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &-\int_0^{2\pi}\int_0^{\theta_0}\left[r\frac{\partial u(r,\theta,\varphi)}{\partial r}-lu(r,\theta,\varphi)\right]r^{l+1}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &=0 \end{aligned} \tag{73}$$

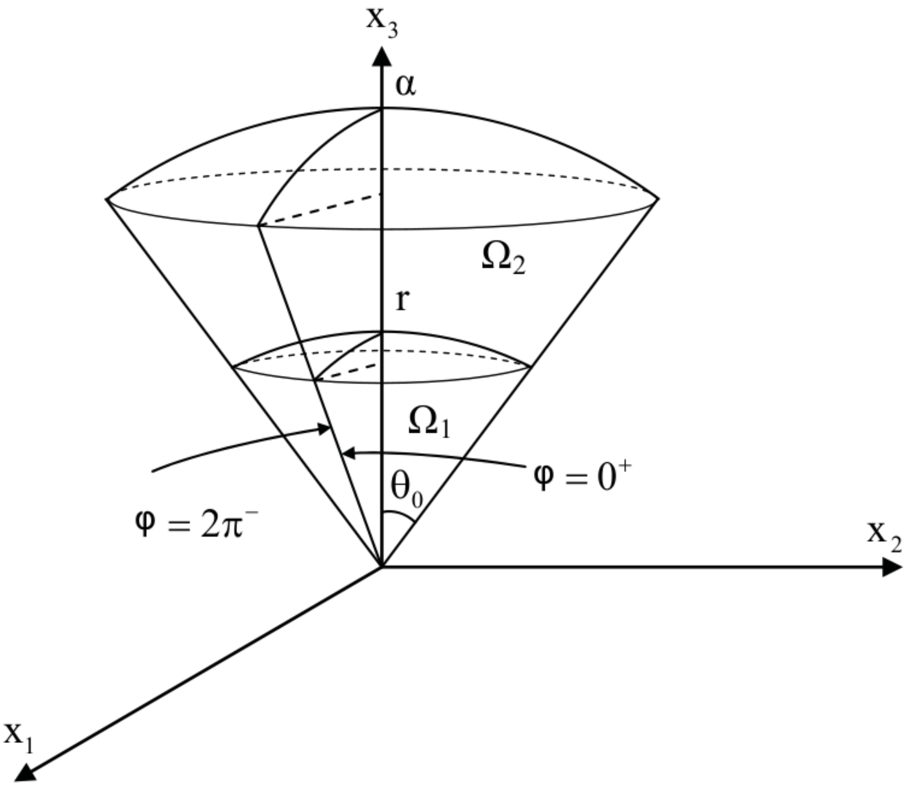


FIG. 2.

and

$$(74)\quad \int_0^{2\pi}\int_r^{\alpha}\left[\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}P_l^m(\cos\theta_0)-u(\rho,\theta_0,\varphi)\frac{\partial P_l^m(\cos\theta_0)}{\partial\theta}\right]\frac{1}{\rho^l}e^{im\varphi}\sin\theta_0\,\mathrm{d}\rho\,\mathrm{d}\varphi$$
$$+\int_0^{2\pi}\int_0^{\theta_0}\left[\alpha\frac{\partial u(\alpha,\theta,\varphi)}{\partial r}+(l+1)u(\alpha,\theta,\varphi)\right]\frac{1}{\alpha^l}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi$$
$$-\int_0^{2\pi}\int_0^{\theta_0}\left[r\frac{\partial u(r,\theta,\varphi)}{\partial r}+(l+1)u(r,\theta,\varphi)\right]\frac{1}{r^l}P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi$$
$$=0.$$

Multiplying (74) by α^{2l+1} and subtracting the resulting equation from (73) we obtain an equation that involves only the integral over the spherical cap of the unknown function $\partial_r u(r,\theta,\varphi)$. But this unknown integral can be eliminated with the use of (72). These steps yield

$$(75)\qquad P_l^m(\cos\theta_0)\sin\theta_0\int_0^{2\pi}\int_0^{r}\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}\left(\frac{\alpha^{2l+1}}{r^{2l+1}}-1\right)\rho^{l+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi$$
$$+P_l^m(\cos\theta_0)\sin\theta_0\int_0^{2\pi}\int_r^{\alpha}\frac{\partial u(\rho,\theta_0,\varphi)}{\partial\theta}\left(\frac{\alpha^{2l+1}}{\rho^{2l+1}}-1\right)\rho^{l+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi$$
$$-\frac{\partial P_l^m(\cos\theta_0)}{\partial\theta}\sin\theta_0\int_0^{2\pi}\int_0^{r}u(\rho,\theta_0,\varphi)\left(\frac{\alpha^{2l+1}}{r^{2l+1}}-1\right)\rho^{l+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi$$

$$
\begin{aligned}
&- \frac{\partial P_l^m(\cos\theta_0)}{\partial\theta}\sin\theta_0 \int_0^{2\pi}\int_r^{\alpha} u(\rho,\theta_0,\varphi)\left(\frac{\alpha^{2l+1}}{\rho^{2l+1}}-1\right)\rho^{l+1}e^{im\varphi}\,\mathrm{d}\rho\,\mathrm{d}\varphi \\
&+ (2l+1)\alpha^{l+1}\int_0^{2\pi}\int_0^{\theta_0}\left(u(\alpha,\theta,\varphi)-\frac{\alpha^l}{r^l}u(r,\theta,\varphi)\right)P_l^m(\cos\theta)e^{im\varphi}\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\
&= 0.
\end{aligned}
$$

In order to eliminate the unknown boundary values $\partial_\theta u(\rho,\theta_0,\varphi)$ we choose l such that

$$P_{l_n}^m(\cos\theta_0) = 0, \quad n = 0,1,2,\dots, \quad m \in \mathbb{Z}. \tag{76}$$

It is known that, for every order m and every angle $\theta_0 \in (0,\pi)$, there exists a sequence $\{l_n\}_{n=0}^{\infty}$ of nonnegative real numbers such that (76) holds; see [10, page 408]. With the above choice of k and l, (75) implies (16) with K_n^m defined by (70). □

Remark 4. The formal inversion of (16) yields

$$u(r,\theta,\varphi) = \frac{1}{2\pi}\sum_{n=0}^{\infty}\sum_{m\in\mathbb{Z}}\frac{1}{c_n^2}K_n^m(r,\theta_0)P_{l_n}^m(\cos\theta)e^{-im\varphi}. \tag{77}$$

Indeed, the inversion with respect to φ is elementary and gives

$$\int_0^{\theta_0} u(r,\theta,\varphi)P_{l_n}^m(\cos\theta)\sin\theta\,\mathrm{d}\theta = \frac{1}{2\pi}\sum_{m\in\mathbb{Z}}K_n^m(\boldsymbol{r})e^{-im\varphi}. \tag{78}$$

In order to invert the left-hand side of (78) we use the following standard Sturm–Liouville technique. We set $x=\cos\theta$ and write the Legendre equation for l_n and $l_{n'}$:

$$\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}P_{l_n}^m(x)\right] + \left[l_n(l_n+1)-\frac{m^2}{1-x^2}\right]P_{l_n}^m(x) = 0 \tag{79}$$

and

$$\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}P_{l_{n'}}^m(x)\right] + \left[l_{n'}(l_{n'}+1)-\frac{m^2}{1-x^2}\right]P_{l_{n'}}^m(x) = 0. \tag{80}$$

Multiplying (79) by $P_{l_{n'}}^m(x)$, (80) by $P_{l_n}^m(x)$, and subtracting the resulting equations we obtain

$$
\begin{aligned}
&\frac{d}{dx}\left[(1-x^2)P_{l_{n'}}^m(x)\frac{d}{dx}P_{l_n}^m(x) - (1-x^2)P_{l_n}^m(x)\frac{d}{dx}P_{l_{n'}}^m(x)\right] \\
&= (l_{n'}-l_n)(l_{n'}+l_n+1)P_{l_n}^m(x)P_{l_{n'}}^m(x).
\end{aligned} \tag{81}
$$

Integrating (81) from $\cos 0 = 1$ to $\cos\theta_0 = x_0$ and using (76) we find that

$$\int_1^{x_0} P_{l_n}^m(x)P_{l_{n'}}^m(x)dx = \begin{cases} 0, & n \neq n', \\ c_n^2, & n = n', \end{cases} \tag{82}$$

where the normalization constants c_n^2 can be evaluated by using L'Hôpital's rule,

(83)

$$
\begin{aligned}
c_n^2 &= \lim_{l\to l_n}\frac{1}{(l-l_n)(l+l_n+1)}\left[(1-x^2)\left(P_l^m(x)\frac{d}{dx}P_{l_n}^m(x) - P_{l_n}^m(x)\frac{d}{dx}P_l^m(x)\right)\right]\Bigg|_{x=1}^{x=x_0} \\
&= \frac{1-x_0^2}{2l_n+1}\left[\frac{d}{dl}P_l^m(x_0)\right]\Bigg|_{l=l_n}\left[\frac{d}{dx}P_{l_n}^m(x)\right]\Bigg|_{x=x_0}.
\end{aligned}
$$

We note that the series in (77) converges due to the factors $(r/\alpha)^{l_n}$, with $r < \alpha$ and $\operatorname{Re} l_n > 0$, which enter the functions K_n^m defined in (70). This result is valid, provided that the set $\{P_{l_n}^m(x)\}$ with l_n defined by (76) forms a complete set. We will not pursue further the question of completeness because in a future work we will present an alternative derivation of this result which bypasses this difficulty (see the discussion in section 5).

Remark 5. As with the previous cases, the extension of this result to the solution of the Dirichlet problem in the exterior of the spherical sector

$$\Omega^c = \{(r, \theta, \varphi) \mid \alpha < r < \infty,\, 0 \leqslant \theta < \theta_0,\, 0 \leqslant \varphi < 2\pi\} \tag{84}$$

is straightforward.

5. Conclusions.

We have presented alternative, apparently simpler, approaches to the classical methods of images and integral transforms for boundary value problems formulated in spherical coordinates.

Regarding the alternative to the method of images presented here, we also note that the formulae for the solution of the Neumann problem in the interior of an n dimensional sphere, $n \geqslant 3$, presented in section 2, are to our knowledge new.

Regarding the alternative to the method of integral transforms presented here, we note that the solution of certain PDEs depends on the *global form* of the boundary values; namely, it depends on certain integrals of the boundary values. The advantage of the global relation is that it yields *directly* these global forms, and in this sense the term *global relation* is justified.

The employment of the global relation in the domain Ω provides the most efficient way of constructing the Dirichlet-to-Neumann map [6]. For example, it is shown in section 4.1 that it is possible, using *only algebraic manipulations*, to obtain a certain *integral* of the Neumann boundary values on a spherical cap in terms of the given Dirichlet data; see (14). Then it is straightforward to obtain the Dirichlet boundary value itself by inverting this integral. It is well known that for a variety of physical applications, one is interested only in the unknown Neumann values on the boundary and *not* on the actual solution in the interior of the domain. For such problems, the approach presented here provides, in our opinion, the most direct and simple way of obtaining the unknown boundary values.

The analysis of the global relation, in addition to constructing the Dirichlet-to-Neumann map, can also yield the solution in Ω. This requires formulating the global relation in appropriate *subdomains* of Ω; see section 4.2.

The new, almost algebraic, alternative to the method of classical transforms should be compared with the latter method, which involves the following steps: (a) Derive the proper transform. (b) Use integration by parts to obtain the PDE satisfied by this transform. (c) Solve the resulting PDE by an integral transform, which in fact involves (c_1) deriving the proper transform, (c_2) using integration by part to obtain the ODE satisfied by this transform, and (c_3) using Green's function techniques to solve this ODE.

It should be noted that integral representations obtained by Green's identity and the appropriate fundamental solution have been used extensively in the literature; see [1, 2, 3, 4, 5, 11, 13, 14, 15]. In most of these approaches, one first characterizes the unknown boundary values by evaluating the integral representation on the boundary and solving the resulting integral equation, and then one inserts the density function for the single or double potential in the integral representation. Our approach has certain conceptual similarities with the method of Waterman and with the null-field method [13] used in scattering theory.

It was mentioned in the introduction that the most important achievement of the method reviewed in [9] is the construction of integral, as opposed to series representations in the spectral domain. The problem of constructing such representations for the problems solved in sections 3 and 4 is under investigation.

REFERENCES

[1] G. Barton, *Elements of Green's Functions and Propagation: Potentials, Diffusion and Waves*, Clarendon Press, Oxford, UK, 1989.

[2] S. Bergman and M. Schiffer, *Kernel Functions and Elliptic Differential Equations in Mathematical Physics*, Academic Press, New York, 1953.

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics* I, John Wiley and Sons, New York, 1989.

[4] R. Courant and D. Hilbert, *Methods of Mathematical Physics* II, John Wiley and Sons, New York, 1989.

[5] G. Dassios and R. E. Kleinman, *Low Frequency Scattering*, Oxford University Press, Oxford, UK, 2000.

[6] G. Dassios and A. S. Fokas, *The basic elliptic equations in an equilateral triangle*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 2721–2748.

[7] A. S. Fokas, *A unified transform method for solving linear and certain non-linear PDE's*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1411–1443.

[8] A. S. Fokas, *Two-dimensional linear partial differential equations in a convex polygon*, Proc. Roy. Soc. London Ser. A, 457 (2001), pp. 371–393.

[9] A. S. Fokas, *A Unified Approach to Boundary Value Problems*, SIAM, Philadelphia, to appear.

[10] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, Cambridge University Press, Cambridge, UK, 1931.

[11] O. D. Kellogg, *Foundations of Potential Theory*, Dover, New York, 1953.

[12] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[13] P. A. Martin, *Multiple Scattering*, Cambridge University Press, Cambridge, UK, 2006.

[14] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* I, McGraw–Hill, New York, 1953.

[15] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* II, McGraw–Hill, New York, 1953.

[16] A. D. Polyanin, *Handbook of Linear Partial Differential Equations for Engineers and Scientists*, Chapman and Hall/CRC Press, Boca Raton, FL, 2001.

SIAM J. Appl. Math.
Vol. 68, No. 4, pp. 1097–1121

THE FACTORIZATION METHOD APPLIED TO THE COMPLETE ELECTRODE MODEL OF IMPEDANCE TOMOGRAPHY*

ARMIN LECHLEITER[†], NUUTTI HYVÖNEN[‡], and HARRI HAKULA[‡]

Abstract. The factorization method is a tool for recovering inclusions inside a body when the Neumann-to-Dirichlet operator, which maps applied currents to measured voltages, is known. In practice this information is never at hand due to the discreteness and physical properties of the measurement devices. The complete electrode model of impedance tomography includes these physical characteristics but leads to a finite-dimensional data set, called the resistivity matrix. The main result of this work is an approximation link relating the resistivity matrix to the Neumann-to-Dirichlet operator in the L^2-operator norm. This result allows us to extend the factorization method to the framework of real-life electrode measurements using a regularized series criterion which is easy to implement in practice. The truncation index of the sequence criterion, which represents the stopping index of the regularization scheme, can be computed solely from the measured, perturbed, and finite-dimensional data. The functionality of the method is demonstrated through numerical experiments.

Key words. impedance tomography, complete electrode model, finite-dimensional approximation, boundary elements, factorization method, perturbation theory

AMS subject classifications. 35R30, 35R25, 47A55, 35Q60, 35J25

DOI. 10.1137/070683295

1. Introduction.

Electrical impedance tomography (EIT) is an imaging technique which has attracted a vast amount of research during the last 20 years, starting with the fundamental paper of Calderón [7]. The nonlinearity and ill-posedness of the problem make it particularly challenging to tackle. The factorization method is one of the rare tools for the inverse EIT problem that is both theoretically founded and suitable for practical application. It stems from the linear sampling method for inverse scattering problems (see Colton and Kirsch [9] and Kirsch [19, 20]) and has been applied to impedance tomography by Brühl and Hanke [3, 5, 4, 14] and by Hyvönen [16, 15]. For an extension to more general inverse elliptic problems we refer the reader to Kirsch [21] and Gebauer [11].

The impedance tomography problem consists of reconstruction of the admittance tensor γ in the elliptic equation

$$\nabla.(\gamma\nabla u) = 0 \quad \text{in } \Omega$$

from boundary measurements of the electric potential u and the corresponding current on the boundary of the bounded domain $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3. Mathematically, this information is given by the Neumann-to-Dirichlet operator that maps the used current pattern on the measured potential. Various imaging problems of practical importance

*Received by the editors February 21, 2007; accepted for publication (in revised form) December 4, 2007; published electronically March 19, 2008.

http://www.siam.org/journals/siap/68-4/68329.html

[†]DFG-Research training group 1294 "Analysis, Simulation and Design of Nanotechnological Processes" and Institut für Algebra und Geometrie, University of Karlsruhe, 76128 Karlsruhe, Germany (lechleiter@math.uni-karlsruhe.de).

[‡]Institute of Mathematics, Helsinki University of Technology, FI-02015 HUT, Finland (nuutti.hyvonen@hut.fi, harri.hakula@tkk.fi). The work of N. Hyvönen was supported by the Academy of Finland (project 115013) and the Finnish Funding Agency for Technology and Innovation (project 40084/06).

deal with locating embedded inhomogeneities in an object with known background admittance. In [3] Brühl applied the factorization method to such a situation and provided explicit characterization of the inclusions under suitable conditions on the inhomogeneities and the background admittance. His results have been generalized in [21, 16].

In [3, 21, 16] the physical characteristics of impedance tomography were modeled using the *continuum model* [8]. This is in practice a poor choice because it does not predict experimental results with satisfactory precision. Especially for medical applications the *complete electrode model* is a much better choice [27]. A main feature of this model is the introduction of a contact impedance z on the boundary between the discrete electrodes and the investigated object. From the viewpoint of the factorization method, the complete model has the fundamental drawback that there are only a limited number of feasible linearly independent electrode current and voltage patterns. This results in a finite-dimensional boundary measurement map, called the resistivity matrix, which makes characterization results like those in [3, 21, 16] unattainable.

Within the continuum model, let us denote by Λ the Neumann-to-Dirichlet map corresponding to the admittance contaminated with inhomogeneities and by Λ_0 the map corresponding to the known background admittance. The main result of this paper (see Theorem 7.1) is that $\Lambda - \Lambda_0$ and the corresponding difference of the finite-dimensional resistivity matrices related to the complete model are arbitrarily close to each other if a large enough number of electrodes is used in the measuring process and the gaps between the electrodes are small enough. This closeness is measured in the operator norm of $L^2(\partial\Omega)$. We give explicit expressions for the rate of convergence in terms of the geometry of the measurement configuration. Having obtained norm estimates in $L^2(\partial\Omega)$, we can apply the perturbation analysis from [22] to construct a factorization method for the complete model; see Theorem 8.2. We emphasize that the difficult and technical norm estimates between the two electrode models are necessary to be able to use results from [22]; compare these with Theorem 2.3. The factorization method we obtain is easy to implement and numerical examples in the last section underline the quality of its reconstructions. However, one has to pay the price that the method requires the knowledge of the resistivity matrix with and without inclusions and the contact impedance of the electrodes (which is for simplicity assumed to be constant). Since we are doing asymptotic analysis in the number of electrodes, a sufficiently large number of them is required to obtain reasonable results. However, our numerical examples indicate that already 16 seems to be an adequate number.

An earlier result concerning the factorization method within the complete model was obtained by Hyvönen in [15], where a Tikhonov regularization approach is used to treat the finite-dimensional situation in the limit case when the electrode configuration gets infinitesimally fine. However, [15] does not treat the numerically simpler series criterion of [21], and it proves only pointwise convergence of the resistivity matrices toward the Neumann-to-Dirichlet maps. It is the sharpness of our estimates that permits us to work here with the series criterion in a manner which is attractive from the viewpoint of practical computations. Numerical experiments with the method developed in [15] have recently been presented in [17], and reference [14] also considers the numerics of the factorization method in the framework of certain simplified electrode models and, to some extent, with real-world data. However, [14] lacks the asymptotic analysis as the number of electrodes is increased. Compared to [17], our method has the advantage that the regularization parameter for regularization by spectral cut-off of the series criterion can, in principle, be chosen in a systematic way by looking at the measurement configuration; in our numerical experiments, we use a substantially simplified way of choosing the spectral cut-off, which results in relatively good reconstructions.

We briefly outline the structure of this paper. In the next section we introduce the factorization method within the continuum model of EIT and state a regularization result [22] which is used as a basic ingredient in what follows. Section 3 recalls some basics of boundary triangulations, used in sections 4 and 5 to introduce the complete electrode model as a perturbation of the continuum model. In section 6 we prove two technical lemmas before the main approximation result of the work is announced in section 7. This result permits us to construct a factorization method using a truncated sequence criterion for the complete electrode model in section 8. Numerical experiments in section 9 confirm the validity of our theoretical findings.

2. Factorization method and its regularization.

Let us consider the factorization method in the framework of the continuum model of EIT, i.e., assuming that there is no contact impedance, that the current can be applied, and that the potential can be measured everywhere on the object boundary. We will follow the presentation of [21]; in particular, we will assume, overly cautiously, that all boundaries are of the class C^2 if not stated otherwise.

Consider admittance tensors $\gamma : \overline{\Omega} \to \mathbb{C}^{n\times n}$ of the form

$$\gamma = \begin{cases} \gamma_0 + \gamma_1 & \text{in } \Omega_{\rm c} \Subset \Omega, \\ \gamma_0 & \text{in } \Omega \setminus \overline{\Omega_{\rm c}}, \end{cases} \tag{2.1}$$

where $\Omega \setminus \overline{\Omega_{\rm c}}$ is connected and the known background admittance tensor $\gamma_0 \in C^{2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, is real, symmetric, and uniformly positive definite. The matrix $\gamma_1 \in L^\infty(\Omega_{\rm c})$ is symmetric with strictly positive definite and negative semidefinite real and imaginary parts, respectively,

$$\operatorname{Re}(\xi^*\gamma_1\xi) \geq c_0|\xi|^2 \quad \text{and} \quad \operatorname{Im}(\xi^*\gamma_1\xi) \leq 0 \quad \text{for } \xi \in \mathbb{C}^n \text{ and some } c_0 > 0. \tag{2.2}$$

Here and in what follows, we denote by $\xi^* = \overline{\xi}^\top$ the transpose conjugate of a vector or a matrix. Notice that for alternating currents with the harmonic time dependence $\exp(-i\omega\cdot)$, $\omega \geq 0$, the admittance γ can be considered to be of the form $\sigma - i\omega\varepsilon$, where σ is the electric conductivity, ε is the electric permittivity, and ω is the angular frequency of the applied current; compare with [8]. As a consequence, if σ and ε are real and symmetric and the angular frequency of the input is not zero, the condition (2.2) is equivalent to saying that the conductivity inside the inclusion is strictly higher than in the background (in the sense of positive definiteness) and the permittivity of the inclusion is positive semidefinite. On the other hand, the conditions set on γ_0 mean that in the background the conductivity is positive and the permittivity vanishes. If $\omega = 0$, i.e., direct current is used, ε does not contribute to the measurements, and so the conditions on the permittivity can be ignored. Also take note that the factorization method retains its functionality also if the conductivity drops in $\Omega_{\rm c}$ [6].

The admittance tensor γ gives rise to a Neumann-to-Dirichlet operator Λ, defined between the $L^2(\partial\Omega)$-based Sobolev spaces $H_\diamond^{\pm 1/2}(\partial\Omega)$ with differentiability index $\pm 1/2$; see [24, 21]. Here, $H_\diamond^{\pm 1/2}(\partial\Omega)$ denotes the closed subspace of zero mean functions in $H^{\pm 1/2}(\partial\Omega)$, i.e., $\int_{\partial\Omega} f \, \mathrm{d}S = 0$ for $f \in H_\diamond^{\pm 1/2}(\partial\Omega)$, and the integral is a shorthand notation for the duality pairing between $H_\diamond^{\pm 1/2}(\partial\Omega)$ in the Gelfand triple $H_\diamond^{1/2}(\partial\Omega) \subset L_\diamond^2(\partial\Omega) \subset H_\diamond^{-1/2}(\partial\Omega)$. We define

$$\Lambda : \ H_\diamond^{-1/2}(\partial\Omega) \to H_\diamond^{1/2}(\partial\Omega), \quad f \mapsto u|_{\partial\Omega},$$

where u is the unique (weak) solution of

$$\begin{cases} \nabla.(\gamma\nabla u) = 0 & \text{in } \Omega, \\ \mathcal{B}_\nu u = f & \text{on } \partial\Omega, \end{cases} \tag{2.3}$$

in $H^1_\diamond(\Omega) = \{v \in H^1(\Omega) \mid \int_{\partial\Omega} v\,\mathrm{d}S = 0\}$. Here $\mathcal{B}_\nu u := (\gamma\nabla u).\nu$ denotes the conormal derivative with respect to the exterior unit normal field ν. Analogously, one defines Λ_0 using γ_0 instead of γ in (2.3). Compactness of the embeddings $H^{1/2}_\diamond(\partial\Omega) \hookrightarrow L^2_\diamond(\partial\Omega) \hookrightarrow H^{-1/2}_\diamond(\partial\Omega)$ implies that Λ and Λ_0 are compact when acting on $L^2_\diamond(\partial\Omega)$.

The inverse problem we consider is to locate Ω_c from measurements of current and voltage on the boundary $\partial\Omega$. This information is within the continuum model given by the Neumann-to-Dirichlet operator Λ. The factorization method solves the inverse problem using the spectral data of the "square root" of the difference $\Lambda - \Lambda_0$. In its original formulation the method required Λ to be self-adjoint. However, Grinberg and Kirsch [13, 21] showed that decomposition of a non–self-adjoint operator into its real and imaginary parts allows us to set up factorization methods for some non–self-adjoint Neumann-to-Dirichlet operators which arise, for instance, when alternating current is used [2].

In brief, the factorization method for EIT works as follows. Let $\mathcal{N}$ be the Neumann function of the differential operator $\nabla.(\gamma_0\nabla\cdot)$ in Ω, i.e., $\mathcal{N}$ satisfies the boundary value problem

$$\nabla.(\gamma_0\nabla\mathcal{N}(\cdot,y)) = \delta_y \quad \text{in } \Omega, \quad \mathcal{B}_\nu\mathcal{N}(\cdot,y) = \frac{1}{|\partial\Omega|} \quad \text{on } \partial\Omega,$$

where δ_y is the Dirac distribution at $y \in \Omega$ and the ground level of potential is chosen so that $\mathcal{N}(\cdot,y)$ integrates to zero over $\partial\Omega$. One constructs test functions $\varphi_y \in L^2_\diamond(\partial\Omega)$ for each point $y \in \Omega$ by

$$\varphi_y(x) = \varphi_{y,a}(x) = a^\top\nabla_y\mathcal{N}(x,y), \quad x \in \partial\Omega, \tag{2.4}$$

where $a \in \mathbb{R}^3$ is the dipole moment. Notice that φ_y can be computed without any information on the inclusion Ω_c. Furthermore, we introduce the—physically meaningless—positive self-adjoint operator

$$(\Lambda - \Lambda_0)_\sharp = |\mathrm{Re}\,(\Lambda - \Lambda_0)| + \mathrm{Im}\,(\Lambda - \Lambda_0) \in \mathcal{L}(L^2_\diamond(\partial\Omega)), \tag{2.5}$$

where the absolute value and the real and imaginary parts are defined as in [21].

Theorem 2.1 (characterization of the inclusion). *Let $(\lambda_j, \psi_j)_{j\in\mathbb{N}}$ be an eigensystem of the compact self-adjoint operator $(\Lambda - \Lambda_0)_\sharp$. Then the sequence*

$$M \mapsto \sum_{j=1}^{M} \frac{|\langle\varphi_y, \psi_j\rangle_{L^2}|^2}{\lambda_j} \tag{2.6}$$

is bounded if and only if $y \in \Omega_\mathrm{c}$.

Let us next consider how the above characterization result can be regularized in case we are given only a noisy incomplete version of $\Lambda - \Lambda_0$. We first recall a perturbation theorem concerning the spectrum of a self-adjoint operator [18, Theorem V.4.10, section 4.3, p. 291], where we denote the spectrum of a linear operator A by $\sigma(A)$.

Theorem 2.2 (continuity of the spectrum). *Let A and B be bounded self-adjoint operators on a Hilbert space H. Then* $\operatorname{dist}(\sigma(A),\sigma(B)) \le \|A-B\|$*; that is,*

$$\sup_{\lambda\in\sigma(B)} \operatorname{dist}(\lambda,\sigma(A)) \le \|A-B\| \quad \textit{and} \quad \sup_{\mu\in\sigma(A)} \operatorname{dist}(\mu,\sigma(B)) \le \|A-B\|.$$

Hence, there is some hope that a measured approximation of the Neumann-to-Dirichlet operator yields good approximations for the eigenvalues of Λ. Due to Rellich's lemma [24, Theorem 3.27], Λ and Λ_0 are compact when acting on $L^2_\diamond(\partial\Omega)$, and hence $(\Lambda-\Lambda_0)_\sharp$ is compact as well; in fact, $(\Lambda-\Lambda_0)_\sharp$ is smoothing if $\partial\Omega$ and the admittances are smooth. Therefore, even a small perturbation of the eigenvalues λ_j of $(\Lambda-\Lambda_0)_\sharp$ might completely destroy the behavior of the sequence $M\mapsto\sum_{j=1}^M |\langle\varphi_y,\psi_j\rangle|^2/\lambda_j$, which characterizes $\Omega_\mathrm{c}\Subset\Omega$ in (2.6). Actually, if $\lambda'_j\approx\lambda_j$, the sequence $M\mapsto\sum_{j=1}^M |\langle\varphi_y,\psi_j\rangle|^2/\lambda'_j$ does not even need to be well defined. Here enters the ill-posedness of the problem. It is quite natural to use a spectral cut-off as a regularization technique or, equivalently, truncate the perturbed Picard series in (2.6). A perturbation analysis of the problem has to deal not only with the perturbed eigenvalues of $(\Lambda-\Lambda_0)_\sharp$ but also with the perturbed eigenvectors which enter the series criterion in (2.6), but above we have neglected this question for simplicity. Interested readers are referred to [22].

For regularization of the Picard series (2.6) in the case of perturbed data, we define a truncation index which depends on two arbitrary, but fixed, parameters $\omega\in(0,1/2)$ and $C>0$. Assume that we have been given the operators $B_M, B_{0M}\in\mathcal{L}(L^2_\diamond(\partial\Omega))$ with the property

$$\|(\Lambda-\Lambda_0)_\sharp-(B_M-B_{0M})_\sharp\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))}=\varepsilon_M\to 0 \quad \text{as } M\to\infty. \tag{2.7}$$

For an eigenvalue $\lambda_j^{(M)}$ of the self-adjoint operator $(B_M-B_{0M})_\sharp$, we define its cluster to be the set of all eigenvalues closer to $\lambda_j^{(M)}$ than $2\varepsilon_M$:

$$\operatorname{clu}(\lambda_j^{(M)}) := \left\{\lambda_l^{(M)}\in\sigma\left((B_M-B_{0M})_\sharp\right) \,\middle|\, \left|\lambda_j^{(M)}-\lambda_l^{(M)}\right|\le 2\varepsilon_M\right\}.$$

Roughly speaking, $\operatorname{clu}(\lambda_j^{(M)})$ contains the eigenvalues of $(B_M-B_{0M})_\sharp$ that might converge to $\lim_{M\to\infty}\lambda_j^{(M)}$ as $M\to\infty$. Now we can define the cut-off index as

$$R(M) := \max\{k\in\mathbb{N} \mid \lambda_k^{(M)}\ge 3\varepsilon_M^\omega,\ \rho_k^{(M)}\ge 8\varepsilon_M^\omega,\ k\varepsilon_M^{1-2\omega}\le C\}, \tag{2.8}$$

where $\rho_k^{(M)}=\min\left\{\left|\lambda_k^{(M)}-\lambda_j^{(M)}\right| \,\middle|\, \operatorname{clu}(\lambda_k^{(M)})\ne\operatorname{clu}(\lambda_j^{(M)})\right\}$. The definition of $R(M)$ has no other motivation than cutting off as many terms in the series criterion as necessary to be able to prove the theorem below. For this asymptotic result it is naturally crucial that $R(M)\to\infty$ as $M\to\infty$. Note that the truncation index prevents, for instance, eigenvalues too close to zero to be investigated in the regularized sequence criterion and that (2.8) implies the more complicated condition (35) of [22].

Theorem 2.3. *Let $\omega\in(0,1/2)$ and $C>0$ be arbitrary, but fixed, parameters, and let the compact operators $B_M, B_{0M}\in\mathcal{L}(L^2_\diamond(\partial\Omega))$, $M\in\mathbb{N}$, be the given data and such that* (2.7) *holds. Let $(\lambda_j^{(M)},\psi_j^{(M)})_{j\in\mathbb{N}}$ be an eigensystem of the compact self-adjoint operator $(B_M-B_{0M})_\sharp$. Then the sequence*

$$M\mapsto\sum_{j=1}^{R(M)}\frac{|\langle\varphi_y,\psi_j^{(M)}\rangle|^2}{\lambda_j^{(M)}}, \quad M\in\mathbb{N},$$

is bounded if and only if $y\in\Omega_\mathrm{c}$.

3. Boundary elements and local projections. Approximation of Neumann-to-Dirichlet operators in a finite-dimensional space motivates us to subdivide the boundary $\partial\Omega$ using a mesh and to consider piecewise polynomial functions on the elements. Such techniques are well known in finite element theory. Our approach is slightly different because we need triangulations (more precisely, quadrangulations) of the boundary and thus curved elements or *panels*. The construction of such meshes is standard in boundary element methods. For completeness we recall some basic results following Sauter and Schwab [25, Kapitel 4]. To be able to reference to this text, we need to make the additional assumption that the boundary $\partial\Omega$ is of the class C^∞ and introduce the reference element

$$\hat{Q} = (0,1) \times (0,1) \subset \mathbb{R}^2.$$

Moreover, we denote the Euclidean norm by $|\cdot|$.

Definition 3.1 (triangulation of $\partial\Omega$). *A triangulation $\mathcal{T}$ of $\partial\Omega$ is a subdivision of $\partial\Omega$ into relatively open disjoint elements such that the following hold:*

1. *$\mathcal{T}$ covers $\partial\Omega$, i.e., $\partial\Omega = \bigcup_{T\in\mathcal{T}} \overline{T}$.*
2. *Each element $T \in \mathcal{T}$ is the image of the reference element $\hat{Q}$ under a diffeomorphism χ_T and there exist constants $C_{\min}$ and $C_{\max}$ such that*

$$0 < C_{\min} < \inf_{x\in\hat{Q}} \inf_{v\in\mathbb{S}^1} |D\chi_T(x)v|^2 \leq \sup_{x\in\hat{Q}} \sup_{v\in\mathbb{S}^1} |D\chi_T(x)v|^2 < C_{\max} < \infty,$$

where $D\chi_T$ denotes the Jacobian of χ_T.

3. *For each reference mapping $\chi_T : \hat{Q} \to T$ there exists an affine mapping $\chi_T^{\mathrm{af}} : \mathbb{R}^2 \to \mathbb{R}^3$ and a smooth map $\chi_{\partial\Omega} : \mathbb{R}^3 \to \mathbb{R}^3$ independent of T such that $\chi_T = \chi_{\partial\Omega} \circ \chi_T^{\mathrm{af}}$ and $\chi_{\partial\Omega} : \chi_T^{\mathrm{af}}(\hat{Q}) \to T$ is a diffeomorphism for all $T \in \mathcal{T}$.*

Finally, $\mathcal{T}$ is a regular triangulation if any two elements contact each other either not at all or in exactly one point or on an entire side. The number of elements of the triangulation is called the size of $\mathcal{T}$.

Any initial mesh $\mathcal{T}$ of $\partial\Omega$ can be refined using a subdivision of the reference element if this subdivision is transported to $\partial\Omega$ via the mapping χ_T. Consecutively refined triangulations lead to a family of triangulations.

Definition 3.2 (family of triangulations). *A set $(\mathcal{T}_M)_{M\in\mathbb{N}}$ is called a family of triangulations if each $\mathcal{T}_M$ is a triangulation of $\partial\Omega$ of size M such that the mesh size δ_M tends to zero:*

$$\delta_M := \max_{T\in\mathcal{T}_M} \sup_{s,s'\in T} |s - s'| \to 0 \quad \textit{as } M \to \infty. \tag{3.1}$$

The family of triangulations $(\mathcal{T}_M)_{M\in\mathbb{N}}$ is shape regular if the quotient of the diameter and the in-circle diameter of $T \in \mathcal{T}_M$ is bounded by some constant κ independent of M [25, *Kapitel* 4.1] *and quasiuniform if*

$$\sup_{M\in\mathbb{N}} \left[\frac{\max_{T\in\mathcal{T}_M} \sup_{s,s'\in T} |s - s'|}{\min_{T\in\mathcal{T}_M} \sup_{s,s'\in T} |s - s'|} \right] < \infty.$$

In what follows, all triangulations are assumed to be regular and all families of triangulations shape regular. It appears later on that quasiuniformity of a family of triangulations is a serious hindrance when dealing with the complete electrode model: Our convergence proof of the finite-dimensional complete model toward the infinite-dimensional continuum model requires the technical assumption that the gaps between the electrodes shrink faster than the electrodes and therefore nonquasiuniform meshes arise naturally. The choice of discontinuous elements makes this special feature no drawback; see [12, 25].

With the help of triangulations it is possible to approximate functions on $\partial\Omega$ in finite-dimensional spaces. We consider polynomial spaces on the reference element and transport them onto $\partial\Omega$ by $\chi_T = \chi_{\partial\Omega} \circ \chi_T^{\mathrm{af}}$. Therefore, we define

$$\hat{\mathcal{P}}^k := \operatorname{span}\{x^\mu \mid \mu \in \mathbb{N}_0^2,\ |\mu_1| + |\mu_2| \leq k\}$$

to be the space of polynomials of degree less than or equal to k in two variables.

Definition 3.3. *Let $\mathcal{T}$ be a triangulation of $\partial\Omega$. Then*

$$\mathcal{P}^k := \{\psi : \partial\Omega \to \mathbb{C} \mid \textit{for all } T \in \mathcal{T} : \psi \circ \chi_T \in \hat{\mathcal{P}}^k\}.$$

Observe that, in general, $f \in \mathcal{P}^k$ is a polynomial not on $\partial\Omega$ but on the reference element $\hat{Q}$ after backtransport with χ_T. Moreover, f does not need to be continuous over the edges of the elements (discontinuous elements).

The *local L^2-projection* is now constructed by lifting the orthogonal projection from $L^2(\hat{Q})$ to $\hat{\mathcal{P}}^k$ to the boundary $\partial\Omega$: For $\psi \in L^2(\hat{Q})$ one defines $\hat{P}^k$ via

$$\hat{P}^k\psi \in \hat{\mathcal{P}}^k \quad \text{and} \quad (\hat{P}^k\psi - \psi) \perp v \quad \text{for all } v \in \hat{\mathcal{P}}^k,$$

where the orthogonality is in the sense of the inner product of $L^2(\hat{Q})$. Furthermore, for $\psi \in H^s(\partial\Omega)$, $s \geq 0$, and a triangulation $\mathcal{T}$ of $\partial\Omega$ we set

$$P^k\psi\big|_T := (\hat{P}^k(\psi|_T \circ \chi_T)) \circ \chi_T^{-1}, \quad T \in \mathcal{T}. \tag{3.2}$$

The lifting of $\hat{P}^k$ to the curved boundary makes the L^2-projector P^k, in general, nonorthogonal on $L^2(\partial\Omega)$. The approximation quality of P^k depends in principal on the magnitude of the Sobolev index of the space $H^s(\partial\Omega)$ [25, Satz 4.3.18].

Theorem 3.4. *Let $(\mathcal{T}_M)_{M\in\mathbb{N}}$ be a shape regular family of triangulations of the boundary of the smooth domain $\Omega \subset \mathbb{R}^3$, $s > 0$, and P_M^k the L^2-projection on $\mathcal{T}_M$. Then*

$$\|\psi - P_M^k\psi\|_{L^2(\partial\Omega)} \leq C\delta_M^{\min(k+1,s)}\|\psi\|_{H^s(\partial\Omega)} \qquad \textit{for } \psi \in H^s(\partial\Omega), \tag{3.3}$$

where C depends on the constant of shape regularity κ and on $\partial\Omega$.

In this work we make use only of the case $k = 0$ (piecewise constant interpolation) since this is the relevant case in impedance tomography when dealing with perfectly conducting electrodes. Thus, we drop the index k in what follows.

4. The complete electrode model. In practice, the current is injected through electrodes attached to the surface of the investigated body Ω. The complete electrode model [8, 27] is now the standard model for this procedure. It takes into account the following four properties of the setup.

First, the electrodes are a discrete set denoted by $E_1, \ldots, E_p$. Each E_j is considered to be a relatively open subset of the boundary $\partial\Omega$ with positive surface measure: $|E_j| > 0$. We assume, furthermore, that the electrodes are connected and well separated, i.e., $\operatorname{dist}(E_k, E_j) > 0$ for $k \neq j$. The set $\{E_1, \ldots, E_p\}$ is called an *electrode configuration.*

Second, the net current through E_j equals the total flux through the surface patch underneath the electrode. Let $I_j \in \mathbb{C}$ be the mean current flux applied to E_j, i.e., $I_j|E_j|$ is the net current, and define $I = (I_1, \ldots, I_p)^\top$. It holds that

$$\frac{1}{|E_j|}\int_{E_j} \mathcal{B}_\nu u \, \mathrm{d}S = I_j \quad \text{for } j = 1, \ldots, p,$$

where $\mathcal{B}_\nu$ is defined as in (2.3). Due to the principle of conservation of charge, we require that $\sum_j I_j\,|E_j| = 0$. The vector I is called a *current pattern* or a *current vector*. For convenience, we denote the space of current patterns of length p by

$$\mathbb{C}^p_E = \left\{ I \in \mathbb{C}^p \;\Big|\; \sum\nolimits_{j=1}^p I_j\,|E_j| = 0 \right\}, \tag{4.1}$$

with the weighted norm $|\cdot|_E$ defined through

$$|I|_E^2 := \sum_{j=1}^p |E_j|\,|I_j|^2\,. \tag{4.2}$$

We remark that $\mathbb{C}^p_E$ can be identified as a subspace of $L^2_\diamond(\partial\Omega)$ via $C^p_E \ni I \mapsto f$, where $f(x) = I_j$ on E_j, $j = 1, \ldots, p$, and 0 elsewhere.

Third, we model the electrodes as perfect conductors; that is, we assume that the potential along an electrode is constant. This is the so-called shunting effect. The set of electrode voltages is denoted by $U = (U_1, \ldots, U_p)^\top$ and assumed to belong to $\mathbb{C}^p_E$. This condition can be seen as a grounding of potential.

Fourth, the complete electrode model includes the effect of contact impedance at the electrodes: When EIT is used in a medical context, a thin layer with high resistivity is formed at the boundary between the electrodes and the skin due to dermal moisture. We incorporate this effect by introducing the surface impedance function $z \in C^\infty(\partial\Omega)$, which denotes the resistivity of the contact layer at the boundary. The real part of z is assumed to be positive. According to Ohm's law, the potential u at E_j drops by $z\mathcal{B}_\nu u|_{E_j}$ over the contact layer.

The complete electrode model gives rise to the following (weak) formulation of the forward problem: Given a current vector $I = (I_1, \ldots, I_p)^\top \in \mathbb{C}^p_E$, an admittance tensor γ, and a contact impedance z, find the potential $u \in H^1(\Omega)$ and the set of electrode voltages $U \in \mathbb{C}^p_E$ that satisfy

$$\nabla.\big(\gamma\nabla u\big) = 0 \quad \text{in } \Omega, \tag{4.3}$$

$$u + z\mathcal{B}_\nu u = U_j \quad \text{on } E_j \quad \text{for } j = 1, \ldots, p, \tag{4.4}$$

$$\frac{1}{|E_j|}\int_{E_j} \mathcal{B}_\nu u \, \mathrm{d}S = I_j \quad \text{for } j = 1, \ldots, p, \tag{4.5}$$

$$\mathcal{B}_\nu u = 0 \quad \text{on } \partial\Omega \setminus \cup_{j=1}^p \overline{E}_j. \tag{4.6}$$

Notice that without the grounding of potential, i.e., the condition $U \in \mathbb{C}^p_E$, the above problem would not have a unique solution. According to [27], the accuracy of this model corresponds to the measurement precision of the physical experiment.

The measurement map, i.e., the resistivity matrix, associated to the complete electrode model is given by

$$\Sigma : \mathbb{C}^p_E \to \mathbb{C}^p_E, \quad I \mapsto U,$$

where U is the second part of the solution to (4.3)–(4.6) corresponding to the current pattern I. The resistivity matrix corresponding to the background admittance is defined by this very same equation when γ is replaced by γ_0 in (4.3).

The existence and uniqueness of the solution $(u, U) \in H^1(\Omega) \oplus \mathbb{C}^p_E$ can be shown using the Lax–Milgram lemma. One starts to look for the solution in the quotient space $(H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C}$ and chooses afterward the unique representative in $H^1(\Omega) \oplus \mathbb{C}^p_E$. In [27, 15] it is shown that $(u, U) \in (H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C}$ satisfies (4.3)–(4.6) if and only if

$$b((u,U),(v,V)) = f(v,V) \quad \text{for all } (v,V) \in (H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C}, \tag{4.7}$$

where the elliptic sesquilinear form $b = b_{\gamma,z}$ is defined by

$$b((u,U),(v,V)) := \int_\Omega \nabla v^* \gamma \nabla u \,\mathrm{d}x + \sum_{j=1}^p \int_{E_j} \frac{1}{z}(u - U_j)(\overline{v} - \overline{V}_j)\,\mathrm{d}S,$$

and

$$f(v,V) := \sum_{j=1}^p |E_j| I_j \overline{V}_j.$$

As in [15], we use the inner product

$$\langle (u,U),(v,V)\rangle_* = \int_\Omega \nabla v^* \nabla u \,\mathrm{d}x + \sum_{j=1}^p \int_{E_j} (U - u)(\overline{V} - \overline{v})\,\mathrm{d}S$$

on $(H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C}$. The associated norm is equivalent to the quotient norm

$$\|(u,U)\|^2_{(H^1(\Omega)\oplus\mathbb{C}^p)/\mathbb{C}} := \inf_{c\in\mathbb{C}} \left\{ \|u + c\|^2_{H^1(\Omega)} + |U + c|^2_E \right\},$$

independently of the number and size of the electrodes [15, Lemma 2.5, Corollary 2.6]. It follows from the conditions set on γ and z that there is a constant of ellipticity $c = c(z,\gamma) > 0$ such that

$$\operatorname{Re} b((u,U),(u,U)) \geq c\|(u,U)\|^2_* \quad \text{for all } (u,U) \in (H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C} \tag{4.8}$$

and for any electrode configuration $\{E_1, \ldots, E_p\}$, $p \in \mathbb{N}$ [15, Corollary 2.6]. Moreover, there exists C independent of the geometry of the electrodes such that

(4.9)
$$|b((u,U),(v,V))| \leq C\|(u,U)\|_*\|(v,V)\|_* \quad \text{for all } (u,U),(v,V) \in (H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C}.$$

Since the functional $f : (H^1(\Omega) \oplus \mathbb{C}^p)/\mathbb{C} \to \mathbb{C}$ is well defined, continuous, and antilinear, the existence and uniqueness of the solution to (4.3)–(4.6) follow now by combining (4.8) and (4.9) with the Lax–Milgram lemma.

5. Electrode configurations and discretization. Let $\mathcal{T} = (T_j)$ be a triangulation of $\partial\Omega$, as introduced in section 3. We choose a subset $\mathcal{E} = (E_j)_{j=1}^p \subset \mathcal{T}$ that we call the electrodes and write $E = \cup_{j=1}^p E_j$. The complementing elements $\mathcal{G} := \mathcal{T} \smallsetminus \mathcal{E}$, $G := \cup_{G_j \notin \mathcal{E}} G_j$, play the role of the gaps between the electrodes. For simplicity, we denote

$$h_T = \sup_{s,s'\in T} |s - s'|, \quad h_\mathcal{E} = \max_{T\in\mathcal{E}} h_T, \quad h_\mathcal{G} = \max_{T\in\mathcal{G}} h_T, \quad h_\mathcal{T} = \max_{T\in\mathcal{T}} h_T.$$

Motivated by the EIT experiment, we assume that between any two electrodes there is a gap (the electrodes are separated)

$$\operatorname{dist}(E_k, E_j) > 0 \quad \text{for all } j, k = 1, \dots, p,\ j \neq k.$$

The space of piecewise polynomials of degree 0 on the electrodes is denoted by $\mathcal{P}^E$,

$$\mathcal{P}^E = \{u : \partial\Omega \to \mathbb{C} \,|\, \text{for all } E_j : u|_{E_j} \equiv \text{const. and } u|_G \equiv 0\},$$

and we write P^E for the local L^2-projector on this space; i.e., P^E acts on the electrodes like a usual L^2-projector (3.2) but vanishes on the gaps. In the same way we define the space $\mathcal{P}^G$ of piecewise constant functions vanishing on the electrodes and associate the projector P^G. Observe that $P^{E+G} := P^E + P^G = P^0_{\mathcal{T}}$ is the piecewise constant L^2-projector on $\mathcal{T}$. In particular, Theorem 3.4 holds for P^{E+G}.

Since $\mathcal{P}^E$ is finite-dimensional, the following vector notation is sometimes useful:

$$\psi = (\psi_j)_{j=1}^p, \quad \psi_j := \psi|_{E_j} \quad \text{for } \psi \in \mathcal{P}^E.$$

That is, we identify a function in $\mathcal{P}^E$ with the associated vector of function values on the electrodes. Likewise, we identify functions in $\mathcal{P}^G$ and $\mathcal{P}^{E+G}$ with their coordinate representation. The $L^2(\partial\Omega)$ norm on $\mathcal{P}^E$ can then be written as

$$\|\psi\|^2_{L^2(\partial\Omega)} = \|\psi\|^2_{L^2(E)} = \sum_{j=1}^p |\psi_j|^2 |E_j| = \left|(\psi_j)_{j=1}^p\right|^2_E \quad \text{for } \psi \in \mathcal{P}^E.$$

Let now $f \in L^2_\diamond(\partial\Omega)$ be the Neumann boundary data for the continuum model problem (2.3). We try to approximate the solution of (2.3) by the solution of the complete electrode model forward problem. Since $P^E f$ does not, in general, belong to $\mathbb{C}^p_E$, it cannot be used as an input of (4.5), and so we are forced to define yet another projector:

$$\tilde{P}^E f = P^E f - K_f \quad \text{for } f \in L^2(\partial\Omega), \tag{5.1}$$

where

$$K_f = \frac{\sum_j (P^E f)_j |E_j|}{\sum_j |E_j|}. \tag{5.2}$$

It is easy to check that $\tilde{P}^E$ maps $L^2(\partial\Omega)$ to $\mathbb{C}^p_E$. In the following two sections we will investigate how well the mapping

$$(\Sigma - \Sigma_0)\tilde{P}^E : L^2_\diamond(\partial\Omega) \to \mathbb{C}^p_E \subset L^2_\diamond(\partial\Omega)$$

approximates the difference $\Lambda - \Lambda_0$ in the L^2-operator norm.

The projection operator $\frac{1}{|E_j|}\int_{E_j}(\cdot)\,\mathrm{d}S$ which appears in the formulation of the complete model (4.5), is abbreviated to $⨍_E(\cdot)$, i.e.,

$$\left[⨍_E f\right](s) = \begin{cases} \frac{1}{|E_j|}\int_{E_j} f\,\mathrm{d}S & \text{if } s \in E_j \in \mathcal{E}, \\ 0 & \text{else,} \end{cases} \qquad \text{for } f \in L^2(\partial\Omega). \tag{5.3}$$

This projection is orthogonal in $L^2(\partial\Omega)$ and it plays a crucial role in our analysis, as does its counterpart on the gaps, namely,

$$\left[⨍_G f\right](s) = \begin{cases} \frac{1}{|G_j|}\int_{G_j} f\,\mathrm{d}S & \text{if } s \in G_j \in \mathcal{G}, \\ 0 & \text{else,} \end{cases} \qquad \text{for } f \in L^2(\partial\Omega).$$

6. Two technical lemmas on projection operators. The aim of this work is to prove that $\Sigma - \Sigma_0$ is an approximation of $\Lambda - \Lambda_0$. The key for this perturbation lemma turns out to be the comparison of the two projection operators P^E and $\fint(\cdot)$ in $L^2(\partial\Omega)$. The two projectors are close to each other if the mesh size is small, as the following technical lemma shows. This statement is by no means surprising since for a polyhedral surface the projectors agree.

Lemma 6.1. *Let Ω be of class C^∞ and $\mathcal{T} = \mathcal{E} \cup \mathcal{G}$ a triangulation of $\partial\Omega$. Then there exists a constant $C(\Omega)$ such that*

$$\left\| \fint_E u - P^E u \right\|_{L^2(E)} \leq C(\Omega)\, h_{\mathcal{E}}^2 \|u\|_{L^2(\partial\Omega)} \quad \textit{for } u \in L^2(\partial\Omega).$$

For simplicity we restrict ourselves in this lemma to smooth domains, although the proof shows that a domain of class C^2 is sufficient.

Proof. 1. The projector P^E has been defined as the L^2-projector on the reference element $\hat{Q} = (0,1)^2$, transported onto the electrodes E_j. Since P^E projects onto constant functions, for $f \in L^2(\partial\Omega)$, $P^E f$ has the form

$$P^E f\big|_{E_j} = \frac{1}{|\hat{Q}|}\int_{\hat{Q}} f \circ \chi_{E_j}\,\mathrm{d}x = \int_{\hat{Q}} f \circ \chi_{E_j}\,\mathrm{d}x, \quad j = 1, \ldots, p,$$

and $P^E f \equiv 0$ between the electrodes. On the other hand, the transformation theorem shows that

$$\fint_E u\bigg|_{E_j} = \frac{1}{|E_j|}\int_{\hat{Q}} f \circ \chi_{E_j} \sqrt{\det\Big(\left(D\chi_{E_j}\right)^* D\chi_{E_j}\Big)}\,\mathrm{d}x, \quad j = 1, \ldots, p.$$

Hence, we have to estimate an expression of the form

$$\frac{1}{|E_j|}\int_{\hat{Q}} f \circ \chi_{E_j} \sqrt{\det\Big(\left(D\chi_{E_j}\right)^* D\chi_{E_j}\Big)}\,\mathrm{d}x - \int_{\hat{Q}} f \circ \chi_{E_j}\,\mathrm{d}x.$$

Our strategy is to exploit the smoothness of $\partial\Omega$.

2. Let us introduce auxiliary quadrilaterals $\hat{E}_j \subset \mathbb{R}^3$ such that $\hat{E}_j$ touches E_j in some point $\xi_j \subset E_j$; i.e., $\hat{E}_j$ lies in the tangential hyperplane $T_{\xi_j}(\partial\Omega)$ (compare with Figure 6.1). In addition, the four corner points of $\hat{E}_j$ are orthogonal projections of the four corner points of E_j onto $T_{\xi_j}(\partial\Omega)$. Without loss of generality we can assume (using a rigid motion) that $T_{\xi_j}(\partial\Omega) = \{x_1 = x_2 = 0\}$ and denote the diameter of $\hat{E}_j$ by

$$h_{\hat{E}_j} = \sup_{\xi,\xi' \in \hat{E}_j} |\xi - \xi'|\,. \tag{6.1}$$

By choosing $h_{\mathcal{E}}$ small enough we can assume that there exists a C^∞ function $\zeta_j : \mathbb{R}^2 \to \mathbb{R}$ such that ζ_j "transports $\hat{E}_j$ onto E_j"; i.e.,

$$\hat{\chi}_{E_j} : \hat{E}_j \ni \xi \mapsto (\xi, \zeta_j(\xi)) \in \mathbb{R}^3 \tag{6.2}$$

is a C^∞ diffeomorphism that maps $\hat{E}_j$ onto E_j. By the compactness of $\partial\Omega$, our smoothness assumption on $\partial\Omega$ implies that there exists $C(\Omega)$ such that

$$\|\hat{\chi}_{E_j}\|_{C^2(\hat{E}_j)} \leq C(\Omega) \quad \text{and} \quad \|\hat{\chi}_{E_j}^{-1}\|_{C^2(E_j)} \leq C(\Omega) \quad \text{for all } j = 1, \ldots, p. \tag{6.3}$$

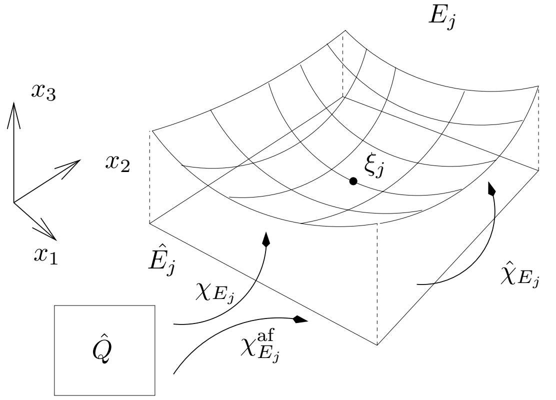


FIG. 6.1. *The quadrilaterals E_j and $\hat{E}_j$. E_j is a C^∞ surface in $\mathbb{R}^3$ with diameter h_{E_j} and $\hat{E}_j \subset \mathbb{R}^2$ with diameter $h_{\hat{E}_j}$. The diffeomorphism $\hat{\chi}_{E_j}$ transports $\hat{E}_j$ onto E_j. The reference element $\hat{Q}$ is transported by χ_{E_j} onto E_j and by $\chi_{E_j}^{\mathrm{af}}$ onto $\hat{E}_j$. We have $\chi_{E_j} = \hat{\chi}_{E_j} \circ \chi_{E_j}^{\mathrm{af}}$.*

Finally, we denote by $\chi_{E_j}^{\mathrm{af}}$ the affine mapping that transports $\hat{Q}$ onto $\hat{E}_j$.

3. As a side computation, we estimate the difference of the surface measures of E_j and $\hat{E}_j$. By the transformation theorem,

$$\begin{aligned}\left||E_j| - |\hat{E}_j|\right| &= \left|\int_{E_j} \mathrm{d}S - \int_{\hat{E}_j} \mathrm{d}\xi\right| \\ &\leq \int_{\hat{E}_j} \left|\sqrt{\det\left(\left(D\hat{\chi}_{E_j}\right)^* D\hat{\chi}_{E_j}\right)} - 1\right| \mathrm{d}\xi \\ &= \int_{\hat{E}_j} \left|\sqrt{1 + |\nabla\zeta_j|^2} - 1\right| \mathrm{d}\xi = \int_{\hat{E}_j} \sqrt{1 + |\nabla\zeta_j|^2} - 1 \, \mathrm{d}\xi.\end{aligned}$$

Recall that we constructed $\hat{E}_j$ so that $\nabla\zeta_j(\xi_j) = 0$ since $\hat{E}_j$ is tangential to E_j at ξ_j. Hence, by the mean value theorem,

$$|\nabla\zeta_j(\xi)| = |\nabla\zeta_j(\xi) - \nabla\zeta_j(\xi_j)| \leq \|\hat{\chi}_{E_j}\|_{C^2(\Omega)} \sup_{\xi \in \hat{E}_j} |\xi - \xi_j| \leq C(\Omega) h_{\hat{E}_j}, \tag{6.4}$$

where $C(\Omega)$ was defined in (6.3) and $h_{\hat{E}_j}$ in (6.1). As a consequence, the mean value theorem for $\sqrt{\cdot}$ implies that

$$\sqrt{1 + |\nabla\zeta_j|^2} - 1 \leq \frac{1}{2}|\nabla\zeta_j|^2 \leq C(\Omega) h_{\hat{E}_j}^2$$

and, in particular, that

$$\int_{\hat{E}_j} \sqrt{1 + |\nabla\zeta_j|^2} - 1 \, \mathrm{d}\xi \leq C(\Omega) h_{\hat{E}_j}^2 \left|\hat{E}_j\right|. \tag{6.5}$$

4. The preceding side computation enables us to estimate as follows:

$$\begin{aligned}
&\left|\frac{1}{|E_j|}\int_{E_j} f\,\mathrm{d}S-(P^E f)_j\right| = \left|\frac{1}{|E_j|}\int_{\hat{E}_j}\left(f\circ\hat{\chi}_{E_j}\right)\sqrt{\det\left(\left(D\hat{\chi}_{E_j}\right)^* D\hat{\chi}_{E_j}\right)}\,\mathrm{d}\xi-(P^E f)_j\right|\\
&=\left|\frac{1}{|E_j|}\int_{\hat{E}_j}\left(f\circ\hat{\chi}_{E_j}\right)\sqrt{1+|\nabla\zeta_j|^2}\,\mathrm{d}\xi-\frac{1}{|\hat{E}_j|}\int_{\hat{E}_j} f\circ\hat{\chi}_{E_j}\,\mathrm{d}\xi\right|\\
&\le\frac{1}{|E_j|}\int_{\hat{E}_j}\left|f\circ\hat{\chi}_{E_j}\right|\left(\sqrt{1+|\nabla\zeta_j|^2}-1\right)\mathrm{d}\xi+\frac{\left||\hat{E}_j|-|E_j|\right|}{|\hat{E}_j|\,|E_j|}\int_{\hat{E}_j}\left|f\circ\hat{\chi}_{E_j}\right|\mathrm{d}\xi\\
&\le\frac{1}{|E_j|}\int_{\hat{E}_j}\left|f\circ\hat{\chi}_{E_j}\right|\left(C(\Omega)h_{\hat{E}_j}^2\right)\mathrm{d}\xi+C(\Omega)\frac{h_{\hat{E}_j}^2}{|E_j|}\int_{\hat{E}_j}\left|f\circ\hat{\chi}_{E_j}\right|\mathrm{d}\xi\\
&\le\frac{C(\Omega)}{|E_j|}h_{\hat{E}_j}^2\sqrt{|\hat{E}_j|}\,\|f\circ\hat{\chi}_{E_j}\|_{L^2(\hat{E}_j)},
\end{aligned}$$

where we used the Cauchy–Schwarz inequality. Note that

$$\|f\circ\hat{\chi}_{E_j}\|_{L^2(\hat{E}_j)}\le C(\Omega)\|f\|_{L^2(E_j)},\qquad h_{\hat{E}_j}\le C(\Omega)h_{E_j},\quad\text{and}\quad|\hat{E}_j|\le|E_j|,$$

where $C(\Omega)$ depends on the curvature of $\partial\Omega$.

5. We are ready for the final estimate of this proof. The preceding parts and the Cauchy–Schwarz inequality allow us to estimate as follows:

$$\begin{aligned}
\left\|\fint_E f\,\mathrm{d}S-P^E f\right\|_{L^2(E)}^2&=\sum_{j=1}^{p}\left\|\frac{1}{|E_j|}\int_{E_j} f\,\mathrm{d}S-(P^E f)_j\right\|_{L^2(E_j)}^2\\
&=\sum_{j=1}^{p}\left|\frac{1}{|E_j|}\int_{E_j} f\,\mathrm{d}S-(P^E f)_j\right|^2|E_j|\\
&\le C(\Omega)\sum_{j=1}^{p}\left[\frac{1}{|E_j|}h_{\hat{E}_j}^2\sqrt{|\hat{E}_j|}\|f\|_{L^2(E_j)}\right]^2|E_j|\\
&\le C(\Omega)\sum_{j=1}^{p}h_{\hat{E}_j}^4\|f\|_{L^2(E_j)}^2\le C(\Omega)h_{\mathcal{E}}^4\|f\|_{L^2(E)}^2.
\end{aligned}\tag{6.6}$$

□

By using the above lemma, we can extend the result of Theorem 3.4.

Corollary 6.2. *Suppose that the assumptions of Theorem* 3.4 *hold. Then*

$$\|\psi-P_M\psi\|_{H^{-1/2}(\partial\Omega)}\le C\delta_M^{1/2}\|\psi\|_{L^2(\partial\Omega)}\qquad\textit{for }\psi\in L^2(\partial\Omega),$$

where C depends on the constant of shape regularity κ and on $\partial\Omega$ and $P_M=P_M^0$.

Proof. Let us introduce the L^2-orthogonal projectors

$$\fint_{\mathcal{T}_M}:L^2(\partial\Omega)\to\mathcal{P}_M\subset L^2(\partial\Omega),\quad M\in\mathbb{N},$$

that are defined in accordance with (5.3) but on the whole triangulations $\mathcal{T}_M$, $M\in\mathbb{N}$, respectively. Here $\mathcal{P}_M$ denotes the space of piecewise constant functions on the triangulation $\mathcal{T}_M$. In the rest of this proof, we will denote by $\psi_M^0\in\mathcal{P}_M$ the image of $\psi\in L^2(\partial\Omega)$ under the Mth of the above-defined projections.

Application of Lemma 6.1 to the whole triangulation $\mathcal{T}_M$ instead of the subset $\mathcal{E}$ shows that

$$\left|\left\langle \psi_M^0 - P_M\psi, \phi\right\rangle_{L^2(\partial\Omega)}\right| \leq C\delta_M^2 \|\psi\|_{L^2(\partial\Omega)}\|\phi\|_{L^2(\partial\Omega)} \quad \text{for } \psi, \phi \in L^2(\partial\Omega),$$

where we used the Cauchy–Schwarz inequality and the notation of Theorem 3.4. In particular, it holds that

$$\begin{aligned}\left|\langle \psi, P_M\phi\rangle_{L^2(\partial\Omega)} - \langle P_M\psi, \phi\rangle_{L^2(\partial\Omega)}\right| &= \left|\left\langle \psi, P_M\phi - \phi_M^0\right\rangle_{L^2(\partial\Omega)} - \left\langle P_M\psi - \psi_M^0, \phi\right\rangle_{L^2(\partial\Omega)}\right| \\ &\leq C\delta_M^2 \|\psi\|_{L^2(\partial\Omega)}\|\phi\|_{L^2(\partial\Omega)}.\end{aligned}$$

As a consequence, it follows from the triangle and Cauchy–Schwarz inequalities and Theorem 3.4 that

$$\begin{aligned}\left|\langle \psi - P_M\psi, \phi\rangle_{L^2(\partial\Omega)}\right| &\leq \left|\langle \psi, \phi - P_M\phi\rangle_{L^2(\partial\Omega)}\right| + \left|\langle \psi, P_M\phi\rangle_{L^2(\partial\Omega)} - \langle P_M\psi, \phi\rangle_{L^2(\partial\Omega)}\right| \\ &\leq C\delta_M^{1/2}\|\psi\|_{L^2(\partial\Omega)}\|\phi\|_{H^{1/2}(\partial\Omega)}\end{aligned}$$

for all $\psi \in L^2(\partial\Omega)$ and $\phi \in H^{1/2}(\partial\Omega)$. Now we can argue by duality as follows:

$$\|\psi - P_M\psi\|_{H^{-1/2}(\partial\Omega)} \leq \sup_{\phi\in H^{1/2}(\partial\Omega),\phi\neq 0} \frac{\left|\langle \psi - P_M\psi, \phi\rangle_{L^2(\partial\Omega)}\right|}{\|\phi\|_{H^{1/2}(\partial\Omega)}} \leq C\delta_M^{1/2}\|\psi\|_{L^2(\partial\Omega)}$$

for all $\psi \in L^2(\partial\Omega)$. This completes the proof. □

When we discretized the boundary current $f \in L^2_\diamond(\partial\Omega)$ of the continuum model to obtain input data for the discrete complete model, we needed to introduce a constant K_f; see (5.2). This constant appeared since for the zero mean current $f \in L^2_\diamond(\partial\Omega)$ the projection $P^E f$ fails to possess zero mean value in general. The following lemma shows that K_f is small for thin gaps and small electrodes.

LEMMA 6.3. *Let $f \in L^2_\diamond(\partial\Omega)$, suppose that the assumptions of Lemma* 6.1 *hold, and define K_f as in* (5.2). *Then there exists $C(\Omega)$ such that*

$$|K_f| \leq \left[|E|^{-1}|G|^{1/2} + C(\Omega)|E|^{-1/2}h_{\mathcal{E}}^2\right]\|f\|_{L^2(\partial\Omega)}. \tag{6.7}$$

Since $|E| + |G| = |\partial\Omega|$ we observe that if the gaps are small, i.e., $|G|$ is small, then $|E| \approx |\partial\Omega|$. If, moreover, the electrodes are thin, i.e., $h_{\mathcal{E}}$ is small, then K_f is small due to (6.7).

Proof. 1. Using techniques similar to those in the proof of Lemma 6.1, one computes that

$$\begin{aligned}\left|\int_E P^E f - f \,\mathrm{d}S\right| &= \left|\sum_{j=1}^{p} |E_j| \left[(P^E f)_j - \frac{1}{|E_j|}\int_{E_j} f \,\mathrm{d}S\right]\right| \\ &\leq \sqrt{|E|}\left\|P^E f - \fint_E f\right\|_{L^2(E)} \\ &\overset{(6.6)}{\leq} C(\Omega)\sqrt{|E|}h_{\mathcal{E}}^2\|f\|_{L^2(\partial\Omega)}.\end{aligned}$$

2. The definition of K_f and part 1 of the proof imply that

$$\begin{aligned}
|K_f| &= \left| \frac{\sum_j (P^E f)_j |E_j|}{\sum_j |E_j|} \right| = \frac{1}{|E|} \left| \int_E P^E f \, \mathrm{d}S \right| \\
&\le \frac{1}{|E|} \left| \int_E f \, \mathrm{d}S \right| + \frac{1}{|E|} \left| \int_E P^E f - f \, \mathrm{d}S \right| \\
&\le \frac{1}{|E|} \left| \int_G f \, \mathrm{d}S \right| + C(\Omega)|E|^{-1/2} h_{\mathcal{E}}^2 \|f\|_{L^2(\partial\Omega)} \\
&\le \left[|E|^{-1} |G|^{1/2} + C(\Omega)|E|^{-1/2} h_{\mathcal{E}}^2 \right] \|f\|_{L^2(\partial\Omega)},
\end{aligned}$$

where we used the zero mean property of f in the second to last intermediate phase. □

7. Approximation of the continuum model by the complete electrode model. We have now collected all tools we require to prove our main theorem. It shows that the complete model approximates the continuum model in the operator norm if the electrodes covering $\partial\Omega$ are fine enough. Moreover, the surface area of the gaps needs to be small. All required geometric assumptions appear quite explicitly in the estimates. Concerning notation, recall that Λ and Λ_0 denote the Neumann-to-Dirichlet operators associated to the admittances γ and γ_0, respectively, whereas Σ and Σ_0 denote the corresponding resistivity matrices. The projection $\tilde{P}^E$ is defined by (5.1). For simplicity, we assume that the contact impedance z is just a complex constant with positive real part. This technical assumption is used to ease the proof of Lemma 7.3.

THEOREM 7.1. *Let $\mathcal{T}$ be a triangulation of the boundary of the smooth bounded domain Ω. We denote as usual by $h_{\mathcal{E}} = \max_j h_{E_j}$, $h_{\mathcal{G}} = \max_j h_{G_j}$, and $h_{\mathcal{T}} = \max(h_{\mathcal{E}}, h_{\mathcal{G}})$ the element size of the electrodes, the gaps, and the triangulation, respectively. Then*

$$\begin{aligned}
&\|(\Lambda - \Lambda_0) - (\Sigma - \Sigma_0)\, \tilde{P}^E\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} \\
&\qquad \le C(\Omega, \gamma, \gamma_0, \kappa_{\mathcal{T}}, z) \left[h_{\mathcal{T}}^{1/2} + \left(\sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1} \right)^{1/2} + |E|^{-1/2} h_{\mathcal{T}}^2 + |E|^{-1} |G|^{1/2} \right]
\end{aligned} \tag{7.1}$$

for $\theta^* \in (4/3, 2)$.

It is crucial that the constant in the above estimate does not depend on the electrode configuration.

Proof. 1. Let us denote by $u \in H^1_\diamond(\Omega)$ the unique solution of the Neumann problem (2.3) for $f \in L^2_\diamond(\partial\Omega)$ and by $(\tilde{u}, U) \in H^1(\Omega) \oplus \mathbb{C}^p_E$ the unique solution to the complete model forward problem (4.3)–(4.6) with the input current $\tilde{P}^E f = P^E f - K_f$; see (5.1). Analogously, we set u_0 and $\tilde{u}_0$ to be the solutions of the same problems for the admittance γ_0 and introduce the constant $c = \int_{\partial\Omega} \tilde{u} \, \mathrm{d}S$ such that $\tilde{u} - c \in H^1_\diamond(\Omega)$. We first observe that there exists a constant $C(\Omega, \gamma)$ such that

$$\begin{aligned}
\|u - \tilde{u} + c\|_{L^2(\partial\Omega)} &\le C(\Omega)\|u - \tilde{u} + c\|_{H^{1/2}(\partial\Omega)} \le C(\Omega)\|u - \tilde{u} + c\|_{H^1(\Omega)} \\
&\le C(\Omega, \gamma)\|\mathcal{B}_\nu u - \mathcal{B}_\nu \tilde{u}\|_{H^{-1/2}(\partial\Omega)}.
\end{aligned} \tag{7.2}$$

This inequality relies on well-posedness of the Neumann problem (2.3) in $H^1_\diamond(\Omega)$: The difference $u - \tilde{u} + c$ solves (2.3) with the Neumann boundary values $\mathcal{B}_\nu u - \mathcal{B}_\nu \tilde{u} \in H^{-1/2}(\partial\Omega)$. Hence, (7.2) follows from the trace theorem and the continuous embedding $H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega)$. An analogous inequality holds of course for $u_0 - \tilde{u}_0$.

Since $\Lambda f = u|_{\partial\Omega}$, $\Sigma\tilde{P}^E f = \tilde{u}|_{\partial\Omega} + z\mathcal{B}_\nu\, \tilde{u}|_{\partial\Omega}$ on the electrodes and $\Sigma\tilde{P}^E f = 0$ in between the electrodes, the triangle inequality implies that

$$\begin{aligned}
&\left\| \left((\Lambda - \Lambda_0) - (\Sigma - \Sigma_0)\, \tilde{P}^E \right) f \right\|^2_{L^2(\partial\Omega)} \\
&\quad = \|u - u_0 - \tilde{u} - z\mathcal{B}_\nu\tilde{u} + \tilde{u}_0 + z\mathcal{B}_\nu\tilde{u}_0\|^2_{L^2(E)} + \|u - u_0\|^2_{L^2(G)} \\
&\quad \le 3\|u - \tilde{u}\|^2_{L^2(E)} + 3\|u_0 - \tilde{u}_0\|^2_{L^2(E)} \\
&\qquad + 3\,\|z\|^2_\infty \|\mathcal{B}_\nu\tilde{u} - \mathcal{B}_\nu\tilde{u}_0\|^2_{L^2(E)} + \|u - u_0\|^2_{L^2(G)}\,.
\end{aligned} \tag{7.3}$$

The terms $\|u - \tilde{u}\|_{L^2(E)}$ and $\|u_0 - \tilde{u}_0\|_{L^2(E)}$ can be estimated in exactly the same manner and we will treat only the first of the two in the following. We set again $c = \int_{\partial\Omega} \tilde{u}\ \mathrm{d}S$ and estimate

$$\begin{aligned}
\|u - \tilde{u}\|_{L^2(E)} &\le \|u - \tilde{u} + c\|_{L^2(E)} + |c||E|^{1/2} \\
&\le C(\Omega, \gamma, z)\|\mathcal{B}_\nu u - \mathcal{B}_\nu\tilde{u}\|_{H^{-1/2}(\partial\Omega)} + |c||E|^{1/2} \\
&\le C(\Omega, \gamma, z) \left(\underbrace{\|f - P^E f\|_{H^{-1/2}(\partial\Omega)}}_{\text{I}} + \underbrace{\|P^E f - \mathcal{B}_\nu\tilde{u}\|_{L^2(\partial\Omega)}}_{\text{II}} \right) + |c||E|^{1/2}.
\end{aligned} \tag{7.4}$$

The absolute value of c can be easily controlled, since

$$\begin{aligned}
c &= \int_{\partial\Omega} \tilde{u}\ \mathrm{d}S = \int_E \tilde{u}\,\mathrm{d}S + \int_G \tilde{u}\,\mathrm{d}S \\
&= \sum_{E_j \in \mathcal{E}} \int_{E_j} (U_j - z\mathcal{B}_\nu u)\ \mathrm{d}S + \int_G \tilde{u}\,\mathrm{d}S \\
&= \sum_{E_j \in \mathcal{E}} U_j|E_j|\ - z \sum_{E_j \in \mathcal{E}} I_j|E_j| + \int_G \tilde{u}\,\mathrm{d}S = \int_G \tilde{u}\,\mathrm{d}S
\end{aligned}$$

and hence $|c| \le \|\tilde{u}\|_{L^2(\partial\Omega)}|G|^{1/2} \le C(\Omega, \gamma)\|f\|_{L^2(\partial\Omega)}|G|^{1/2}$. Note that

$$\|\mathcal{B}_\nu\tilde{u} - \mathcal{B}_\nu\tilde{u}_0\|_{L^2(E)} \le \|\mathcal{B}_\nu\tilde{u} - P^E f\|_{L^2(\Omega)} + \|P^E f - \mathcal{B}_\nu\tilde{u}_0\|_{L^2(\Omega)},$$

and both of the latter terms correspond to the term II of (7.4). Terms I and II of (7.4) will be estimated in parts 2 and 3 of the proof. The only remaining term $\|u - u_0\|_{L^2(G)}$, appearing at the end of (7.3), will be treated in the fourth part.

2. We bound term I using the approximation property of $P^{E+G} = P^0_{\mathcal{T}}$ on $L^2(\partial\Omega)$; see Theorem 3.4. For convenience, we formulate this result as a lemma.

LEMMA 7.2. *Under the assumptions of Theorem* 7.1*, it holds that*

$$\|f - P^E f\|_{H^{-1/2}(\partial\Omega)} \le C(\Omega, \kappa_{\mathcal{T}}) \left(h^{1/2}_{\mathcal{T}} + \sqrt{\sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1}} \right) \|f\|_{L^2(\partial\Omega)}.$$

We estimate as follows:

$$\begin{aligned}
\mathrm{I} &= \|f - (P^E + P^G)f + P^G f\|_{H^{-1/2}(\partial\Omega)} \\
&\leq \|f - P^{E+G} f\|_{H^{-1/2}(\partial\Omega)} + \left\|P^G f - \textstyle⨍_G f \,\mathrm{d}S\right\|_{H^{-1/2}(\partial\Omega)} + \left\|\textstyle⨍_G f \,\mathrm{d}S\right\|_{H^{-1/2}(\partial\Omega)} \\
&\leq C(\Omega, \kappa_{\mathcal{T}})\, h_{\mathcal{T}}^{1/2} \|f\|_{L^2(\partial\Omega)} + C(\Omega) \left\|P^G f - \textstyle⨍_G f \,\mathrm{d}S\right\|_{L^2(G)} + \left\|\textstyle⨍_G f \,\mathrm{d}S\right\|_{H^{-1/2}(\partial\Omega)} \\
&\leq C(\Omega, \kappa_{\mathcal{T}})\, h_{\mathcal{T}}^{1/2} \|f\|_{L^2(\partial\Omega)} + C(\Omega)\, h_{\mathcal{G}}^2 \|f\|_{L^2(\partial\Omega)} + \left\|\textstyle⨍_G f \,\mathrm{d}S\right\|_{H^{-1/2}(\partial\Omega)},
\end{aligned}$$

where we used Corollary 6.2 and Lemma 6.1 for the gaps instead of the electrodes. The constant $C(\Omega, \kappa_{\mathcal{T}})$ depends on the curvature of $\partial\Omega$ and the shape regularity of the triangulation $\mathcal{T}$.

The last term on the right will be estimated using a duality argument, and so we start with the "dual" case $\phi \in H^{1/2}(\partial\Omega)$. Due to a Sobolev embedding theorem, we know that for $v \in W^{1,p}(\Omega)$, $p < n$, the trace $v|_{\partial\Omega}$ belongs to $L^\theta(\partial\Omega)$ for $1 \leq \theta < (n-1)p/(n-p)$; see [26, Abschnitt 117, Satz 4; Abschnitt 118] or [1, Theorem 7.43]. In our case, $p = 2$ and $n = 3$ and we find that $H^{1/2}(\partial\Omega) = H^1(\Omega)\big|_{\partial\Omega} \subset L^\theta(\partial\Omega)$ for $1 \leq \theta < 4$, with continuous embedding.[1] We fix $2 < \theta < 4$ and define the conjugate exponent θ^* in $(4/3, 2)$ by $1/\theta + 1/\theta^* = 1$. Then we estimate as follows:

$$\begin{aligned}
\left\|⨍_G \phi\right\|^2_{L^2(\partial\Omega)} &\leq \sum_{G_j \in \mathcal{G}} \left|\frac{1}{|G_j|} \int_{G_j} \phi \,\mathrm{d}S\right|^2 |G_j| \\
&\leq \sum_{G_j \in \mathcal{G}} \frac{1}{|G_j|} \|1\|^2_{L^{\theta^*}(G_j)} \|\phi\|^2_{L^\theta(G_j)} \qquad (7.5)\\
&\leq C(\Omega) \|\phi\|^2_{H^{1/2}(\partial\Omega)} \sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1} \quad \text{for } \phi \in H^{1/2}(\partial\Omega).
\end{aligned}$$

A duality argument is now used to bound the $L^2(\partial\Omega)$-orthogonal mean value projection $⨍(\cdot)$ on G in $H^{-1/2}(\partial\Omega)$:

$$\begin{aligned}
\left\|⨍_G f\right\|_{H^{-1/2}(\partial\Omega)} &= \sup_{\phi \in H^{1/2}(\partial\Omega), \phi \neq 0} \frac{\left|\langle ⨍ f, \phi\rangle_{L^2(\partial\Omega)}\right|}{\|\phi\|_{H^{1/2}(\partial\Omega)}} \\
&= \sup_{\phi \in H^{1/2}(\partial\Omega), \phi \neq 0} \frac{\left|\langle f, ⨍ \phi\rangle_{L^2(\partial\Omega)}\right|}{\|\phi\|_{H^{1/2}(\partial\Omega)}} \\
&\leq \sup_{\phi \in H^{1/2}(\partial\Omega), \phi \neq 0} \frac{\|f\|_{L^2(\partial\Omega)} \|⨍ \phi\|_{L^2(\partial\Omega)}}{\|\phi\|_{H^{1/2}(\partial\Omega)}} \\
&\overset{(7.5)}{\leq} C(\Omega) \sqrt{\sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1}} \|f\|_{L^2(\partial\Omega)}. \qquad \square
\end{aligned}$$

3. In this part of the proof we estimate term II of (7.4). Again, we announce the bound in a lemma.

Lemma 7.3. *Under the assumptions of Theorem* 7.1, *it holds that*

$$\|P^E f - \mathcal{B}_\nu \tilde{u}\|_{L^2(\partial\Omega)} \leq C(\Omega, \kappa_{\mathcal{T}}, z, c) \left(h_{\mathcal{T}}^{1/2} + |E|^{-1/2} h_{\mathcal{T}}^2 + |E|^{-1} |G|^{1/2}\right) \|f\|_{L^2(\partial\Omega)}.$$

[1] At this point, an extension to the two-dimensional case is not obvious without suitable modification, but see Remark 7.7.

We first observe that both $\mathcal{B}_\nu \tilde{u}$ and $P^E f$ are identically zero on G. Thus,

$$\begin{aligned}
\mathrm{II}^2 &= \|P^E f - \mathcal{B}_\nu \tilde{u}\|^2_{L^2(\partial\Omega)} \\
&\overset{(4.5),(5.1)}{=} \sum_{E_j \in \mathcal{E}} \left\| \left(\frac{1}{|E_j|} \int_{E_j} \mathcal{B}_\nu \tilde{u} \, \mathrm{d}S + K_f \right) - \mathcal{B}_\nu \tilde{u} \right\|^2_{L^2(E_j)} \\
&\le \underbrace{2 \sum_{E_j \in \mathcal{E}} \left\| \frac{1}{|E_j|} \int_{E_j} \mathcal{B}_\nu \tilde{u} \, \mathrm{d}S - \mathcal{B}_\nu \tilde{u} \right\|^2_{L^2(E_j)}}_{\mathrm{III}^2} + 2|\partial\Omega| \, |K_f|^2.
\end{aligned}$$

Lemma 6.3 gives a bound for the latter term on the right side of this inequality.

The term III can be bounded using the formulation of the complete electrode model. The following estimate is very strong, since we start with Neumann boundary values and end up with Dirichlet ones. Additionally, we obtain a factor $\sqrt{h_\mathcal{E}}$.

$$\begin{aligned}
\mathrm{III}^2 &= \sum_{E_j \in \mathcal{E}} \left\| \frac{1}{|E_j|} \int_{E_j} \left(\frac{U_j - \tilde{u}}{z} \right) \mathrm{d}S - \left(\frac{U_j - \tilde{u}|_{E_j}}{z} \right) \right\|^2_{L^2(E_j)} \qquad \text{by (4.4)} \\
&\le C(z) \sum_{E_j \in \mathcal{E}} \left\| \frac{1}{|E_j|} \int_{E_j} \tilde{u} \, \mathrm{d}S - \tilde{u} \right\|^2_{L^2(E_j)} \qquad \text{since } U_j/z \text{ constant} \\
&\le 2C(z) \left[\left\| \fint_E \tilde{u} - P^E \tilde{u} \right\|^2_{L^2(E)} + \|P^E \tilde{u} - \tilde{u}\|^2_{L^2(E)} \right] \\
&\le 2C(\Omega, \kappa_\mathcal{T}, z) \left[h^2_\mathcal{E} \|\tilde{u}\|^2_{L^2(E)} + h_\mathcal{E} \|\tilde{u}\|^2_{H^{1/2}(E)} \right] \quad \text{due to Theorems 3.4, 6.1} \\
&\le 2C(\Omega, \kappa_\mathcal{T}, z) h_\mathcal{E} \|\tilde{u}\|^2_{H^{1/2}(\partial\Omega)} \le 2C(\Omega, \kappa_\mathcal{T}, z) h_\mathcal{E} \|\tilde{u}\|^2_{H^1(\Omega)}
\end{aligned}$$

for $h_\mathcal{E}$ small enough, due to the boundedness of the trace mapping. The first inequality in the above chain of estimates is the only reason why we assumed that z is a constant in the beginning of this section. It is crucial to observe that this estimate remains valid if $\tilde{u}$ on the right-hand side is replaced by any $\tilde{u} + d$, $d \in \mathbb{C}$. This follows, for example, by noticing that the value of the term following the first of the above inequalities is not affected if we write $\tilde{u} + d$ in the place of $\tilde{u}$. As a consequence, with the help of Theorem 2.3 of [15] and (5.1), we see that

$$\begin{aligned}
\mathrm{III} &\le C(\Omega, \kappa_\mathcal{T}, z) h^{1/2}_\mathcal{E} \inf_{d \in \mathbb{C}} \|\tilde{u} + d\|_{H^1(\Omega)} \\
&\le C(\Omega, \kappa_\mathcal{T}, \gamma, z) h^{1/2}_\mathcal{E} \|\tilde{P}^E f\|_{L^2(\partial\Omega)} \\
&\le C(\Omega, \kappa_\mathcal{T}, \gamma, z) h^{1/2}_\mathcal{E} \left(\|f\|_{L^2(\partial\Omega)} + \|K_f\|_{L^2(\partial\Omega)} \right),
\end{aligned}$$

the last step being valid because $P^E : L^2(\partial\Omega) \to L^2(\partial\Omega)$ is bounded. Once again, we can bound the latter term on the right-hand side of the last inequality with the help of Lemma 6.3. □

4. In order to complete the proof, we still need to estimate the term $\|u - u_0\|_{L^2(G)}$ in (7.3). By the triangle inequality, it is sufficient to bound $\|u\|_{L^2(G)}$ and $\|u_0\|_{L^2(G)}$ separately. Again, we restrict ourselves to the first of these two terms. Using the approximation properties of different projections and the second part of this proof,

we find that

$$\begin{aligned}\|u\|_{L^2(G)} &\leq \|u - P^G u\|_{L^2(G)} + \left\|P^G u - \fint_G u\right\|_{L^2(G)} + \left\|\fint_G u\right\|_{L^2(G)} \\ &\leq C(\Omega, \kappa_{\mathcal{T}}) \left(h_{\mathcal{T}}^{1/2} + h_{\mathcal{T}}^2\right) \|u\|_{H^{1/2}(\partial\Omega)} + C(\Omega) \sqrt{\sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1}} \|u\|_{H^{1/2}(\partial\Omega)} \\ &\leq C(\Omega, \kappa_{\mathcal{T}}, \gamma) \left[h_{\mathcal{T}}^{1/2} + \sqrt{\sum_{G_j \in \mathcal{G}} |G_j|^{2/\theta^* - 1}} \right] \|f\|_{L^2(\partial\Omega)}.\end{aligned}$$

5. Now, the initial claim follows by combining part 4 and Lemmas 7.2 and 7.3 with the considerations in the beginning of this proof. □

REMARK 7.4. *Theorem* 7.1 *provides an approximation result suitable for the factorization method as the differences* $\Lambda - \Lambda_0$ *and* $(\Sigma - \Sigma_0)\tilde{P}^E$ *are considered. If one merely considers the difference between* Λ *and* $\Sigma\tilde{P}^E$, *the analogous approximation result fails. The reason for this is that the resistivity matrix does not approximate the Neumann-to-Dirichlet operator but the Neumann-to-Robin operator, defined as*

$$\Upsilon : H^s_\diamond(\partial\Omega) \to H^s_\diamond(\partial\Omega), \quad f \mapsto u|_{\partial\Omega} + zf \qquad \textit{for } s \in [-1/2, 1/2],$$

with a suitable grounding of the potential. This operator "of the second kind" does not share the smoothing properties of Λ. *However, if one considers the difference* $\|\Upsilon - \Sigma P^E\|_{\mathcal{L}(L^2_\diamond(\partial\Omega), H^s_\diamond(\partial\Omega))}$ *in weaker norms, i.e., for* $s < 0$, *then a similar bound as given in Theorem* 7.1 *holds.*

REMARK 7.5. *By taking advantage of Lemma* 6.1 *and the continuity of the mappings* $\Sigma, \Sigma_0 : \mathbb{C}^p_E \to \mathbb{C}^p_E$, *Theorem* 7.1 *could also be formulated using the projection that is related to* $\fint_E$ *in the same way as* $\tilde{P}_E$ *is related to* P_E. *This observation can be useful in practical considerations since constructing* $\fint_E$ *does not necessarily require construction of the diffeomorphisms* $(\chi_T)_{T \in \mathcal{E}}$.

REMARK 7.6. *Although we formulated Theorem* 7.1 *for the admittance* γ, *defined by* (2.1), *and the background admittance* γ_0, *the result holds true for any other two smooth enough admittances as well. In particular, the two admittances do not need to be related in the same way as the ones we used above.*

REMARK 7.7. *The crucial point for extension of the last theorem to two dimensions is the Sobolev embedding theorem used in the proof of Lemma* 7.2. *All other steps are independent of dimension. As proven in* [23, *section* 8.5], $H^{1/2}(\mathbb{R}) \hookrightarrow L^\theta(\mathbb{R})$ *for any* $\theta \in [2, \infty)$. *Using standard localization techniques this embedding can be transported to* $\partial\Omega$ *and serves, in two dimensions, as a replacement of the Sobolev embeddings used in* (7.5).

Let us now consider a family of triangulations $(\mathcal{T}_M)_{M \in \mathbb{N}}$ of $\partial\Omega$ and denote the electrodes of $\mathcal{T}_M$ by $\mathcal{E}^M$ and the gaps by $\mathcal{G}^M$. The spaces of constant functions on the electrodes and on the gaps are denoted $\mathcal{P}^E_M$ and $\mathcal{P}^G_M$, respectively, and the local L^2-projection operators by P^E_M and P^G_M. Furthermore, the auxiliary projection operator $\tilde{P}^E_M$ is defined in accordance with (5.1). By the definition of a family of triangulations the mesh size of $\mathcal{T}_M$ tends to zero as M tends to infinity. This implies that the mesh size of the electrodes, as well as the mesh size of the gaps, tends to zero. In view of our aim to prove that $(\Sigma_M - \Sigma_{0M})\tilde{P}^E_M$ approximates $\Lambda - \Lambda_0$, we need to require that the surface measure of the gaps tends to zero as M tends to infinity. Such a condition can be set up as

$$(7.6)\qquad \sum_{G_j^M \in \mathcal{G}^M} \left|G_j^M\right| = \left|\partial\Omega \setminus \bigcup_{E_j^M \in \mathcal{E}^M} E_j^M\right| \to 0 \quad \text{as } M \to \infty$$

and is inspired by Hyvönen's notion of a sequence of electrode configurations [15]. However, in view of (7.1), we are forced to set up a stronger assumption, namely,

$$(7.7)\qquad \sum_{G_j \in \mathcal{G}_M} \left|G_j^M\right|^{2/\theta^* - 1} \to 0 \quad \text{as } M \to \infty.$$

For $\theta^* = 1$ this condition is equivalent to (7.6), but the proof of Theorem 7.1 requires that $\theta^* \in (4/3, 2)$. Hence, fast shrinking of the gaps is assumed. If (7.7) is satisfied, we have norm convergence of the difference of the two resistivity matrices toward the difference of the corresponding Neumann-to-Dirichlet operators.

Theorem 7.8. *Let $(\mathcal{T}_M)_{M\in\mathbb{N}}$ be a family of triangulations of $\partial\Omega$ satisfying condition* (7.7)*. Then*

$$\|(\Lambda - \Lambda_0) - (\Sigma_M - \Sigma_{0M})\tilde{P}_M^E\|_{\mathcal{L}\left(L^2_\diamond(\partial\Omega)\right)} =: \delta_M \to 0 \quad \textit{as } M \to \infty.$$

Proof. Since $(\mathcal{T}_M)_{M\in\mathbb{N}}$ is shape regular, we can bound $\kappa_{\mathcal{T}_M}$ by a common constant κ. Hence, the constant in (7.1) is uniformly bounded for all $M \in \mathbb{N}$. Condition (7.7) implies, in particular, that the surface area of the gaps $|G| = \sum |G_j|$ tends to zero. Consequently, $|E| = \sum |E_j| \to |\partial\Omega|$, and the claim follows from (7.1). □

The rest of this work concentrates on using the above result in building a factorization algorithm for the complete electrode model of EIT; thorough numerical testing of the approximation link provided by Theorems 7.1 and 7.8 is left for future articles.

8. A factorization method for the complete electrode model. In this section we combine the approximation result of Theorem 7.8 with the perturbation result of Theorem 2.3. In view of the series criterion (2.6) of the factorization method, we need to consider the self-adjoint operator $(\Lambda - \Lambda_0)_\sharp$ instead of $\Lambda - \Lambda_0$. However, if $(\Sigma_M - \Sigma_{0M})\tilde{P}_M^E$ approaches $\Lambda - \Lambda_0$ in $\mathcal{L}(L^2_\diamond(\partial\Omega))$, then $((\Sigma_M - \Sigma_{0M})\tilde{P}_M^E)_\sharp$ approaches $(\Lambda - \Lambda_0)_\sharp$ in $\mathcal{L}(L^2_\diamond(\partial\Omega))$, as the following result shows.

Proposition 8.1. *Under the assumptions and notation of Theorem* 7.8*,*

$$\left\|(\Lambda - \Lambda_0)_\sharp - \big((\Sigma_M - \Sigma_{0M})P_M^E\big)_\sharp\right\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} \le C(2 + |\ln\delta_M|)\delta_M \to 0 \quad \textit{as } M \to \infty.$$

Proof. The proof uses the following estimate of Vainikko [28]. If $A, B \in \mathcal{L}(H)$ are bounded operators on a Hilbert space H and $\|A - B\| \le \varepsilon$, then

$$(8.1)\qquad \||A|^p - |B|^p\| \le C_p(1 + |\ln\varepsilon|)\varepsilon^{\min(1,p)} \quad \text{for any real } p > 0.$$

We abbreviate $A = \Lambda - \Lambda_0$ and $B_M = (\Sigma_M - \Sigma_{0M})\tilde{P}_M^E$ and observe that

$$\begin{aligned}\|\operatorname{Re} A - \operatorname{Re} B_M\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} &= \frac{1}{2}\|A + A^* - (B_M + B_M^*)\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))}\\ &\le \|A - B_M\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} \le \delta_M,\end{aligned}$$

and the same estimate also holds for $\operatorname{Im} A - \operatorname{Im} B_M$. Therefore, we can use (8.1) with $p = 1$ to estimate as follows:

$$\begin{aligned}\|A_\sharp - B_{M\sharp}\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} &\le \||\operatorname{Re} A| - |\operatorname{Re} B_M|\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} + \|\operatorname{Im}(A - B_M)\|_{\mathcal{L}(L^2_\diamond(\partial\Omega))} \\ &\le C_1(1 + |\ln(\delta_M)|)\delta_M + \delta_M \le \max(C_1, 1)\,(2 + |\ln(\delta_M)|)\delta_M.\end{aligned}$$

It holds that $|\ln(\delta_M)|\delta_M \to 0$ as $M \to \infty$ since δ_M is a zero sequence. Hence, $B_{M\sharp}$ approximates $A_\sharp$ if B_M approximates A. □

The $L^2_\diamond(\partial\Omega)$ estimate of the preceding proposition enables to us to use the material from section 2 to construct a factorization method for the complete electrode model. The idea is to approximate the infinite series (2.6) using the eigenvalues λ_j^M and eigenvectors ψ_j^M of $((\Sigma_M - \Sigma_{0M})\tilde{P}_M^E)_\sharp$.

THEOREM 8.2. *Let $\omega \in (0, 1/2)$ and $C > 0$ be arbitrary, but fixed, parameters, and consider the finite-dimensional operators $(\Sigma_M - \Sigma_{0M})\tilde{P}_M^E \in \mathcal{L}(L^2_\diamond(\partial\Omega))$, $M \in \mathbb{N}$. We denote by $(\lambda_j^{(M)}, \psi_j^{(M)})_{j\in\mathbb{N}}$ an eigensystem of $((\Sigma_M - \Sigma_{0M})\tilde{P}_M^E)_\sharp$ and define the truncation index $R(M)$ according to* (2.8), *with the noise level ε_M set as*

$$\varepsilon_M := C(2 + |\ln \delta_M|)\delta_M, \quad M \in \mathbb{N},$$

where C is the constant from Proposition 8.1 *and δ_M is defined as in Theorem* 7.8. *Then the sequence*

$$M \mapsto \sum_{j=1}^{R(M)} \frac{|\langle \varphi_y, \psi_j^{(M)}\rangle|^2}{\lambda_j^{(M)}}, \quad M \in \mathbb{N},$$

is bounded if and only if $y \in \Omega_c$.

In the following section, we will demonstrate that Theorem 8.2 is not only of theoretical interest; it can also lead to practical reconstruction algorithms.

9. Numerical experiments.

In this section, we present numerical experiments that implement a simplified version of the series criterion of Theorem 8.2 with simulated electrode data. Although the theoretical results of this paper have been formulated in three spatial dimensions (but note Remark 7.7), the numerical tests are carried out in two dimensions: Our object of interest Ω is an isotropic unit square characterized by unit background admittance and unit contact impedance. Moreover, there are sixteen electrodes that cover 44 percent of $\partial\Omega$. We consider two different inclusion geometries: Ω is contaminated in the first test by one kite-shaped inclusion and, in the second test, by a kite-shaped inclusion and a circular inhomogeneity. The admittance inside the inclusions is two. Our experimental settings are exactly the ones used in the third and the fourth numerical tests of [17], where the reader can find the precise details on the measurement geometry, the computation of the forward data, and the simulation of the measurement noise. This arrangement also gives the reader an opportunity to compare the series criterion and the factorization-type algorithm of [17] in a straightforward manner. More extensive studies on numerical implementation of factorization-type algorithms for different electrode models can be found in [14, 17].

The test functions φ_y, defined on the boundary of the unit square, are computed with the help of a suitable analytic map and the known functional form of the test functions in the unit disk [3, 5, 17]. To be more precise, we introduce

$$\tilde{\varphi}_y(x) = \tilde{\varphi}_{y,b}(x) := \frac{1}{\pi}\frac{(y - x)\cdot b}{|y - x|^2}, \quad |x| = 1,\ |y| < 1,\ b \in \mathbb{R}^2, \tag{9.1}$$

and let $\eta : \Omega \to \{z \in \mathbb{R}^2 \mid |z| < 1\}$ be a bijective analytic function whose derivative does not vanish anywhere on Ω. We set

$$\varphi_y = \tilde{\varphi}_{\eta(y)} \circ \eta + c,$$

where the constant c is chosen in such a way that φ_y integrates to zero over $\partial\Omega$. We compute the auxiliary test functions $\tilde{\varphi}_{y,b}$, $|y| < 1$, with a fixed dipole moment b, which results in $\varphi_y = \varphi_{y,a}$, where the dipole moment $a \in \mathbb{R}^2$ (see (2.4)) depends on the probe location $y \in \Omega$ through η (cf. [3, 17]). The analytic mapping η needed above is provided by the Schwarz–Christoffel toolbox for MATLAB [10].

Using the notation of Theorem 8.2, we define a preliminary indicator function through

$$\alpha_{R,b}(y) = \sum_{j=1}^{R} \frac{|\langle \varphi_y, \psi_j \rangle|^2}{\lambda_j} \Big/ \sum_{j=1}^{R} |\langle \varphi_y, \psi_j \rangle|^2, \quad y \in \Omega, \ 1 \leq R \leq 15, \tag{9.2}$$

where b is the dipole moment in the unit disk (see (9.1)) and we have left out the parameter M appearing in the formulae of Theorem 8.2 since the electrode configuration is fixed. Take note that we have introduced the normalizing factor $\sum_{j=1}^{R} |\langle \varphi_y, \psi_j \rangle|^2$ in the denominator of (9.2) since one is ultimately interested in the shape of the test function φ_y and not in its magnitude; it is easy to check that all above theoretical results remain valid if φ_y is replaced by $\varphi_y / \|\varphi_y\|_{L^2(\partial\Omega)}$. In order to average out artifacts, we take the mean of $\alpha_{R,b}(y)$ over three dipole moments:

$$\alpha_R(y) = \frac{1}{3} \sum_{k=0}^{2} \alpha_{R,b_k}(y), \quad y \in \Omega,$$

where $b_k = (\cos(2\pi k/3), \sin(2\pi k/3))$.

For visualization purposes, i.e., to obtain better contrast, we introduce one more indicator function, namely,

$$\operatorname{ind}(y) = \frac{1}{\alpha_R(y)}, \quad y \in \Omega.$$

The reconstructions of Figures 9.1 and 9.2 were obtained by plotting ind over Ω. Theorem 8.2 suggests that $\alpha_R(y)$ is probably larger when $y \in \Omega \setminus D$ than when $y \in D$ if $1 \leq R \leq 15$ is chosen suitably. Hence, $\operatorname{ind}(y)$ should be larger when $y \in D$ than when $y \in \Omega \setminus D$.

In both tests, we use three different noise levels: $\varepsilon = 0$, 2×10^{-4}, and 2×10^{-3} (cf. [17]). For the noiseless cases the cut-off parameter is chosen to be $R = 12$. When working with noisy resistivity matrices, we use a significantly simplified and less conservative version of (2.8): Since the admittances are real and the inclusions are more conductive than the background, the operator $\Sigma - \Sigma_0$ is self-adjoint and negative definite [15]. In particular, it holds that $((\Sigma - \Sigma_0)\tilde{P}^E)_\sharp = (\Sigma_0 - \Sigma)\tilde{P}^E$. When noise is added to Σ (cf. [17]), some eigenvalues of $\Sigma_0 - \Sigma$ may become negative. We choose the cut-off parameter R and rearrange the eigenvalues if necessary, so that the eigenvectors corresponding to the negative eigenvalues of the noisy $\Sigma_0 - \Sigma$ do not contribute to (9.2). In the first experiment, this approach produced the cut-off parameters $R = 11$ and 8 for the noise levels $\varepsilon = 2 \times 10^{-4}$ and 2×10^{-3}, respectively. In the second test, the corresponding values were $R = 10$ and 9. Notice that different realizations of measurement noise produce different cut-off parameters.

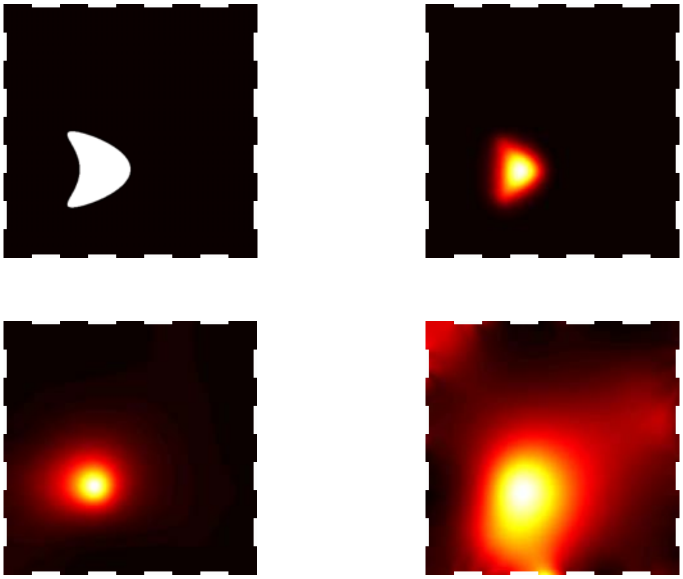


Fig. 9.1. *The first experiment. The parameters ε and R represent the noise level and the spectral cut-off, respectively. Top left: The inclusion support. Top right: $\varepsilon = 0$ and $R = 12$. Bottom left: $\varepsilon = 2 \times 10^{-4}$ and $R = 11$. Bottom right: $\varepsilon = 2 \times 10^{-3}$ and $R = 8$.*

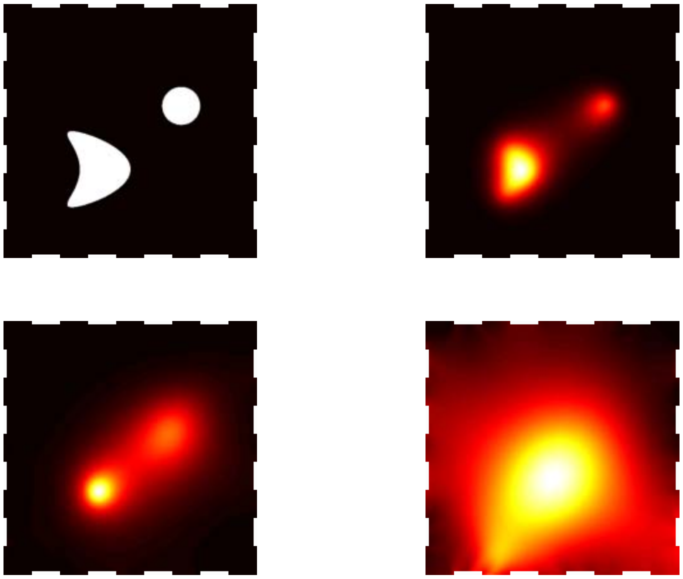


Fig. 9.2. *The second experiment. The parameters ε and R represent the noise level and the spectral cut-off, respectively. Top left: The inclusion support. Top right: $\varepsilon = 0$ and $R = 12$. Bottom left: $\varepsilon = 2 \times 10^{-4}$ and $R = 10$. Bottom right: $\varepsilon = 2 \times 10^{-3}$ and $R = 9$.*

The findings of the first numerical experiment, where Ω contains one kite-shaped inclusion, are presented in Figure 9.1, and the reconstructions of the second experiment, where Ω is contaminated by a kite-shaped inclusion and a circular inhomogeneity, are shown in Figure 9.2. As the figures illustrate, the noiseless reconstructions capture the shapes and the locations of the inclusions quite well. When $\varepsilon = 2 \times 10^{-4}$, the function ind : $\Omega \to \mathbb{R}$ still provides information on the whereabouts of the inhomogeneities. Unfortunately, with the highest noise level $\varepsilon = 2 \times 10^{-3}$, the reconstructions are badly blurred and their information content is rather low. We conclude that the above-described simplified version of the method implicated by Theorem 8.2 provides useful information on the inclusions if the signal to noise ratio of the measurements is not too low.

The quality of our reconstructions is approximately the same as of those obtained by a factorization-type algorithm in [17]. However, the above-introduced algorithm has a slight advantage in the noisy case: Its regularization parameter, i.e., the spectral cut-off R, is chosen in a systematic way by looking at the measurement data. The algorithm used in [14], even more notably, has this advantageous property but lacks the asymptotic analysis, which is provided for the series criterion by this work, and it has not been properly studied within the complete electrode model.

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, 2nd ed., Academic Press, New York, 2003.

[2] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[3] M. Brühl, *Gebietserkennung in der elektrischen Impedanztomographie*, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1999.

[4] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[5] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[6] M. Brühl, M. Hanke, and M. Pidcock, *Crack detection using electrostatic measurements*, Math. Model. Numer. Anal., 35 (2001), pp. 595–605.

[7] A. Calderón, *On an inverse boundary value problem*, in Seminar of Numerical Analysis and Its Applications to Continuum Physics (Rio de Janeiro, 1980), Sociedade Brasileira de Mathemàtica, Rio de Janeiro, Brazil, 1980, pp. 65–73.

[8] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[9] D. Colton and A. Kirsch, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.

[10] T. A. Driscoll, *Algorithm 756: A MATLAB toolbox for Schwarz-Christoffel mapping*, ACM Trans. Math. Software, 22 (1996), pp. 168–186.

[11] B. Gebauer, *The factorization method for real elliptic problems*, Z. Anal. Anwend., 25 (2006), pp. 81–102.

[12] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, New York, 1986.

[13] N. Grinberg and A. Kirsch, *The linear sampling method in inverse obstacle scattering for impedance boundary conditions*, J. Inverse Ill-Posed Probl., 10 (2002), pp. 171–185.

[14] M. Hanke and M. Brühl, *Recent progress in electrical impedance tomography*, Inverse Problems, 19 (2003), pp. S65–S90.

[15] N. Hyvönen, *Complete electrode model of electrical impedance tomography: Approximation properties and characterization of inclusions*, SIAM J. Appl. Math., 64 (2004), pp. 902–931.

[16] N. Hyvönen, *Application of the factorization method to the characterization of weak inclusions in electrical impedance tomography*, Adv. in Appl. Math., 39 (2007), pp. 197–221.

[17] N. Hyvönen, H. Hakula, and S. Pursiainen, *Numerical implementation of the factorization method within the complete electrode model of impedance tomography*, Inverse Problems and Imaging, 1 (2007), pp. 299–317.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1995.

[19] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[20] A. Kirsch, *Factorization of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory*, Inverse Problems, 15 (1999), pp. 413–429.

[21] A. Kirsch, *The factorization method for a class of inverse elliptic problems*, Math. Nachr., 278 (2004), pp. 258–277.

[22] A. Lechleiter, *A regularization technique for the factorization method*, Inverse Problems, 22 (2006), pp. 1605–1625.

[23] E. H. Lieb and M. Loss, *Analysis*, 2nd ed., AMS, Providence, RI, 2001.

[24] W. McLean, *Strongly Elliptic Systems and Boundary Integral Operators*, Cambridge University Press, Cambridge, UK, 2000.

[25] S. Sauter and C. Schwab, *Randelementmethoden*, Teubner, Berlin, 2004.

[26] W. Smirnow, *Lehrgang der höheren Mathematik, Teil V*, 2nd ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.

[27] E. Somersalo, M. Cheney, and D. Isaacson, *Existence and uniqueness for electrode models for electric current computed tomography*, SIAM J. Appl. Math., 52 (1992), pp. 1023–1040.

[28] G. Vainikko, *The discrepancy principle for a class of regularization methods*, U.S.S.R. Comput. Maths. Math. Phys., 21 (1982), pp. 1–19.

SIAM J. Appl. Math.
Vol. 68, No. 4, pp. 1122–1145

FINGERING FROM IONIZATION FRONTS IN PLASMAS*

MANUEL ARRAYÁS†, SANTIAGO BETELÚ‡, MARCO A. FONTELOS§, AND JOSÉ L. TRUEBA†

Abstract. In this paper we describe the formation of fingers from ionization fronts for a hydrodynamic plasma model. The fingers result from a balance between the destabilizing effect of impact ionization and the stabilizing effect of electron diffusion on ionization fronts. We show that electron diffusion acts as an effective surface tension on moving fronts and we estimate analytically the size of the fingers and its dependence on both the electric field and electron diffusion coefficient. We perform direct numerical simulation of the model and compute finger-like traveling waves analogous to structures such as Saffman–Taylor fingers and Ivantsov paraboloid in the context of Hele–Shaw and Stefan problems, respectively.

Key words. ionization fronts, fingering instabilities, plasmas, pattern formation

AMS subject classifications. 35K55, 35K57, 76X05, 65N06

DOI. 10.1137/050647074

1. Introduction.

Lightning is a stream of electrified air, known as plasma. Charged particles are bound in the air by powerful electric forces to form electrically neutral atoms and molecules. As a result, the air is an excellent insulator. This means that if we apply an electric field to a volume filled with neutral particles, electric currents will not flow. However, if a very strong electric field is applied to matter of low conductivity and some electrons or ions are created, then the few mobile charges can generate an avalanche of more charges by impact ionization. A low temperature plasma is created, resulting in an electric discharge. The change in the properties of a dielectric that causes it to become conductive is known as electric breakdown. Breakdown is a threshold process: no changes in the state of the medium are noticeable while the electric field across a discharge gap is gradually increased but, suddenly, at a certain value of the electric field, a current is detected.

Discharges can assume different appearances depending on the characteristics of the electric field and the properties of the medium. Phenomenologically, discharges can be classified into stationary ones, such as arc, glow, or dark discharges, and transient ones, such as sparks and leaders [18].

At atmospheric pressure and at distances over 1 cm between anode and cathode, the discharge channels are sharp and narrow, and we have a streamer discharge. A streamer is a sharp ionization wave that propagates into a nonionized gas, leaving a nonequilibrium plasma behind. Streamers have been also reported in early stages of atmospheric discharges [15, 17]. They can split into branches spontaneously, but how this branching is determined by the underlying physics is one of the greatest unsolved problems in the physics of electric discharges. The pattern of this branching resembles

*Received by the editors December 8, 2005; accepted for publication (in revised form) December 10, 2007; published electronically March 19, 2008.
http://www.siam.org/journals/siap/68-4/64707.html

†Area de Electromagnetismo, Universidad Rey Juan Carlos, Camino del Molino s/n, 28943 Fuenlabrada, Madrid, Spain (manuel.arrayas@urjc.es, joseluis.trueba@urjc.es).

‡Department of Mathematics, University of North Texas, P.O. Box 311430, Denton, TX 76203-1430 (betelu@unt.edu).

§Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain (marco.fontelos@uam.es).

the ones observed in the propagation of cracks, dendritic growth and viscous fingering. Those phenomena are known to be governed by deterministic equations rather than by stochastic events. In this paper, we extend and generalize the results announced in [8] and implement direct numerical simulations of these deterministic models.

1.1. The minimal model for streamers. We consider a fluid description of a low-ionized plasma. The electron density N_e^d varies in time as

$$\frac{\partial N_e^d}{\partial \tau^d} + \nabla_{\mathbf{R}}^d \cdot \mathbf{J}_e^d = S_e^d. \tag{1.1}$$

In this expression, the superscript d means that the quantity has physical dimensions so that τ^d is the physical time, $\nabla_{\mathbf{R}}^d$ is the gradient operator, S_e^d is the source term, i.e., the net creation rate of electrons per unit volume, and

$$\mathbf{J}_e^d(\mathbf{R}^d, \tau^d) = N_e^d(\mathbf{R}^d, \tau^d)\, \mathbf{U}_e^d(\mathbf{R}^d, \tau^d)$$

is the electron current density, with $\mathbf{U}_e^d$ being the average velocity of electrons. Similar expressions can be obtained for positive N_p^d and negative N_n^d ion densities. On timescales of interest for the case of negative streamers, the ion currents can be neglected because they are more than two orders of magnitude smaller than the electron one, so we will take

$$\frac{\partial N_p^d}{\partial \tau^d} = S_p^d, \tag{1.2}$$

$$\frac{\partial N_n^d}{\partial \tau^d} = S_n^d, \tag{1.3}$$

with $S_{p,n}^d$ being source terms for positive and negative ions. Conservation of charge has to be imposed in all processes so that the condition $S_p^d = S_e^d + S_n^d$ holds.

A usual procedure is to approximate the electron current $\mathbf{J}_e^d$ as the sum of a drift (electric force) and a diffusion term

$$\mathbf{J}_e^d = -\mu_e \boldsymbol{\mathcal{E}}^d N_e^d - D_e^d \nabla_{\mathbf{R}}^d N_e^d, \tag{1.4}$$

where $\boldsymbol{\mathcal{E}}^d$ is the total electric field (the sum of the external electric field applied to initiate the propagation of a ionization wave and the electric field created by the local point charges) and μ_e^d and D_e^d are the mobility and diffusion coefficients of the electrons. Note that, as the initial charge density is low and there is no applied magnetic field, the magnetic effects in (1.4) are neglected. In principle, the diffusion coefficient is not completely determined but, in the case of equilibrium, diffusion is linked to mobility through the Einstein relation $D_e^d/\mu_e = kT/e$, with k being the Boltzmann constant, T the temperature, and e the absolute value of the electron charge.

Several physical processes can be considered to give rise to the source terms $S_{e,p,n}^d$. The most important of them are impact ionization (an accelerated electron collides with a neutral molecule and ionizes it), attachment (an electron may become attached when it collides with a neutral gas atom or molecule, forming a negative ion), recombination (of a free electron with a positive ion or a negative ion with a positive ion), and photoionization (when photons created by recombination or scattering processes interact with a neutral atom or molecule, producing a free electron and a positive ion; see [7] and references therein).

A model to describe streamers is obtained when explicit expressions for the source terms, the electron mobility μ_e, and the diffusion coefficient D_e^d are provided. It is also necessary to impose equations for the evolution of the electric field $\boldsymbol{\mathcal{E}}^d$. It is usual to consider that this evolution is given by Poisson's law,

$$\nabla_{\mathbf{R}}^d \cdot \boldsymbol{\mathcal{E}}^d = \frac{e}{\varepsilon_0} \left(N_p^d - N_n^d - N_e^d \right), \tag{1.5}$$

where ε_0 is the permittivity of the gas and we are assuming that the absolute value of the charge of positive and negative ions is e.

A simplification occurs when the streamer development out of a macroscopic initial ionization seed is considered in a nonattaching gas such as argon or nitrogen [12]. In this case, attachment, recombination, and photoionization processes can be neglected. As a consequence, the negative ion density N_n^d can be considered constant. The balance equations turn out to be

$$\frac{\partial N_e^d}{\partial \tau^d} = \nabla_{\mathbf{R}}^d \cdot \left(\mu_e \boldsymbol{\mathcal{E}}^d N_e^d + D_e^d \nabla_{\mathbf{R}}^d N_e^d \right) + \nu_i N_e^d, \tag{1.6}$$

$$\frac{\partial N_p^d}{\partial \tau^d} = \nu_i N_e^d. \tag{1.7}$$

This is called the minimal streamer model for a nonattaching gas. In these equations, $\nu_i N_e^d$ is a model for the impact ionization source term, in which the ionization coefficient ν_i is given by the phenomenological Townsend's approximation,

$$\nu_i = \mu_e |\boldsymbol{\mathcal{E}}^d| \alpha_0 e^{-\mathcal{E}_0 / |\boldsymbol{\mathcal{E}}^d|}, \tag{1.8}$$

where α_0 is the inverse of ionization length. The ionization length is the distance, on average, that a free electron travels before ionizing a molecule. The value of α_0 is proportional to the pressure of the ambient gas according to Townsend's theory [9]. $\mathcal{E}_0$ is the characteristic impact ionization electric field.

Townsend's approximation provides physical scales and intrinsic parameters for the model as long as only impact ionization is present in the gas. It is then convenient to reduce the equations to a dimensionless form. The natural units for nitrogen are functions of the gas pressure p (in bars). These units are the ionization length

$$R_0 = \frac{1}{\alpha_0} = 2.3 \times 10^{-6}\,\mathrm{m} \left(\frac{p}{1\,bar} \right)^{-1}, \tag{1.9}$$

as a length unit, the characteristic impact ionization field

$$\mathcal{E}_0 = 2 \times 10^7\,\mathrm{V/m} \left(\frac{p}{1\,bar} \right), \tag{1.10}$$

as an electric field unit, and the electron mobility

$$\mu_e = 3.8 \times 10^{-2}\,\mathrm{m}^2/(\mathrm{V} \cdot \mathrm{s}) \left(\frac{p}{1\,bar} \right)^{-1}, \tag{1.11}$$

as a unit of velocity divided by electric field. These natural units lead to the velocity scale

$$U_0 = \mu_e \mathcal{E}_0 = 7.6 \times 10^5\,\mathrm{m/s}, \tag{1.12}$$

the time scale

$$\tau_0 = \frac{R_0}{U_0} = 3 \times 10^{-12}\,\mathrm{s}\,\left(\frac{p}{1\,bar}\right)^{-1}, \tag{1.13}$$

the particle density scale

$$N_0 = \frac{\varepsilon_0 \mathcal{E}_0}{eR_0} = 4.7 \times 10^{20}\,\mathrm{m}^{-3}\,\left(\frac{p}{1\,bar}\right)^{2}, \tag{1.14}$$

and the electron diffusion scale

$$D_0 = R_0 U_0 = 1.8\,\mathrm{m}^2/\mathrm{s}\,\left(\frac{p}{1\,bar}\right)^{-1}. \tag{1.15}$$

We introduce the dimensionless variables $\mathbf{r} = \mathbf{R}^d/R_0$, $\tau = \tau^d/\tau_0$, the dimensionless field $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}^d/\mathcal{E}_0$, the dimensionless electron and positive ion particle densities $N_e = N_e^d/N_0$ and $N_p = N_p^d/N_0$, and the dimensionless diffusion constant $D_e = D_e^d/D_0$.

The dimensionless minimal model reads

$$\frac{\partial N_e}{\partial \tau} = \nabla \cdot (N_e \boldsymbol{\mathcal{E}} + D_e\,\nabla N_e) + N_e |\boldsymbol{\mathcal{E}}| e^{-1/|\boldsymbol{\mathcal{E}}|}, \tag{1.16}$$

$$\frac{\partial N_p}{\partial \tau} = N_e |\boldsymbol{\mathcal{E}}| e^{-1/|\boldsymbol{\mathcal{E}}|}, \tag{1.17}$$

$$N_p - N_e = \nabla \cdot \boldsymbol{\mathcal{E}}. \tag{1.18}$$

This model exhibits spontaneous branching of the streamers, as indicated by numerical simulations [4], in agreement with experimental situations [17]. In order to understand this branching, Arrayás and Ebert [5] derived the dispersion relation for transversal Fourier-modes of planar negative shock fronts without diffusion (D_e). For perturbations of small wave number k, the planar shock front becomes unstable with a linear growth rate proportional to k. It has been also shown that all the modes with large enough wave number k (small wave length perturbations) grow at the same rate (the growth rate does not depend on k when k is large). However, it could be expected from the physics of the problem that a particular mode would be selected. To address this problem, we consider in this paper the effect of diffusion.

1.2. Outline of this paper. Our analysis will show that the electron density N_e may develop sharp fronts of thickness $O(\sqrt{D_e})$. Moreover, it satisfies an equation analogous to the Fisher equation, which is a well-known model in some biological contexts (see [14]). A surprising fact established during the last 30 years is that the combination of sharp interfaces with small diffusive effects may result in asymptotic limits (for $D_e \ll 1$) in which the motion of the interface is described by equations involving solely geometrical properties such as its mean curvature. A pioneer attempt to achieve such a description is due to Allen and Cahn [2] and concerns a model, today known as the Allen–Cahn equation, for the kinetics of melted Fe-Al alloys. Subsequent work by Rubinstein, Sternberg, and Keller [19] showed that the points of the interface separating both species move along the normal direction with a velocity proportional to its mean curvature. This kind of dynamics is termed "mean curvature flow." Many mathematicians have contributed to providing a rigorous proof of the convergence of the Allen–Cahn model to motion by mean curvature. These ideas have also been extended to some other, rather different, contexts. An improvement of the above model is the so-called Cahn–Hilliard model [10], described by a fourth order differential equation. This model leads to an asymptotic limit given by the motion of sharp interfaces in the Hele–Shaw (or Mullins–Sekerka) problem for the evolution of a fluid between two plates separated by a small distance [1]. A biological model consisting of reaction-diffusion equations [11] for competing species separated by a sharp interface gives rise to a limiting problem similar to the Stefan problem for phase transformation (for example, ice solidifying water). Remarkably, some of these limiting models have solutions that develop branch-like patterns, such as fingers in Hele–Shaw or dendrites in the Stefan problem.

In this paper we exploit some of the ideas introduced in the references above in order to study the motion of ionization fronts. We will show that a planar front separating a partly ionized region from a region without charge is affected by two opposing effects: electrostatic repulsion of electrons and electron diffusion. The first effect tends to destabilize the front, while the second acts effectively as a mean curvature contribution to the velocity of the front that stabilizes it. The net result is the appearance of fingers with a characteristic thickness determined by the balance of these two opposing actions. The common underlying mathematical structure among the minimal streamer model and other pattern-forming systems such as the Hele–Shaw and Stefan problems strongly suggests that the basic mechanisms governing important phenomena such as the development of complex patterns through branching of single "fingers" should be very similar.

2. Streamer evolution in strong electric fields. In order to study the evolution and branching of ionization fronts, we consider the following experimental situation. The space between two plates is filled with a nonattaching gas such as nitrogen. A stationary potential difference is applied to these plates so that an electric field is produced in the gas. The electric field is directed from the anode to the cathode. To initiate the avalanche, an initial seed of ionization is set near the cathode. We study the evolution of negative ionization fronts towards the anode.

We will assume that the distance between the cathode and the anode is much larger than the space scale R_0 (in experiments, this distance is more than one thousand times larger than R_0) so that we can consider the anode to be at an infinite distance from the initial seed of ionization. Moreover, we will concentrate on the study of the dynamics under the effect of strong external electric fields, which are larger than the electric field unit $\mathcal{E}_0$. This means that the modulus of the dimensionless electric field $|\boldsymbol{\mathcal{E}}|$ is larger than 1. Strictly speaking, if we denote by $\mathcal{E}_\infty$ the modulus of the dimensionless electric field at large distance from the cathode, we will assume that $\mathcal{E}_\infty \gg 1$. Under these circumstances, it is natural to rescale the dimensionless quantities in the minimal model as

$$\boldsymbol{\mathcal{E}} = \mathcal{E}_\infty \, \mathbf{E}, \tag{2.1}$$

$$N_e = \mathcal{E}_\infty \, n_e, \tag{2.2}$$

$$N_p = \mathcal{E}_\infty \, n_p, \tag{2.3}$$

$$\tau = \frac{t}{\mathcal{E}_\infty}, \tag{2.4}$$

so that we have

$$\frac{\partial n_e}{\partial t} - \nabla \cdot (n_e \mathbf{E} + D \, \nabla n_e) = n_e |\mathbf{E}| e^{-1/(\mathcal{E}_\infty |\mathbf{E}|)}, \tag{2.5}$$

$$\frac{\partial n_p}{\partial t} = n_e |\mathbf{E}| e^{-1/(\mathcal{E}_\infty |\mathbf{E}|)}, \tag{2.6}$$

$$\nabla \cdot \mathbf{E} = n_p - n_e, \tag{2.7}$$

where

$$D = \frac{D_e}{\mathcal{E}_\infty} \tag{2.8}$$

is, in general, a small parameter. For $\mathcal{E}_\infty \gg 1$, this system can be approximated by

$$\frac{\partial n_e}{\partial t} - \nabla \cdot (n_e \mathbf{E} + D \, \nabla n_e) = n_e |\mathbf{E}|, \tag{2.9}$$

$$\frac{\partial n_p}{\partial t} = n_e |\mathbf{E}|, \tag{2.10}$$

$$\nabla \cdot \mathbf{E} = n_p - n_e. \tag{2.11}$$

Our approximation will be valid in all regions where $\mathcal{E}_\infty |\mathbf{E}| \gg 1$. These are the regions of interest in the situations studied in this paper since by (2.11) the intensity of the electric field varies continuously as long as n_e and n_p are bounded, and hence should not vary much in the neighborhood of the ionization front. We will show that this is indeed the case and it is in this region that the mechanisms leading to branching occur.

3. Planar fronts.

We will concentrate on the planar case. Experimentally, this means that we have two large planar plates situated at $x = 0$ (cathode) and $x = d$ (anode), respectively (x is the horizontal axis and we suppose that $d \gg 1$). The space between the plates is filled with a nonattaching gas such as nitrogen. A stationary electric potential difference is applied to the plates so that an electric field is produced in the gas. The initial electric field is directed from the anode to the cathode and is uniform in the space between the plates with a value $\mathcal{E}_\infty \gg 1$. As in this section we are interested in the evolution of the ionization wave along the x axis, the rescaled electric field can be written as $\mathbf{E} = E\mathbf{u}_x$, where $E < 0$ so that $|\mathbf{E}| = |E| = -E$, and $\mathbf{u}_x$ is a unitary vector in the x direction. We are left then with the following system:

$$\frac{\partial n_e}{\partial t} = \frac{\partial}{\partial x}\left(n_e E + D \frac{\partial n_e}{\partial x} \right) + n_e |E|, \tag{3.1}$$

$$\frac{\partial n_p}{\partial t} = n_e |E|, \tag{3.2}$$

$$\frac{\partial E}{\partial x} = n_p - n_e. \tag{3.3}$$

3.1. The traveling waves with $D = 0$.

It is very simple to compute traveling wave solutions when $D = 0$. In this case, the equation for the evolution of the electron density is

$$\frac{\partial n_e}{\partial t} = \frac{\partial (n_e E)}{\partial x} - n_e E. \tag{3.4}$$

Subtracting (3.1) from (3.2) with $D = 0$, and taking the time derivative of (3.3), we obtain the equation

$$\frac{\partial^2 E}{\partial x \partial t} + \frac{\partial}{\partial x}(n_e E) = 0. \tag{3.5}$$

Integrating this expression once in x, one obtains

$$\frac{\partial E}{\partial t} + n_e E = C(t), \tag{3.6}$$

where $C(t)$ can be fixed by the boundary conditions at infinity, $E \to -1$ and $n_e \to 0$. This implies $C(t) = 0$ so that

$$\frac{\partial E}{\partial t} = -n_e E. \tag{3.7}$$

In physical terms, the left-hand side of (3.6), due to Ampére's law, is the curl of the magnetic field which is zero because the magnetic effects are neglected in the framework of the minimal model.

We look for traveling wave solutions of the system (3.4)–(3.7), introducing the ansatz

$$n_e = f(x-ct), \; E = -g(x-ct) \tag{3.8}$$

into the above system. The minus sign in the electric field is due to the fact that the electric field is negative, so g is a positive function. Introducing (3.8) into (3.4) and (3.7), we obtain

$$-c\frac{df}{d\xi} = \frac{d}{d\xi}(fg) + fg, \tag{3.9}$$

$$c\frac{dg}{d\xi} = fg. \tag{3.10}$$

Introducing $dg/d\xi$ given by (3.10) into (3.9), we obtain an equation for $df/d\xi$, and hence we obtain the following system of ODEs:

$$\frac{df}{d\xi} = \frac{-fg + \frac{1}{c}f^2 g}{c-g}, \tag{3.11}$$

$$\frac{dg}{d\xi} = \frac{1}{c}fg, \tag{3.12}$$

where $\xi = x - ct$. This system can be explicitly solved by noticing that

$$\frac{df}{dg} = -\frac{c-f}{c-g} \tag{3.13}$$

so that

$$(c-f)(c-g) = c(c-1), \tag{3.14}$$

with the constant $c(c-1)$ being given by conditions at $\xi \to \infty$, namely, that the electron density vanishes and the electric field is equal to -1 there. Therefore,

$$\frac{dg}{d\xi} = \frac{g(1-g)}{c-g}, \tag{3.15}$$

allowing direct integration to yield the implicit solution (up to translations in ξ),

$$c\log g + (1-c)\log(1-g) = \xi. \tag{3.16}$$

This expression yields solutions for any $c \geq 1$. We will be interested in the limit $c \to 1$ since it is well known [6] that compactly supported initial data (representing a seed

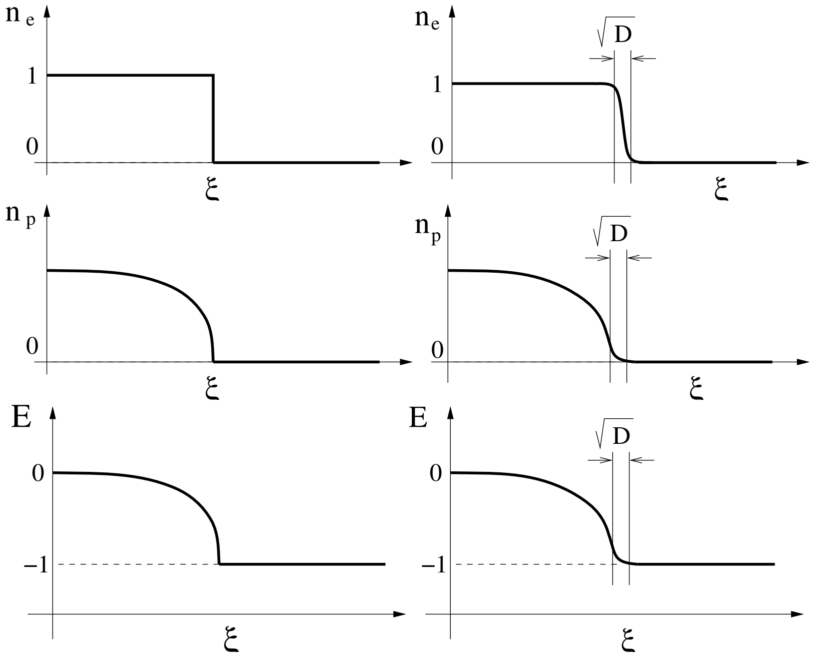


FIG. 1. *The moving fronts with $D = 0$ and $c = 1$. The moving fronts when $0 < D \ll 1$ and $c = 1 + 2\sqrt{D}$.*

of ionization located in some region) develop fronts traveling with this velocity. In the case $c = 1$ the solution can be obtained straightforwardly, giving

$$g(\xi) = \begin{cases} e^{\xi}, & \text{for } \xi < 0, \\ 1, & \text{for } \xi \geq 0, \end{cases} \quad f(\xi) = \begin{cases} 1, & \text{for } \xi < 0, \\ 0, & \text{for } \xi \geq 0. \end{cases} \tag{3.17}$$

We can also find the solution for the positive ion density n_p in the case $c = 1$. Taking $n_p = h(x - t)$, we have

$$h(\xi) = \begin{cases} 1 - e^{\xi} \text{ for } \xi < 0, \\ 0 \text{ for } \xi \geq 0. \end{cases} \tag{3.18}$$

This solution for n_e represents a shock front moving with velocity $c = 1$ (see Figure 1).

3.2. The traveling waves with $D \neq 0$. We proceed now to investigate the traveling waves for $0 < D \ll 1$. As D is a small parameter, the traveling wave solutions for the electron and positive ion densities and the electric field are expected to be not very different to that corresponding to $D = 0$ found in the previous subsection. Consequently, we look for solutions such that n_e and n_p decay exponentially at infinity and E is also an exponentially small correction of -1 at infinity. This means that we can take

$$n_e = Ae^{-\lambda(x-ct)}, \tag{3.19}$$

$$n_p = Be^{-\lambda(x-ct)}, \tag{3.20}$$

$$E = -1 + Ce^{-\lambda(x-ct)} \tag{3.21}$$

asymptotically far behind the wave. If we introduce these expressions into (3.1) we obtain, for $x - ct \to \infty$, the relation

$$-c\lambda + \lambda + D\lambda^2 = -1, \tag{3.22}$$

which has real solutions if and only if $(c-1)^2 - 4D \geq 0$. Therefore,

$$c \geq 1 + 2\sqrt{D}. \tag{3.23}$$

All initial data decaying at infinity faster than $Ae^{-\lambda^* x}$, with $\lambda^* = 1/\sqrt{D}$, will develop traveling waves [12] with velocity $c = 1 + 2\sqrt{D}$. If $D \ll 1$, the profiles for n_p and E will vary very little from the profiles with $D = 0$. On the other hand, n_e will develop a boundary layer at the front, smoothing the jump from $n_e = 1$ to $n_e = 0$. If we write the equation for the traveling wave $n_e = f(x - (1 + 2\sqrt{D})t)$ into the expression

$$\frac{\partial n_e}{\partial t} - n_e \frac{\partial E}{\partial x} - E\frac{\partial n_e}{\partial x} - D\frac{\partial^2 n_e}{\partial x^2} = n_e|E|, \tag{3.24}$$

and we take, from (3.3), $\partial_x E = n_p - n_e$, approximating at the boundary layer $n_p = 0$, $E = -1$, we obtain the equation

$$-2\sqrt{D}\frac{\partial f}{\partial \xi} - D\frac{\partial^2 f}{\partial \xi^2} = f(1-f), \tag{3.25}$$

where $\xi = x - (1 + 2\sqrt{D})t$. Defining $\chi = \xi/\sqrt{D}$, we obtain an equation for the boundary layer,

$$-2\frac{\partial f}{\partial \chi} - \frac{\partial^2 f}{\partial \chi^2} = f(1-f), \tag{3.26}$$

together with the matching conditions,

$$f(-\infty) = 1, \; f(+\infty) = 0. \tag{3.27}$$

Expression (3.26) is the well-known equation for traveling waves of Fisher's equation. It appears in the context of mathematical biology [16] and is known to have solutions subject to (3.27). This means that we have a boundary layer of width $\sqrt{D}$ at $\xi = 0$ in which (3.26) gives the solution for the electron density n_e. Before this layer, we have $n_e \approx 1$, and after the layer, $n_e \approx 0$. When $D = 0$, this is the shock front found in the previous subsection. It will be useful to analyze the structure of n_p at the boundary layer. Introducing

$$n_p = \sqrt{D}h(\chi), \tag{3.28}$$

one obtains from (3.2) the following formula at zero order in D, with $\chi = [x - (1 + 2\sqrt{D})t]/\sqrt{D}$:

$$\frac{dh(\chi)}{d\chi} = f(\chi) \tag{3.29}$$

so that

$$h(\chi) = -\int_{\chi}^{\infty} f(z)dz. \tag{3.30}$$

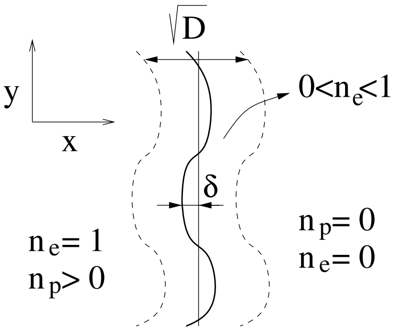


FIG. 2. *Schematic representation of the perturbed front.*

Notice that we now have

$$\frac{\partial n_p}{\partial x} = f(\chi) = O(1) \text{ at the boundary layer.} \tag{3.31}$$

Analogously, from Poisson's equation $\partial_x E = n_p - n_e$, we can deduce $E = -1 + O(\sqrt{D})$ across the boundary layer. We will write this solution as

$$E = -1 + \sqrt{D} E_{bl} + O(D). \tag{3.32}$$

These solutions can be seen in Figure 1.

4. The dispersion relation.

The planar front studied in the previous sections may be unstable with respect to perturbations on the boundary layer, which then forms "ripples" or "corrugations." Consequently, we are interested in obtaining the dispersion relation to find which transversal mode will grow faster and eventually determine the characteristic shape of the streamer. So we let the planar front that propagates in the x-direction receive a small perturbation with an initial arbitrary dependence on the transversal coordinates.

Next we introduce a perturbation in the transversal direction y (see Figure 2). We will do it by introducing a new system of coordinates in the form

$$\bar{t} = t, \tag{4.1}$$

$$\overline{y} = y, \tag{4.2}$$

$$\overline{x} = x - \delta\,\varphi(x, y, t) \tag{4.3}$$

so that, at $t = 0$, $n_e^{(0)}(\overline{x})$ and $E^{(0)}(\overline{x})$ correspond to the profiles of a traveling wave computed in the previous section, and δ is a sufficiently small parameter compared to $\sqrt{D}$. By doing this, we follow a strategy analogous to the one used in [19] to deduce the asymptotic approximation of the Allen–Cahn equation by mean curvature flow.

We can compute straightforwardly the relations between derivatives up to order δ^2,

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \overline{x}} - \delta\, \frac{\partial \varphi}{\partial x} \frac{\partial}{\partial \overline{x}}, \tag{4.4}$$

$$\frac{\partial}{\partial y} = \frac{\partial}{\partial \overline{y}} - \delta\, \frac{\partial \varphi}{\partial y} \frac{\partial}{\partial \overline{x}}, \tag{4.5}$$

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial \overline{t}} - \delta\, \frac{\partial \varphi}{\partial t} \frac{\partial}{\partial \overline{x}}, \tag{4.6}$$

$$\frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial \overline{x}^2} - \delta\, \frac{\partial^2 \varphi}{\partial x^2} \frac{\partial}{\partial \overline{x}} - 2\delta\, \frac{\partial \varphi}{\partial x} \frac{\partial^2}{\partial \overline{x}^2} + O(\delta^2), \tag{4.7}$$

$$\frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial \overline{y}^2} - \delta\, \frac{\partial^2 \varphi}{\partial y^2} \frac{\partial}{\partial \overline{x}} - 2\delta\, \frac{\partial \varphi}{\partial y} \frac{\partial^2}{\partial \overline{x} \partial \overline{y}} + O(\delta^2). \tag{4.8}$$

We introduce the perturbed electric field and electron density as

$$\mathbf{E} = E^{(0)}\, \mathbf{u}_x + \delta \left(E_x^{(1)}\, \mathbf{u}_x + E_y^{(1)} \mathbf{u}_y \right), \tag{4.9}$$

$$n_e = n_e^{(0)} + \delta\, n_e^{(1)}, \tag{4.10}$$

$$n_p = n_p^{(0)} + \delta\, n_p^{(1)}. \tag{4.11}$$

For example, the perturbed electron density (4.10) reads

$$n_e = n_e^{(0)} \left(x - \delta \varphi(x, y, t) \right) + \delta\, n_e^{(1)}, \tag{4.12}$$

where $\varphi(x, y, t)$ is a purely geometrical perturbation and $\delta\, n_e^{(1)}$ is a perturbation of the electron density behind the front. Note that, in this sense, the kind of perturbation we are introducing is similar to the perturbations that are usually introduced in the study of the stability of other pattern-forming systems. This is the case, for instance, of the propagation of solidification fronts [13]. The difference here with respect to those other systems is the fact that our interface is not sharp but a boundary layer of thickness $\sqrt{D}$.

We shall assume here that $\varphi(x, y, 0)$ is an initial perturbation independent of x. Note, at this point, that an x-dependence of the perturbation to the electron density is allowed in the term $n_e^{(1)}$.

4.1. Equations for the corrections at first order. Inserting these expressions into (2.9), we obtain

$$\begin{aligned}
\frac{\partial n_e^{(0)}}{\partial \overline{t}} - E^{(0)} \frac{\partial n_e^{(0)}}{\partial \overline{x}} &= n_e^{(0)} |E^{(0)}| + n_e^{(0)} \left(n_p^{(0)} - n_e^{(0)} \right) \\
&\quad + D \left(1 - 2\delta\, \frac{\partial \varphi}{\partial x} \right) \frac{\partial^2 n_e^{(0)}}{\partial \overline{x}^2} \\
&\quad + \delta \left[\frac{\partial \varphi}{\partial t} + E_x^{(1)} - E^{(0)} \frac{\partial \varphi}{\partial x} - D\, \Delta_{(x,y)} \varphi \right] \frac{\partial n_e^{(0)}}{\partial \overline{x}} \\
&\quad + \delta \left(|E_x^{(1)}| + \left(n_p^{(1)} - n_e^{(1)} \right) \right) n_e^{(0)} \\
&\quad + \delta \left(- \frac{\partial n_e^{(1)}}{\partial \overline{t}} + n_e^{(1)} \left(n_p^{(0)} - n_e^{(0)} \right) + E^{(0)} \frac{\partial n_e^{(1)}}{\partial \overline{x}} + n_e^{(1)} |E^{(0)}| \right) \\
&\quad + \delta\, D\, \Delta_{(\overline{x}, \overline{y})} n_e^{(1)} + O(\delta^2),
\end{aligned} \tag{4.13}$$

where $\Delta_{(x,y)} = \partial^2/\partial x^2 + \partial^2/\partial y^2$ and $\Delta_{(\overline{x},\overline{y})} = \partial^2/\partial \overline{x}^2 + \partial^2/\partial \overline{y}^2$. From (2.10) we obtain

$$\frac{\partial n_p^{(0)}}{\partial \overline{t}} + \delta\,\frac{\partial n_p^{(1)}}{\partial \overline{t}} = n_e^{(0)}|E^{(0)}| + \delta\,\frac{\partial \varphi}{\partial t}\frac{\partial n_p^{(0)}}{\partial \overline{x}} + \delta\,|E_x^{(1)}|n_e^{(0)} + \delta\,|E^{(0)}|n_e^{(1)} + O(\delta^2), \tag{4.14}$$

and from (2.11),

$$\frac{\partial E^{(0)}}{\partial \overline{x}} + \delta\left(\frac{\partial E_x^{(1)}}{\partial \overline{x}} + \frac{\partial E_y^{(1)}}{\partial \overline{y}}\right) = n_p^{(0)} - n_e^{(0)} + \delta\left(n_p^{(1)} - n_e^{(1)}\right) + \delta\,\frac{\partial \varphi}{\partial x}\frac{\partial E^{(0)}}{\partial \overline{x}} + O(\delta^2). \tag{4.15}$$

We can construct a solution up to $O(\delta^2)$ by imposing that $O(\delta^0)$ terms and $O(\delta^1)$ terms in (4.13), (4.14), and (4.15) vanish. The $O(\delta^0)$ terms give

$$\frac{\partial n_e^{(0)}}{\partial \overline{t}} = E^{(0)}\frac{\partial n_e^{(0)}}{\partial \overline{x}} + n_e^{(0)}|E^{(0)}| + n_e^{(0)}\left(n_p^{(0)} - n_e^{(0)}\right) + D\,\frac{\partial^2 n_e^{(0)}}{\partial \overline{x}^2}, \tag{4.16}$$

$$\frac{\partial n_p^{(0)}}{\partial \overline{t}} = n_e^{(0)}|E^{(0)}|, \tag{4.17}$$

$$\frac{\partial E^{(0)}}{\partial \overline{x}} = n_p^{(0)} - n_e^{(0)}, \tag{4.18}$$

and the $O(\delta^1)$ terms give

$$\begin{aligned} 0 = {} & \left[\frac{\partial \varphi}{\partial \overline{t}} + E_x^{(1)} - E^{(0)}\frac{\partial \varphi}{\partial \overline{x}} - D\,\Delta_{(\overline{x},\overline{y})}\varphi\right]\frac{\partial n_e^{(0)}}{\partial \overline{x}} \\ & -2D\,\frac{\partial \varphi}{\partial \overline{x}}\frac{\partial^2 n_e^{(0)}}{\partial \overline{x}^2} + \left(|E_x^{(1)}| + n_p^{(1)} - n_e^{(1)}\right)n_e^{(0)} \\ & -\frac{\partial n_e^{(1)}}{\partial \overline{t}} + n_e^{(1)}\left(n_p^{(0)} - n_e^{(0)}\right) + E^{(0)}\frac{\partial n_e^{(1)}}{\partial \overline{x}} + n_e^{(1)}|E^{(0)}| \\ & +D\,\Delta_{(\overline{x},\overline{y})}n_e^{(1)}, \end{aligned} \tag{4.19}$$

$$0 = \frac{\partial n_p^{(1)}}{\partial \overline{t}} + \frac{1}{1+2\sqrt{D}}\frac{\partial \varphi}{\partial \overline{t}}n_e^{(0)} - |E_x^{(1)}|n_e^{(0)} - |E^{(0)}|n_e^{(1)}, \tag{4.20}$$

$$0 = \frac{\partial E_x^{(1)}}{\partial \overline{x}} + \frac{\partial E_y^{(1)}}{\partial \overline{y}} - \left(n_p^{(1)} - n_e^{(1)}\right) - \frac{\partial \varphi}{\partial \overline{x}}\left(n_p^{(0)} - n_e^{(0)}\right), \tag{4.21}$$

in which we have replaced, at order δ, derivatives with respect to x by derivatives with respect to $\overline{x}$, used (4.18) to replace $\partial E^{(0)}/\partial \overline{x}$ by $n_p^{(0)} - n_e^{(0)}$, and used (4.17), (3.29), and (3.30) to replace $\partial n_p^{(0)}/\partial \overline{x}$ by

$$\frac{\partial n_p^{(0)}}{\partial \overline{x}} = \frac{-1}{1+2\sqrt{D}}\frac{\partial n_p^{(0)}}{\partial \overline{t}} = \frac{-1}{1+2\sqrt{D}}n_e^{(0)}. \tag{4.22}$$

The solution of the system given by (4.16), (4.17), and (4.18) is the traveling wave found in the previous section so that

$$n_e^{(0)} = f(\overline{\xi}), \tag{4.23}$$

where $\overline{\xi} = \overline{x} - c\overline{t}$.

In order to analyze the system (4.19)–(4.21), we introduce changes of coordinates in two stages: first, we change coordinates into a frame in which the planar front remains stationary and, second, we rescale coordinates in the boundary layer in order to make it of $O(1)$ size.

The first change of coordinates is of the form

$$x' = \overline{x} - c\overline{t},\ y' = \overline{y},\ t' = \overline{t}, \tag{4.24}$$

where $c = 1 + 2\sqrt{D}$. Hence, the system (4.19)–(4.21) transforms into

$$\begin{aligned} 0 = &\left[\frac{\partial \varphi}{\partial t'} + E_x^{(1)} - D\,\Delta_{(x',y')}\varphi\right]\frac{\partial n_e^{(0)}}{\partial x'} - (E^{(0)}+c)\frac{\partial \varphi}{\partial x'}\frac{\partial n_e^{(0)}}{\partial x'} \\ &-2D\,\frac{\partial \varphi}{\partial x'}\frac{\partial^2 n_e^{(0)}}{\partial x'^2} + \left(|E_x^{(1)}| + n_p^{(1)}\right)n_e^{(0)} - \frac{\partial n_e^{(1)}}{\partial t'} + (E^{(0)}+c)\frac{\partial n_e^{(1)}}{\partial x'} \\ &+\left(|E^{(0)}| + n_p^{(0)} - 2n_e^{(0)}\right)n_e^{(1)} + D\,\Delta_{(x',y')}n_e^{(1)}, \end{aligned} \tag{4.25}$$

$$\begin{aligned} 0 = &\left(\frac{\partial}{\partial t'} - c\frac{\partial}{\partial x'}\right)n_p^{(1)} + \frac{1}{1+2\sqrt{D}}\left[\left(\frac{\partial}{\partial t'} - c\frac{\partial}{\partial x'}\right)\varphi\right]n_e^{(0)} \\ &-|E_x^{(1)}|n_e^{(0)} - |E^{(0)}|n_e^{(1)}, \end{aligned} \tag{4.26}$$

$$0 = \frac{\partial E_x^{(1)}}{\partial x'} + \frac{\partial E_y^{(1)}}{\partial y'} - \left(n_p^{(1)} - n_e^{(1)}\right) - \frac{\partial \varphi}{\partial x'}\left(n_p^{(0)} - n_e^{(0)}\right). \tag{4.27}$$

Second, noticing that x' is of order $\sqrt{D}$ at the boundary layer, as we saw in the previous section, we write

$$x' = \sqrt{D}\,\tilde{x},\ y' = \sqrt{D}\,\tilde{y},\ t' = \tilde{t} \tag{4.28}$$

to obtain the rescaled system

$$\begin{aligned} 0 = &\left[\frac{\partial \varphi}{\partial \tilde{t}} + E_x^{(1)} - \Delta_{(\tilde{x},\tilde{y})}\varphi\right]\frac{\partial n_e^{(0)}}{\partial \tilde{x}} - 2\frac{\partial \varphi}{\partial \tilde{x}}\frac{\partial^2 n_e^{(0)}}{\partial \tilde{x}^2} \\ &-(E_{bl}+2)\frac{\partial \varphi}{\partial \tilde{x}}\frac{\partial n_e^{(0)}}{\partial \tilde{x}} + \sqrt{D}\left(|E_x^{(1)}| + n_p^{(1)}\right)n_e^{(0)} \\ &-\sqrt{D}\frac{\partial n_e^{(1)}}{\partial \tilde{t}} + \sqrt{D}(E_{bl}+2)\frac{\partial n_e^{(1)}}{\partial \tilde{x}} \\ &+\sqrt{D}\left(|E^{(0)}| + n_p^{(0)} - 2n_e^{(0)}\right)n_e^{(1)} + \sqrt{D}\,\Delta_{(\tilde{x},\tilde{y})}n_e^{(1)}, \end{aligned} \tag{4.29}$$

$$\begin{aligned} 0 = &\left(\frac{\partial}{\partial \tilde{t}} - \frac{c}{\sqrt{D}}\frac{\partial}{\partial \tilde{x}}\right)n_p^{(1)} \\ &+\frac{1}{1+2\sqrt{D}}\left[\left(\frac{\partial}{\partial \tilde{t}} - \frac{c}{\sqrt{D}}\frac{\partial}{\partial \tilde{x}}\right)\varphi\right]n_e^{(0)} - |E_x^{(1)}|n_e^{(0)} - |E^{(0)}|n_e^{(1)}, \end{aligned} \tag{4.30}$$

$$0 = \frac{\partial E_x^{(1)}}{\partial \tilde{x}} + \frac{\partial E_y^{(1)}}{\partial \tilde{y}} - \sqrt{D}\left(n_p^{(1)} - n_e^{(1)}\right) - \frac{\partial \varphi}{\partial \tilde{x}}\left(n_p^{(0)} - n_e^{(0)}\right), \tag{4.31}$$

where we have used that, at the boundary layer, by (3.32), $(E^{(0)} + c)/\sqrt{D} = E_{bl} + 2 + O(\sqrt{D})$.

The terms in (4.29) involving $n_e^{(1)}$ lead to a PDE for $n_e^{(1)}$. Namely,

$$\frac{\partial n_e^{(1)}}{\partial \tilde{t}} - (E_{bl}+2)\frac{\partial n_e^{(1)}}{\partial \tilde{x}} - \Delta_{(\tilde{x},\tilde{y})}n_e^{(1)} = \left(|E^{(0)}| + n_p^{(0)} - 2n_e^{(0)}\right)n_e^{(1)}. \tag{4.32}$$

Notice that (4.32) is an advection-diffusion equation with a source term of the form $(|E^{(0)}| + n_p^{(0)} - 2n_e^{(0)})\,n_e^{(1)}$. Since the source term is, from the expression for the traveling waves found in subsection 3.2, negative behind the front, $n_e^{(1)}$ will decay exponentially fast, provided it lays in the ionized region, which is a basic assumption for our perturbation.

Hence, at leading order in (4.29), when $D \ll 1$, one obtains the equation

$$0 = \left[\frac{\partial\varphi}{\partial\tilde{t}} + E_x^{(1)} - \Delta_{(\tilde{x},\tilde{y})}\varphi\right]\frac{\partial n_e^{(0)}}{\partial\tilde{x}} - 2\frac{\partial\varphi}{\partial\tilde{x}}\frac{\partial^2 n_e^{(0)}}{\partial\tilde{x}^2} - (E_{bl}+2)\frac{\partial\varphi}{\partial\tilde{x}}\frac{\partial n_e^{(0)}}{\partial\tilde{x}}. \tag{4.33}$$

Equation (4.31) is, at leading order in D,

$$0 = \frac{\partial E_x^{(1)}}{\partial\tilde{x}} + \frac{\partial E_y^{(1)}}{\partial\tilde{y}} - \frac{\partial\varphi}{\partial\tilde{x}}\left(n_p^{(0)} - n_e^{(0)}\right), \tag{4.34}$$

so that (4.33) and (4.34) are independent of $n_p^{(1)}$, and we can describe the evolution of the perturbed system as

$$0 = \frac{\partial\varphi}{\partial\tilde{t}} + E_x^{(1)} - \Delta_{(\tilde{x},\tilde{y})}\varphi - 2\frac{\partial\varphi}{\partial\tilde{x}}\frac{\partial^2 n_e^{(0)}/\partial\tilde{x}^2}{\partial n_e^{(0)}/\partial\tilde{x}} - (E_{bl}+2)\frac{\partial\varphi}{\partial\tilde{x}}, \tag{4.35}$$

$$0 = \frac{\partial E_x^{(1)}}{\partial\tilde{x}} + \frac{\partial E_y^{(1)}}{\partial\tilde{y}} - \frac{\partial\varphi}{\partial\tilde{x}}\left(n_p^{(0)} - n_e^{(0)}\right). \tag{4.36}$$

It will be more convenient for us to formulate (4.36) in terms of the electric potential in the next subsection.

4.2. The first order correction to the electric field. To establish conditions for the behavior of the perturbation of the electric field, we first note that the total electric field has to be irrotational since the magnetic field is negligible. So we will write $\mathbf{E} = -\nabla V$, where V is an electric potential. Then, (2.11) can be written as

$$-\Delta_{(x,y)}V = n_p^{(0)} - n_e^{(0)} + \delta\, n_p^{(1)} + O(\delta^2). \tag{4.37}$$

Changing variables, we have

$$\begin{aligned} &- \Delta_{(\overline{x},\overline{y})}V + \delta\,\Delta_{(x,y)}\varphi\frac{\partial V}{\partial\overline{x}} + 2\delta\left(\frac{\partial\varphi}{\partial\overline{y}}\frac{\partial^2 V}{\partial\overline{x}\partial\overline{y}} + \frac{\partial\varphi}{\partial\overline{x}}\frac{\partial^2 V}{\partial\overline{x}^2}\right)\\ &\qquad = n_p^{(0)} - n_e^{(0)} + \delta\, n_p^{(1)} + O(\delta^2). \end{aligned} \tag{4.38}$$

We write the electric potential as

$$V(\overline{x},\overline{y}) = V^{(0)}(\overline{x}) + \delta\, V^{(1)}(\overline{x},\overline{y}), \tag{4.39}$$

so that (4.38) can be written, at the first two orders in δ, as

$$-\frac{\partial^2 V^{(0)}(\overline{x})}{\partial\overline{x}^2} = n_p^{(0)} - n_e^{(0)}, \tag{4.40}$$

$$-\Delta_{(\overline{x},\overline{y})}V^{(1)}(\overline{x},\overline{y}) = -\Delta_{(\overline{x},\overline{y})}\varphi\frac{\partial V^{(0)}(\overline{x})}{\partial\overline{x}} - 2\frac{\partial\varphi}{\partial\overline{x}}\frac{\partial^2 V^{(0)}(\overline{x})}{\partial\overline{x}^2} + n_p^{(1)}. \tag{4.41}$$

Expression (4.40) implies that $V^{(0)}(\overline{x})$ is an electric potential associated with the electric field $E^{(0)}(\overline{x})$. The electric potential $V^{(1)}$ satisfies (4.41) with the condition of decaying at $|\overline{x}| \to \infty$. Using (4.40) and (4.41), and the relation $E^{(0)} = -\partial V^{(0)}/\partial \overline{x}$, we arrive at

$$-\Delta_{(\overline{x},\overline{y})} V^{(1)}(\overline{x},\overline{y}) = \Delta_{(\overline{x},\overline{y})}\varphi E^{(0)} + 2\frac{\partial \varphi}{\partial \overline{x}}\left(n_p^{(0)} - n_e^{(0)}\right) + n_p^{(1)}. \tag{4.42}$$

Changing coordinates as in (4.24) and (4.28), in terms of $(\tilde{x},\tilde{y})$ coordinates that are $O(1)$ at the diffusion boundary layer, we obtain

$$\begin{aligned} - \Delta_{(\tilde{x},\tilde{y})} V^{(1)} &= \left(\Delta_{(\tilde{x},\tilde{y})}\varphi\right) E^{(0)}(\sqrt{D}\tilde{x}) \\ &+2\sqrt{D}\,\frac{\partial \varphi}{\partial \tilde{x}}\left(n_p^{(0)}(\sqrt{D}\tilde{x}) - n_e^{(0)}(\sqrt{D}\tilde{x})\right) + D\, n_p^{(1)}. \end{aligned} \tag{4.43}$$

Neglecting $O(\sqrt{D})$ and $O(D)$ terms, and using (4.35), we can finally describe the evolution of the perturbed system as

$$0 = \frac{\partial \varphi}{\partial \tilde{t}} + E_x^{(1)} - \Delta_{(\tilde{x},\tilde{y})}\varphi - 2\frac{\partial \varphi}{\partial \tilde{x}}\frac{\partial^2 n_e^{(0)}/\partial \tilde{x}^2}{\partial n_e^{(0)}/\partial \tilde{x}} - (E_{bl}+2)\frac{\partial \varphi}{\partial \tilde{x}}, \tag{4.44}$$

$$0 = \Delta_{(\tilde{x},\tilde{y})} V^{(1)} + \left(\Delta_{(\tilde{x},\tilde{y})}\varphi\right) E^{(0)}(\sqrt{D}\tilde{x}). \tag{4.45}$$

In the following subsections we shall analyze the system (4.44)–(4.45).

4.3. Analysis of the perturbed system. It proves convenient, since the system (4.44)–(4.45) is linear, to use Fourier transforms in the coordinate $\tilde{y}$ (associated with the wave number k). Denoting the Fourier transform of function f as $\hat{f}$, we find

$$0 = \frac{\partial \hat{\varphi}}{\partial \tilde{t}} + \hat{E}_x^{(1)} - \left(\frac{\partial^2 \hat{\varphi}}{\partial \tilde{x}^2} - k^2\hat{\varphi}\right) - 2\frac{\partial \hat{\varphi}}{\partial \tilde{x}}\frac{\partial^2 n_e^{(0)}/\partial \tilde{x}^2}{\partial n_e^{(0)}/\partial \tilde{x}} - (E_{bl}+2)\frac{\partial \hat{\varphi}}{\partial \tilde{x}}, \tag{4.46}$$

$$0 = \frac{\partial^2 \hat{V}^{(1)}}{\partial \tilde{x}^2} - k^2\hat{V}^{(1)} + \left(\frac{\partial^2 \hat{\varphi}}{\partial \tilde{x}^2} - k^2\hat{\varphi}\right) E^{(0)}(\sqrt{D}\tilde{x}), \tag{4.47}$$

a linear system that can be represented symbolically by

$$\mathcal{L}(\hat{\varphi}, \hat{V}^{(1)}) = 0. \tag{4.48}$$

In principle, once $\hat{V}^{(1)}$ has been calculated from (4.47), we can calculate $\hat{E}_x^{(1)}$ as

$$\frac{\partial \hat{V}^{(1)}}{\partial \overline{x}} = -\hat{E}_x^{(1)}, \tag{4.49}$$

insert it into (4.46), and obtain an equation for $\hat{\varphi}$. Since $\hat{\varphi}$ is initially independent of $\tilde{x}$, we shall write

$$\hat{\varphi} = \hat{\varphi}_0(k,t) + \hat{\varphi}_1(x,k,t), \tag{4.50}$$

with $\hat{\varphi}_1(x,k,t=0) = 0$ and, accordingly,

$$\hat{V}^{(1)} = \hat{V}_0^{(1)}(x,k,t) + \hat{V}_1^{(1)}(x,k,t). \tag{4.51}$$

We will require the following equations to be fulfilled: (i) for $\hat{\varphi}_0$ and $\hat{V}_0^{(1)}$,

$$0 = \frac{\partial \hat{\varphi}_0}{\partial \tilde{t}} + \hat{E}_{0x}^{(1)}(\tilde{x}=0) + k^2\hat{\varphi}_0, \tag{4.52}$$

$$0 = \frac{\partial^2 \hat{V}_0^{(1)}}{\partial \tilde{x}^2} - k^2\hat{V}_0^{(1)} - k^2\hat{\varphi}_0 E^{(0)}(\sqrt{D}\tilde{x}), \tag{4.53}$$

where

$$\hat{E}^{(1)}_{0x} = -\frac{\partial \hat{V}^{(1)}_0}{\partial \overline{x}}; \tag{4.54}$$

and (ii)

$$\mathcal{L}(\hat{\varphi}_1, \hat{V}^{(1)}_1) = \begin{pmatrix} \hat{E}^{(1)}_{0x}(\tilde{x}=0,k,t) - \hat{E}^{(1)}_{0x}(\tilde{x},k,t) \\ 0 \end{pmatrix} \tag{4.55}$$

for $\hat{\varphi}_1$ and $\hat{V}^{(1)}_1$. We shall solve first (4.52)–(4.53) and proceed later to show that the φ_1 solution of (4.55) is merely a small perturbation of φ_0.

If we take the derivative of (4.53) with respect to $\overline{x}$, taking into account (4.49) and the relation between $E^{(0)} = -\partial V^{(0)}/\partial \overline{x}$ and $n^{(0)}_p - n^{(0)}_e$ given by (4.40), we find

$$\frac{\partial^2 \hat{E}^{(1)}_{0x}}{\partial \tilde{x}^2} - k^2 \hat{E}^{(1)}_{0x} = -k^2 \hat{\varphi}_0 \left(n^{(0)}_p(\sqrt{D}\tilde{x}) - n^{(0)}_e(\sqrt{D}\tilde{x}) \right). \tag{4.56}$$

Taking the Fourier transform in $\tilde{x}$ (associated with the wave number ω) and denoting the double Fourier transform as $\hat{\hat{f}}$, one obtains

$$(k^2+\omega^2)\hat{\hat{E}}^{(1)}_{0x}(\omega,k) = \frac{k^2\hat{\varphi}_0(k)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ds\, e^{-i\omega s} q(\sqrt{D}s), \tag{4.57}$$

where we have defined the net charge density as $q(\sqrt{D}\tilde{x}) = n^{(0)}_p(\sqrt{D}\tilde{x}) - n^{(0)}_e(\sqrt{D}\tilde{x})$. Taking the inverse Fourier transform in ω of (4.57), it follows that

$$\begin{aligned} \hat{E}^{(1)}_{0x}(\tilde{x},k) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega\tilde{x}} \frac{k^2\hat{\varphi}_0(k)}{k^2+\omega^2} \int_{-\infty}^{\infty} ds\, e^{-i\omega s} q(\sqrt{D}s) \\ &= \frac{1}{2\pi} k^2 \hat{\varphi}_0(k) \int_{-\infty}^{\infty} ds\, q(\sqrt{D}s) \int_{-\infty}^{\infty} d\omega\, \frac{e^{i\omega(\tilde{x}-s)}}{k^2+\omega^2}. \end{aligned} \tag{4.58}$$

The integral in ω can be done by deforming the integration contour in the complex plane. The result is

$$\hat{E}^{(1)}_{0x}(\tilde{x},k) = \frac{|k|\hat{\varphi}_0(k)}{2} \int_{-\infty}^{\infty} ds\, q(\sqrt{D}s) e^{-|k|\,|\tilde{x}-s|}. \tag{4.59}$$

Since the value of $q = n^{(0)}_p - n^{(0)}_e$ in the case $D \ll 1$ differs from the same quantity in the case $D = 0$ only in the region of the boundary layer, that is, $O(D^{1/2})$, we can approximate the profile for the net charge density $q(\sqrt{D}\tilde{x})$ by the profile for the diffusionless traveling waves calculated in the previous section, i.e.,

$$q(\sqrt{D}\tilde{x}) = \begin{cases} -e^{\sqrt{D}\tilde{x}} \text{ for } \tilde{x} < 0, \\ \quad 0 \text{ for } \tilde{x} > 0. \end{cases} \tag{4.60}$$

With this approximation, (4.59) reads

$$\hat{E}^{(1)}_{0x}(\tilde{x},k) = -\frac{|k|\hat{\varphi}_0(k)}{2} \int_{-\infty}^{0} ds\, e^{\sqrt{D}s} e^{-|k|\,|\tilde{x}-s|}. \tag{4.61}$$

The integral in (4.61) can be computed for both $\tilde{x} < 0$ and $\tilde{x} > 0$. The result is

$$\hat{E}^{(1)}_{0x}(\tilde{x}, k) = -\frac{|k|\hat{\varphi}_0(k)}{2\sqrt{D}} \times \begin{cases} \frac{1}{1+|k|/\sqrt{D}} e^{-|k|\tilde{x}} \text{ for } \tilde{x} \geq 0, \\ \frac{-2|k|/\sqrt{D}}{1-|k|^2/D} e^{\sqrt{D}\tilde{x}} + \frac{1}{1-|k|/\sqrt{D}} e^{|k|\tilde{x}} \text{ for } \tilde{x} \leq 0. \end{cases} \tag{4.62}$$

Therefore,

$$\hat{E}^{(1)}_{0x}(0, k) = -\frac{|k|\hat{\varphi}_0(k)}{2(\sqrt{D} + |k|)}, \tag{4.63}$$

and

$$R \equiv |\hat{E}^{(1)}_{0x}(\tilde{x} = 0, k, t) - \hat{E}^{(1)}_{0x}(\tilde{x}, k, t)| = O(e^{-|k|\tilde{x}} - 1)|\hat{\varphi}_0|. \tag{4.64}$$

Notice that

$$R \leq |k\tilde{x}||\hat{\varphi}_0|, \tag{4.65}$$

a fact that we shall use below.

Let us remark that considering the profile of q with diffusion would change the integral in (4.61) by only an $O(1)$ amount, which is negligible in comparison with $1/(\sqrt{D} + |k|)$, provided $|k| \ll 1$.

4.4. The dispersion relation.

Inserting the result (4.63) into (4.52), we find

$$\frac{\partial \hat{\varphi}_0(k)}{\partial \tilde{t}} - \frac{|k|\hat{\varphi}_0(k)}{2(\sqrt{D} + |k|)} + |k|^2 \hat{\varphi}_0(k) = 0. \tag{4.66}$$

Let us write the following ansatz for $\hat{\varphi}_0$:

$$\hat{\varphi}_0(k, \tilde{t}) = e^{m\tilde{t}} \hat{\phi}(k). \tag{4.67}$$

Introducing this expression into (4.66), we obtain the relation

$$m = \frac{|k|}{2(\sqrt{D} + |k|)} - |k|^2, \tag{4.68}$$

that is, the dispersion relation of the perturbations of the fronts. Note that there exists a maximum of $m(|k|)$ that selects the wavelength of the perturbation. It is easy to obtain the following expansion (in D) for the location of the maximum of m:

$$k_{max} = \frac{1}{2^{2/3}} D^{1/6} - \frac{2}{3} D^{1/2} + \frac{2^{2/3}}{9} D^{5/6} + \frac{2^{7/3}}{81} D^{7/6} + O(D^{3/2}). \tag{4.69}$$

When D is a small parameter, this maximum is approximately located at

$$k_{max} \approx \frac{D^{1/6}}{2^{2/3}}. \tag{4.70}$$

Notice that k_{max} is $O(D^{1/6})$ so that in the front, and for the fastest growing mode, where $\tilde{x} = O(1)$, we will have R defined in (4.64) of order $O(D^{1/6})|\hat{\varphi}_0|$.

Hence the solution to (4.55), with initial condition $\hat{\varphi}_1(\tilde{x}, k, 0) = 0$, will be such that $|\hat{\varphi}_1| = O(D^{1/6})t$ and will constitute merely a small perturbation of the leading order $\hat{\varphi}_0$. In fact, $\hat{\varphi}_1$ will be a small perturbation of $\hat{\varphi}_0$ whenever $|k| \ll 1$ and not only in the neighborhood of the fastest growing mode.

The value of k_{max} in (4.70) corresponds to a typical spacing between fingers in the coordinate y given by

$$\lambda_{max} = \frac{2\pi D^{1/2}}{k_{max}} \approx 10\, D^{1/3}. \tag{4.71}$$

In the original nondimensional quantities, this is

$$\lambda_{max} \approx 10 \left(\frac{D_e}{\mathcal{E}_\infty}\right)^{1/3}. \tag{4.72}$$

The typical spacing can be put into physical quantities for nitrogen using the relations (1.9), (1.10), and (1.15). In this way, we can give the dependence of the physical spacing λ^d between consecutive fingers in terms of the gas pressure p (in bars), the physical external electric field $\mathcal{E}^d_\infty$, and the diffusion coefficient D^d_e. We obtain

$$\begin{aligned} \lambda^d_{max} &\approx 10 R_0 \left(\frac{\mathcal{E}_0}{D_0}\right)^{1/3} \left(\frac{D^d_e}{\mathcal{E}^d_\infty}\right)^{1/3} \\ &\approx 2.3 \times 10^{-5}\,\mathrm{m} \left(\frac{2 \times 10^7\,\mathrm{V} \cdot \mathit{bar}/\mathrm{m}}{1.8\,\mathrm{m}^2/\mathrm{s}}\right)^{1/3} \left(\frac{D^d_e}{p\,\mathcal{E}^d_\infty}\right)^{1/3} \end{aligned} \tag{4.73}$$

so that the spacing decreases as the gas pressure or the external electric field increases, and increases as the diffusion coefficient increases. This expression shows the possibility of validating the main results of this work through experiments of electric discharges in nitrogen.

5. Numerical studies of stability of planar fronts and nonplanar waves.

The theory developed in the previous sections applies solely to waves traveling at velocity $c = 1$ in the nondiffusive case and $c = 1 + 2\sqrt{D}$ when $D \neq 0$. These traveling waves appear only for a certain class of initial data, namely, those for which n_e is identically zero beyond a certain point in space. From the numerical point of view, solutions tend to develop traveling waves which do not propagate exactly with that velocity. Nevertheless, we will show in this numerical section that the main stability/instability features of our theoretical results remain valid in general. Specifically, we show the existence of traveling waves in the form of fingers when the diffusion coefficient is small enough, and show that for a given diffusion coefficient, stability of planar fronts depends critically on the wavelength of the perturbations.

We developed a numerical code to solve the initial value problem and study the evolution of nonplanar traveling waves. We discretized the equations with finite differences on a domain of size $L_x \times L_y$ with a uniform square grid of spacing h. For the temporal integration we used an improved Euler scheme. We first compute an approximation for the solution of the system (2.9)–(2.11) at $t + \delta t/2$ as

$$\Delta_a \phi^{(k)} = -\left(n_p - n_e\right)^{(k)}, \tag{5.1}$$

$$n_e^{(k+1/2)} = n_e^{(k)} + \frac{\delta t}{2}\left(\mathbf{E} \cdot \nabla_u n_e + n_e(n_p - n_e) + D + \Delta_a n_e + n_e|\mathbf{E}|\right)^{(k)}, \tag{5.2}$$

$$n_p^{(k+1/2)} = n_p^{(k)} + \frac{\delta t}{2}\left(n_e|\mathbf{E}|\right)^{(k)}, \tag{5.3}$$

and then we obtain a second order approximation by using the derivatives at the center of the interval $(t, t + \delta t)$,

$$\Delta_a \phi^{(k+1/2)} = -\left(n_p - n_e\right)^{(k+1/2)}, \tag{5.4}$$

$$n_e^{(k+1)} = n_e^{(k)} + \delta t \left(\mathbf{E} \cdot \nabla_c n_e + n_e(n_p - n_e) + D\Delta_a n_e + n_e|\mathbf{E}|\right)^{(k+1/2)}, \tag{5.5}$$

$$n_p^{(k+1)} = n_p^{(k)} + \delta t \left(n_e|\mathbf{E}|\right)^{(k+1/2)}, \tag{5.6}$$

where the superscript (k) denotes the time step at time $k\delta t$, $\mathbf{E} = -\nabla_c \phi$, and

$$\begin{aligned}\Delta_a \phi = \frac{1}{6h^2}[&\phi_{i+1,j+1} + \phi_{i+1,j-1} + \phi_{i-1,j+1} + \phi_{i-1,j-1} \\ &+4(\phi_{i+1,j}\phi_{i-1,j} + \phi_{i,j+1} + \phi_{i,j-1}) - 20\phi_{ij}]\end{aligned} \tag{5.7}$$

is the second order accurate approximation of the Laplacian that is symmetrical up to third order. In (5.2) and (5.5), ∇_u is the upwind gradient with respect to the electric field, and ∇_c is the centered second order accurate gradient. In order to solve the Poisson equations (5.1) and (5.4), we used successive overrelaxations (SOR) [3], which in our case is convenient because at each time step we have a good approximation of the solution from the previous step.

We found empirically that the scheme is stable, provided we satisfy the following CFL-like condition:

$$\delta t < \min(h/(2E_m), h^2/4D), \tag{5.8}$$

where E_m is the maximum value of the absolute value of the electric field in the domain of integration (which in our equations plays the role of velocity).

We enforce the following boundary conditions:

$$\frac{\partial \phi}{\partial x}(L_x, y) = 1, \qquad \phi(0, y) = 0, \qquad \frac{\partial \phi}{\partial y}(x, L_y) = \frac{\partial \phi}{\partial y}(x, 0) = 0, \tag{5.9}$$

$$n_e(L_x, y) = \frac{\partial n_e}{\partial x}(0, y) = 0, \qquad \frac{\partial n_e}{\partial y}(x, L_y) = \frac{\partial n_e}{\partial y}(x, 0) = 0, \tag{5.10}$$

$$n_p(L_x, y) = \frac{\partial n_p}{\partial x}(0, y) = 0, \qquad \frac{\partial n_p}{\partial y}(x, L_y) = \frac{\partial n_p}{\partial y}(x, 0) = 0, \tag{5.11}$$

which correspond to a constant electric field on the top end of the domain and zero-flux conditions on the sides.

5.1. Validation with traveling waves. We validated the scheme by comparing the numerical solution with the following exact solutions for the traveling waves without diffusion:

$$n_e^{ex}(\zeta) = 1 - \frac{e^\zeta}{\sqrt{e^\zeta(4 + e^\zeta)}}, \tag{5.12}$$

$$\begin{aligned}n_p^{ex}(\zeta) = 1 &+ \frac{e^\zeta}{2} - \frac{e^{\zeta/2}\sqrt{4 + e^\zeta}}{2} + \frac{\log 2}{2} \\ &- \log\left(e^{\zeta/2} + \sqrt{4 + e^\zeta}\right) + \frac{1}{2}\log\left(2 + e^\zeta + e^{\zeta/2}\sqrt{4 + e^\zeta}\right),\end{aligned} \tag{5.13}$$

where $\zeta = x - 2t$. This solution is convenient for the validation because it is smooth, and our numerical scheme is best suited to calculating differentiable solutions. We

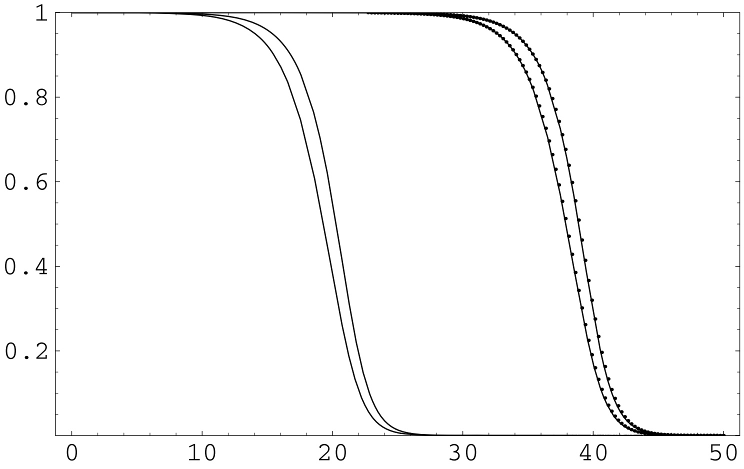


FIG. 3. *Validation with an explicit solution for $c = 2$ and $D = 0$. The size of the domain is $L_x = 50$ and we used 200 gridpoints. The curves on the left indicate the initial condition for n_e (upper line) and n_p (lower line), and the curves on the right show the comparison between the numerical calculation and the exact solution at $t = t_q$.*

first set as an initial condition the exact solution at $t = 0$ and then we compute the numerical solution at $t_q = 9.5$. In Figure 3 we show the comparison between both solutions. In Figure 4 we show the total error calculated as

$$\text{error} = \int_0^{L_x} \left(n_e(x, t_q) - n_e^{ex}(x, t_q)\right)^2 dx. \tag{5.14}$$

This measure of error takes into account the accumulation of all arithmetic and truncation errors on the time interval $(0, t_q)$. Figure 4 shows that the error is proportional to the square of the interspacing h, indicating that the scheme is second order accurate.

5.2. Computing two-dimensional traveling waves. One difficulty that arises with a finite computational domain is that traveling waves eventually arrive at the end of the domain of integration. This is a problem because, given an arbitrary initial condition, sometimes it takes a long time for traveling waves to converge to a steady state.

We solve this difficulty by making use of a displacement technique that keeps the waves near the center of the domain at all times. Each time that the position of the front of a wave (defined, for example, as the point where $n_e = 0.1$) is beyond the middle of the domain, we then translate the solution backwards by exactly one gridpoint,

$$n_{ei,j} \leftarrow n_{ei+1,j}, \qquad n_{p_{i,j}} \leftarrow n_{p_{i+1,j}}. \tag{5.15}$$

At the end of the domain ($i = n_x$) we set zero values for the charge densities. Using this procedure, we can compute two-dimensional traveling waves. In the following calculations, we have $\lambda = 10$, $L_y = 2\lambda$, $L_x = 3L_y$ and the domain is discretized by 300×100 points. The initial condition has a plane front perturbed with a cosenoidal perturbation of wavelength λ and amplitude $\lambda/40$.

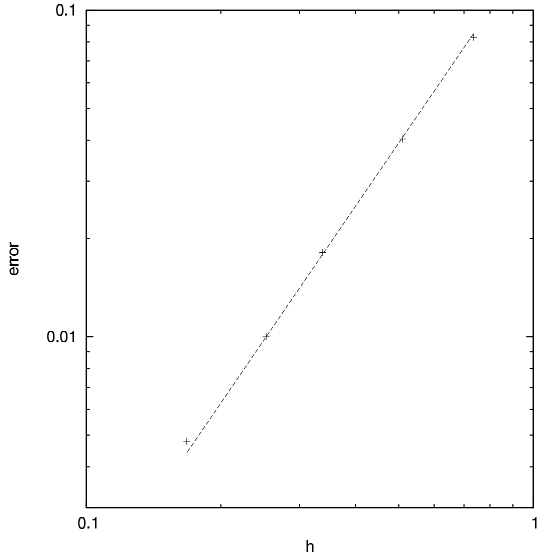


FIG. 4. *Errors integrated along the domain of integration for $D = 0$ at time $t_q = 9.5$. The size of the domain is $L = 50$ and we used 200 gridpoints. The points indicate the resulting numerical errors, and the line is a power with exponent 2, indicating that the scheme is second order accurate.*

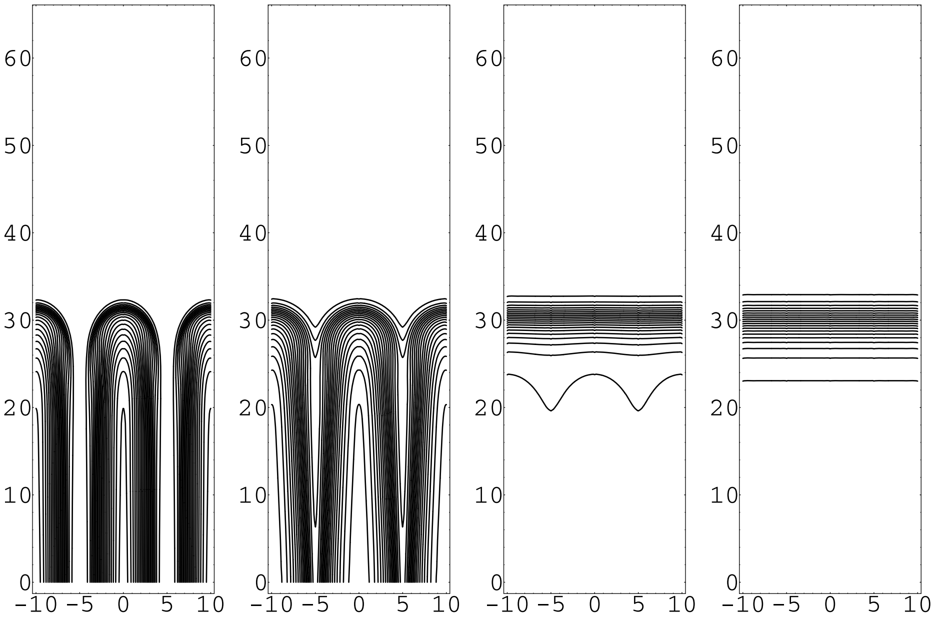


FIG. 5. *Two-dimensional contour plots for the electronic charge density for the traveling waves with $D = 0$, 0.1, 0.2, and 0.3. The x-axis is in the vertical direction.*

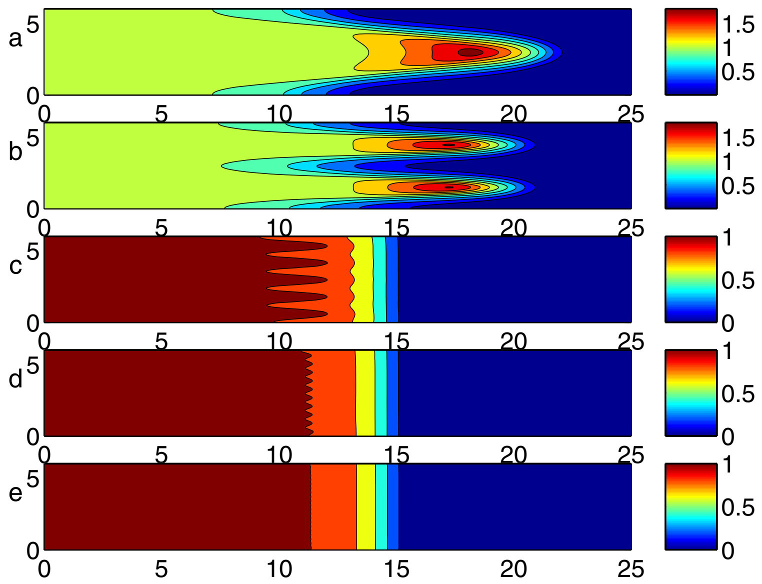


Fig. 6. *Level curves of the electron density n_e with diffusion coefficient $D = 0.1$ and time interval* 2. *The wavelength of the perturbation is, in each case,* (a) $\lambda = 6$, (b) $\lambda = 3$, (c) $\lambda = 5/6$, (d) $\lambda = 10/6$, *and* (e) $\lambda = 20/6$. *These values correspond to wave numbers $k = 2\pi/\lambda$.*

We observed that after the wave travels a distance equivalent to ten times the length of the computational domain, the numerical solution reaches a steady state, which is insensitive to the initial conditions. In Figure 5 we show traveling waves with $D = 0$, 0.1, 0.2, and 0.3. Notice that the aspect of the traveling waves is very sensitive to the value of the diffusion coefficient. In particular, when D is close to zero, well-developed fingers do appear, while the fronts remain essentially planar when D is large enough.

In Figure 6 we perturb a planar traveling wave, which was found with the displacement procedure described above. The perturbation was introduced by translating all the contour lines a distance $\cos(2\pi y/\lambda)$ on the x-direction. Then we evolved the solution on a time interval of length 2. In all cases $D = 0.1$, and we take several wavelengths λ. It is evident from the figures that there is a tendency to form fingers when the wavelength is above some critical value while the perturbation decays and disappears for small enough wavelengths. In Figure 7, the cases (b) and (c) of Figure 6 have been plotted in perspective. This confirms the results obtained in previous sections concerning stability.

6. Conclusions. In this paper we have used a fluid approximation to describe the process of electric breakdown in nonattaching gases such as nitrogen. We have shown that a planar negative front separating an ionized region from a region without charge may become unstable under the combined action of the external electric field and the electron diffusion. The common underlying mathematical structure allows us to exploit some of the ideas developed for other pattern-forming systems such as the Hele–Shaw and Stefan problems.

We have calculated the dispersion relation for a perturbation in the transversal direction of a planar traveling wave in the limit of small diffusion. An analytical expression for the typical spacing between fingers is obtained.

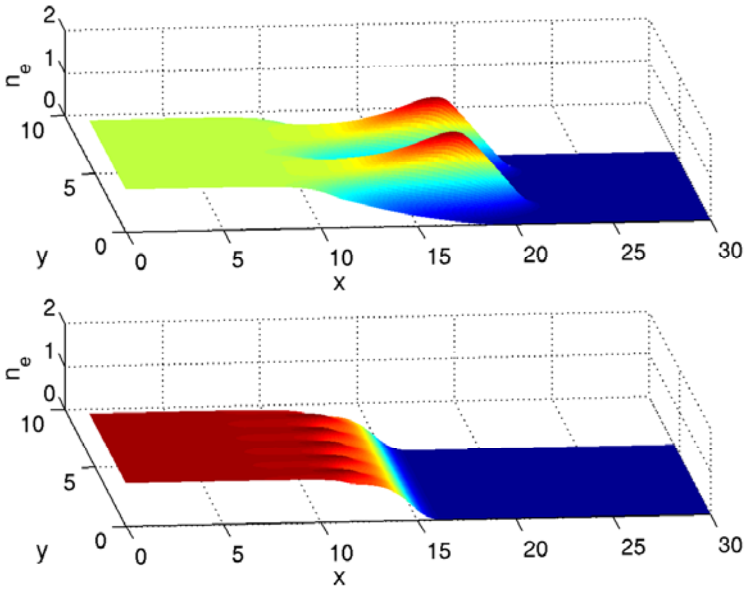


Fig. 7. *Representations in perspective of the electron density in the cases* (b) *and* (c) *of Figure* 6*, respectively.*

In order to test the analytical results, we have developed a numerical code to study the evolution of planar traveling waves. The traveling waves are then perturbed and we follow the evolution after that. Under some circumstances the solutions converge to traveling waves in the form of fingers that we have computed numerically for several diffusion coefficients. Our numerical results clearly support the conclusions on the branching and stability developed analytically.

Acknowledgments. We thank Irene Sendiña for stimulating discussions on some parts of this paper. M. A. and J. L. T. thank the Spanish Ministerio de Educación y Ciencia for its support under project ESP2007-66542-C04-03.

REFERENCES

[1] N. D. Alikakos, P. W. Bates, and X. Chen, *The convergence of solutions of the Cahn-Hilliard equation to the solution of Hele-Shaw model*, Arch. Ration. Mech. Anal., 128 (1994), pp. 165–205.

[2] S. M. Allen and J. W. Cahn, *A macroscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta. Metal., 27 (1979), pp. 1085–1095.

[3] W. F. Ames, *Numerical Methods for Partial Differential Equations*, 3rd ed., Academic Press, New York, 1992.

[4] M. Arrayás, U. Ebert, and W. Hundsdorfer, *Spontaneous branching of anode-directed streamers between planar electrodes*, Phys. Rev. Lett., 88 (2002), article 174502.

[5] M. Arrayás and U. Ebert, *Stability of negative ionization fronts: Regularization by electric screening?*, Phys. Rev. E, 69 (2004), article 036214.

[6] M. Arrayás, M. A. Fontelos, and J. L. Trueba, *Power laws and self-similar behavior in negative ionization fronts*, J. Phys. A: Math. Gen., 39 (2006), pp. 7561–7578.

[7] M. Arrayás, M. A. Fontelos, and J. L. Trueba, *Photoionization effects in ionization fronts*, J. Phys. D: Appl. Phys., 39 (2006), pp. 5176–5182.

[8] M. Arrayás, M. A. Fontelos, and J. L. Trueba, *Mechanism of branching in negative ionization fronts*, Phys. Rev. Lett., 95 (2005), article 165001.

[9] M. Arrayás and J. L. Trueba, *Investigations of pre-breakdown phenomena: Streamer discharges*, Cont. Phys., 46 (2005), pp. 265–276.

[10] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system* I: *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–351.

[11] E. N. Dancer, D. Hilhorst, M. Mimura, and L. A. Peletier, *Spatial segregation limit of a competition-diffusion system*, European J. Appl. Math., 10 (1999), pp. 97–115.

[12] U. Ebert, W. van Saarloos, and C. Caroli, *Propagation and structure of planar streamer fronts*, Phys. Rev. E, 55 (1997), pp. 1530–1549.

[13] C. Godrèche, ed., *Solids Far from Equilibrium*, Cambridge University Press, Cambridge, UK, 1991.

[14] A. N. Kolmogorov, I. G. Petrovskii, and N. S. Piskunov, *Study of the diffusion equation with growth of the quantity of matter and its application to a biology problem*, in Selected Works of A. N. Kolmogorov, Vol. I, Kluwer Academic, Amsterdam, 1991, pp. 242–270.

[15] N. Liu and V. P. Pasko, *Effects of photoionization on propagation and branching of positive and negative streamers in sprites*, J. Geophys. Res., 109 (2004), pp. 1–17.

[16] J. D. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1990.

[17] V. P. Pasko, M. A. Stanley, J. D. Mathews, U. S. Inan, and T. G. Wood, *Electrical discharge from a thundercloud top to the lower ionosphere*, Nature, 416 (2002), pp. 152–154.

[18] Y. P. Raizer, *Gas Discharge Physics*, Springer, Berlin, 1991.

[19] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion, and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1146–1167

EFFECTIVE MOTION OF A VIRUS TRAFFICKING INSIDE A BIOLOGICAL CELL[*]

THIBAULT LAGACHE[†] AND DAVID HOLCMAN[‡]

Abstract. Virus trafficking is fundamental for infection success, and plasmid cytosolic trafficking is a key step of gene delivery. Based on the main physical properties of the cellular transport machinery such as microtubules and motor proteins, our goal here is to derive a mathematical model to study cytoplasmic trafficking. Because experimental results reveal that both active and passive movements are necessary for a virus to reach the cell nucleus, by taking into account the complex interactions of the virus with the microtubules, we derive here an estimate of the mean time a virus reaches the nucleus. In particular, we present a mathematical procedure in which the complex viral movement, oscillating between pure diffusion and a deterministic movement along microtubules, can be approximated by a steady state stochastic equation with a constant effective drift. An explicit expression for the drift amplitude is given as a function of the real drift, the density of microtubules, and other physical parameters. The present approach can be used to model viral trafficking inside the cytoplasm, which is a fundamental step of viral infection, leading to viral replication and, in some cases, to cell damage.

Key words. virus trafficking, cytoplasmic transport, mean first passage time, exit points distribution, stochastic processes, wedge geometry

AMS subject classification. 92B05

DOI. 10.1137/060672820

1. Introduction.

Because cytosolic transport has been identified as a critical barrier for synthetic gene delivery [1], the delivery of plasmids or viral DNAs from the cell membrane to the nuclear pores has attracted the attention of many biologists. The cell cytosol contains many types of organelles, actin filaments, microtubules, etc., so that to reach the nucleus, a viral DNA has to travel through a crowded and risky environment. We are interested here in studying the efficiency of the delivery process and we present a mathematical model of virus trafficking inside the cell cytoplasm. We model the viral movement as a Brownian motion. However, the density of actin filaments and microtubules inside the cell can hinder diffusion, as demonstrated experimentally [2]. In a crowded environment, we will model the virus as a material point. This reduction is simplistic for several reasons: an actin filament network can trap a diffusing object that is beyond a certain size, and, as observed experimentally, a DNA fragment cannot find its way across the actin filaments [2]. Active directional transport along microtubules or actin filaments seems then the only way to deliver a plasmid to the nucleus. The active transport of the virus generally involves motor proteins, such as kinesin (to travel in the direction of the cell membrane) or dynein (to travel toward the nucleus). Once a virus is attached to a dynein protein, its movement can be modeled as a deterministic drift toward the nucleus.

Recently, a macroscopic model has been developed to describe the dynamics of

[*] Received by the editors October 20, 2006; accepted for publication (in revised form) December 14, 2007; published electronically March 19, 2008.
http://www.siam.org/journals/siap/68-4/67282.html

[†] Department of Biology, Ecole Normale Supérieure, 75005 Paris, France (lagache@biologie.ens.fr).

[‡] Department of Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel, and Department of Biology and Mathematics, Ecole Normale Supérieure, 75005 Paris, France (holcman@biologie.ens.fr). This author is partially supported by the program “Chaire d’Excellence” from the French Ministry of Research.

adenovirus concentration inside the cell cytoplasm [3]. This approach offers very interesting results about the effect of microtubules, but neglects the complexity of the geometry and cannot be used to describe the movement of a single virus, which might be enough to cause cellular infection. Modeling virus trafficking requires the use of a stochastic description. We model here the motion of a virus as that of a material point, so the probability of it being trapped by actin filaments or microtubules is neglected. In the present approximation, the viral movement has two main components: a Brownian one, which accounts for its free movement, and a drift directed toward the centrosome or MTOC (microtubules organization center), an organelle located near the nucleus. The magnitude of the drift along microtubules depends on many parameters such as the binding and unbinding rates and the velocity of the motor proteins [4].

In the present approach, we present a method to approximate a time-dependent dynamics of virus trafficking by an effective stochastic equation with a radial steady state drift. The main difficulties we have to overcome arise from the time-dependent nature of the trajectories which consists of intermittent epochs of drifts and free diffusion. We propose to derive an explicit expression for the steady state drift amplitude. In this approximation, the effective drift will gather the mean properties of the cytoplasmic organization such as the density of microtubules and its off binding rate.

Our method for finding the effective drift can be described as follows. First, we approximate the cell geometry as a two-dimensional disk and use a pure Brownian description to approximate the virus diffusion step. This geometrical approximation is valid for any two-dimensional cell such as the *in vitro* flat skin fibroblast culture cells [3]: indeed, due to their adhesion to the substrate, the thickness of these cells can be neglected in first approximation. Second, when the distribution of the initial viral position is uniform on the cell surface, we will estimate, during the diffusing period, the hitting position on a microtubule. By solving a partial differential equation, inside a sliced shape domain, delimited by two neighboring microtubules, we will provide an estimate of the mean time to the most likely hitting point. Finally, the amplitude of the radial steady state drift will be obtained by an iterative method which assumes that, after a virus has moved a certain distance along a microtubule, it is released at a point uniformly distributed on the final radial distance from the nucleus, ready for a new random walk. This scenario repeats until the virus reaches the nucleus surface. Finally, we will compute the mean time, the mean number of steps before a virus reaches the nucleus, and the amplitude of the effective drift by using the following criteria: The mean first passage time (MFPT) to the nucleus of the iterative approximation is equal to the MFPT obtained by directly solving an Ornstein–Uhlenbeck stochastic equation. The explicit computation of the effective drift is a key result in the estimation of the probability and the mean time a single virus or DNA molecule takes to reach a small nuclear pore [5].

2. Modeling stochastic viral movement inside a biological cell.

We approximate the cell as a two-dimensional geometrical domain Ω, which is here a disk of radius R, and the nucleus located inside is a concentric disk of much smaller radius $\delta \ll R$. We model the motion of an unattached DNA fragment as a material point so that the probability of it being trapped by actin filaments or microtubules is neglected. The motion of a (DNA) molecule of mass m is described by the overdamped limit of the Langevin equation (Smoluchowski's limit) [6] for the position $\mathbf{X}(t)$ of the molecule at time t. When the particle is not bound to a microtubule filament, its movement is described as pure Brownian with a diffusion constant D. When the particle hits

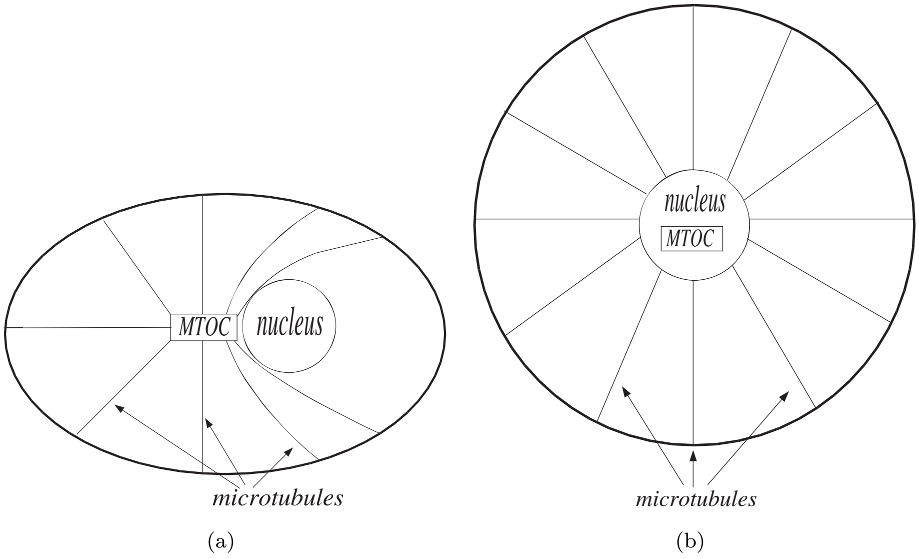


FIG. 2.1. *Cell geometry.* (a) *Cell's microtubules network. All microtubules starting from the cell membrane converge to the MTOC, located near the nucleus.* (b) *Simplified cell's microtubules network organization. The MTOC coincides with the nucleus.*

a filament, it binds for a certain random time and moves along with a deterministic drift. We take into account only the movement toward the nucleus, which coincides here with the MTOC, an organelle in which all microtubules converge (see Figure 2.1). For $\delta < |\mathbf{X}(t)| < R$, we describe the overall movement by the stochastic rule

$$\dot{\mathbf{X}} = \begin{cases} \sqrt{2D}\dot{\mathbf{w}} & \text{for} \quad \mathbf{X}(t) \quad \text{free}, \\ V\frac{\mathbf{r}}{|\mathbf{r}|} & \text{for} \quad \mathbf{X}(t) \quad \text{bound}, \end{cases} \tag{2.1}$$

where V is a constant velocity, $\dot{\mathbf{w}}$ a δ-correlated standard white noise, and $\mathbf{r}$ the $\mathbf{X}$ radial coordinate, the origin of which is the center of the cell. We assume that all filaments starting from the cell surface end on the nuclear surface. The binding time corresponds to a chemical reaction event; we assume that it is exponentially distributed, and for simplicity we approximate it by a constant t_m.

Once a virus enters the cell membrane, it moves according to the rule (2.1) until it hits a nuclear pore. Although nuclear pores occupy a small portion of the nuclear surface, we consider only the virus movement until it hits the nuclear surface $D(\delta)$. In this article, our goal is to replace (2.1) by a steady state stochastic equation

$$\dot{\mathbf{X}} = \mathbf{b}(\mathbf{X}) + \sqrt{2D}\dot{\mathbf{w}}, \tag{2.2}$$

where the drift $\mathbf{b}$ is radially symmetric. In a first approximation, we consider a constant radial drift $\mathbf{b}(\mathbf{X}) = -B\frac{\mathbf{r}}{|\mathbf{r}|}$ and compute hereafter the value of the constant amplitude B such that the MFPTs of the processes (2.2) and (2.1) to the nucleus are equal.

2.1. Modeling viral dynamics in the cytoplasm. Inside the cytosol, microtubules are distributed on the cell surface and converge radially to the MTOC. In the present analysis, we do not take into account the effect of organelle crowding due

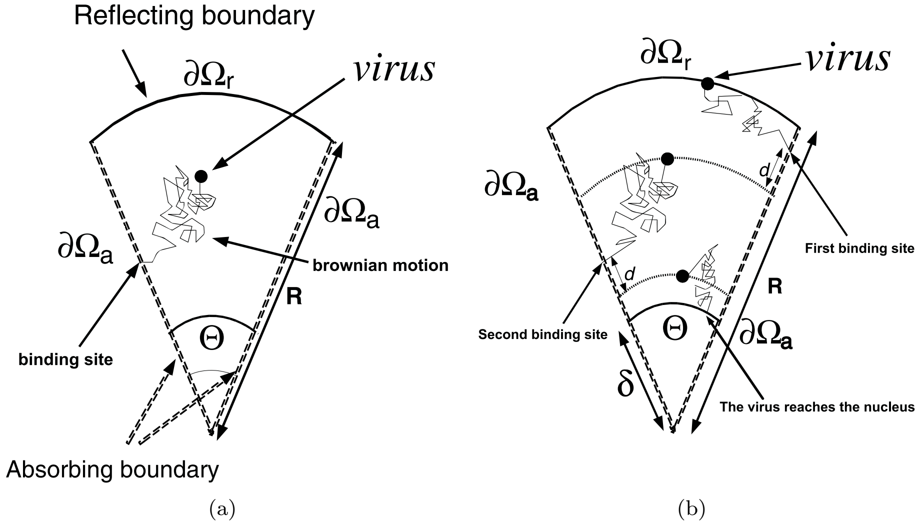


FIG. 2.2. *Virus trafficking inside a cell.* (a) *Representation of the cell portion between two microtubules.* (b) *Transport along microtubules: Two fundamental steps are represented. A fundamental step is made of the two intermediate steps which are first the diffusion inside the domain and then the directed motion along the microtubule.*

to the endoplasmic reticulum, the Golgi apparatus, etc. However, it is always possible to include them indirectly by using an apparent diffusion constant. We consider the fundamental domain $\tilde{\Omega}$ defined as the two-dimensional slice of angle Θ between two neighboring microtubules. We consider here that microtubules are uniformly distributed, and thus $\Theta = \frac{2\pi}{N}$, where N is the total number of microtubules.

Although a virus can drift along microtubules in both directions by using dynein (resp., kinesin) motor proteins for the inward (resp., outward) movement, we only take into account the drift toward the nucleus [7]. It is still unclear what is the precise mechanism used by a virus to select a direction of motion. Attached to a dynein molecule, the virus transport consists of several steps of few nanometers: the length of each step depends on the load of the transported cargo and adenosine tryphosphate (ATP) concentration [8]. We neglect here the complexity of this process, assuming that ATP molecules are abundant, uniformly distributed over the cell, and not a limiting factor. We thus assume the bound particle moves toward the nucleus with the mean constant velocity V. When the particle is released from the microtubule, inside the domain, the process can start afresh and the particle diffuses freely. Because the Smoluchowski limit of the Langevin equation does not account for the change in velocity, we release the particle at a certain distance away from the microtubule, but at a fixed distance from the nucleus (at an angle chosen uniformly distributed); see Figure 2.2.

Because microtubules are taken uniformly distributed, we can always release the virus inside the slice $\tilde{\Omega}$, between two neighboring microtubules. Thus the movement of the virus will be studied in $\tilde{\Omega}$: inside the cytosol, the viral movement is purely Brownian until it hits a microtubule, which is now the lateral boundary of $\tilde{\Omega}$ (see Figure 2.2). We assume that once a virus hits a microtubule, with probability one, the dynamics switches from diffusion to a deterministic motion with a constant drift. A virus spends on a microtubule a time that we consider to be exponentially distributed, since this time is the sum of escape time from deep potential wells. We approximate the total time on a microtubule by the mean time t_m. Thus a virus moves at a distance $d_m = Vt_m$ along a microtubule, which depends only on the characteristic of the virus–microtubule interactions. To summarize, the virus trajectory is a succession of diffusion steps mixed with some periods of attaching and detaching to microtubules. This scenario repeats until the virus hits the nuclear surface (Figure 2.2).

2.2. Computing the MFPT to reach the nucleus. We define the *mean time to infection* as the MFPT a virus reaches the surface of the disk $D(\delta)$ inside the domain $\tilde{\Omega}$ (see Figure 2.2).

To estimate the mean time to infection, we note that we can decompose the overall motion as a repeated fundamental step. This step consists of the free diffusion of the particle inside the domain followed by the motion along the microtubule. The total time of infection τ_i is then the sum of times the particle spends in each step. Although the time on a microtubule is deterministically equal to t_m, the diffusing time is not easy to compute and depends on the initial condition. Ultimately τ_i depends on the number of times the fundamental step repeats before the particle reaches the nucleus.

Let us now describe each step. The first step starts when the virus enters the cell at the periphery $r = R = R_0$ (at a random angle $\theta \in [0;\Theta]$) and ends when the virus hits either the lateral boundary or the nucleus. We now consider the first passage time $u(R_0)$ to the absorbing boundary and denote by $r(R_0)$ the hitting position. To account for the deterministic drift, during a deterministic time t_m we move the virus from a distance d_m along the microtubule. In that case, the initial random position for the next step is given by $r = R_1 = r(R_0) - d_m$ and the total time in step 1 is $u(R_0) + t_m$.

We iterate the process as follows and consider in each step k the distance $R_k = r(R_{k-1}) - d_m$ from which the particle starts and the time $u(R_k) + t_m$ it spends inside the step. If we denote by n_s the random number of steps necessary to reach the nucleus $r = \delta$, the time to infection τ_i is given by

$$\tau_i = \sum_{k=0}^{n_s-1} u(R_k) + n_s t_m + t_r, \tag{2.3}$$

where t_r is a residual time, which is the time to reach the nucleus before a full step is completed.

We are interested in estimating the MFPT τ of τ_i, given by

$$\tau = E(\tau_i) = E\left(\sum_{k=0}^{n_s-1} u(R_k)\right) + \langle n_s\rangle t_m + \langle t_r\rangle, \tag{2.4}$$

where $\langle n_s\rangle$ is the mean number of steps and $\langle t_r\rangle$ is the mean residual time. If we introduce the probability distribution $p_m = \Pr\{n_s = m\}$, which states that the number of steps is exactly equal to m, we can write

$$\tau = E(\tau_i) = \sum_{m=1}^{\infty} E\left(\sum_{k=0}^{n_s-1} u(R_k)|n_s = m\right) p_m + \langle n_s\rangle t_m + \langle t_r\rangle. \tag{2.5}$$

To estimate the MFPT τ, we shall approximate the previous sum by using the MFPT $\bar{u}(R_k)$ in each step k. To estimate $\bar{u}(R_k)$, we will solve (in the next paragraph) the Dynkin's equation with the following boundary conditions: Inside $\tilde{\Omega}$, the particle is reflected at the periphery $r = R$ and absorbed at the nucleus $\partial\tilde{\Omega}_a$ and at $\theta = 0$ and $\theta = \Theta$. We will also estimate the mean distance $\bar{d}_k$ covered during step k. For that purpose we will estimate the mean exit position $r_m(R_k)$, conditioned on the initial position $r = R_k$. Indeed, we will thus get $\bar{d}_k = R_k - r_m(R_k) - d_m$. The estimates of the mean distances covered for each fundamental step will ultimately lead to an approximation of the mean number of steps $n = \langle n_s \rangle$: n will be computed such that $R_n \geq \delta$ and $R_{n+1} < \delta$ (where $R_n = r_m(R_{n-1}) - d_m$ is defined recursively). Finally, we will obtain the following approximation for the infection time:

$$\tau \approx \sum_{k=0}^{n-1} \bar{u}(R_k) + nt_m + \langle t_r \rangle. \tag{2.6}$$

The mean residual time $\langle t_r \rangle$ can be equal either to $\bar{u}(R_n) + \alpha t_m$, where $0 \leq \alpha < 1$ if the virus binds to a microtubule in the last step and travels a distance αd_m on the microtubule, or to the MFPT to the nuclear boundary if $r_m(R_n) < \delta$.

3. MFPT and exit point distribution.

In a first approximation, under the assumptions of a sufficiently small radius $\delta \ll R$ and an angle $\Theta \ll 1$ for the computation of the MFPT and the distribution of exit points, we neglect the nuclear area. We define the full pie wedge Ω^R domain of angle Θ. Inside Ω^R, we use the boundary conditions described above. Consequently, the MFPT to a microtubule $u = u(r, \theta)$ of a virus starting initially at position (r, θ) is a solution of the Dynkin's equations [6]

$$\begin{aligned} D\Delta u(\boldsymbol{x}) &= -1 \text{ for } \boldsymbol{x} \in \Omega^R, \\ u(\boldsymbol{x}) &= 0 \text{ for } \boldsymbol{x} \in \partial\Omega^R_a, \\ \frac{\partial u}{\partial \boldsymbol{n}} &= 0 \text{ for } \boldsymbol{x} \in \partial\Omega^R_r, \end{aligned} \tag{3.1}$$

where $\partial\Omega^R_a = \{\theta = 0\} \cup \{\theta = \Theta\}$ and $\Omega^R_r = \{r = R\}$.

3.1. The general solution for the MFPT.

In this paragraph only we reparametrize the domain by $-\Theta/2 \leq \theta \leq \Theta/2$. By writing (3.1) in polar coordinates and using the separation of variables, the general solution of equation

$$\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \right)(r, \theta) = -1 \text{ for } (r, \theta) \in \Omega^R, \tag{3.2}$$

$$u(r, \theta) = 0 \text{ for } (r, \theta) \in \partial\Omega^R_a, \tag{3.3}$$

is given by [9],

$$u(r, \theta) = \frac{r^2}{4D}\left(\frac{\cos(2\theta)}{\cos(\Theta)} - 1 \right) + \sum_{n=0}^{\infty} A_n r^{\lambda_n} \cos(\lambda_n \theta) \text{ for } \frac{-\Theta}{2} \leq \theta \leq \frac{\Theta}{2}, \tag{3.4}$$

where the edge boundary is here located at position $\theta = \pm\Theta/2$. The sum in the right-hand side is the general solution of the homogeneous problem $\Delta u = 0$ in Ω^R. The boundary conditions on the sides of the wedge impose that

$$\lambda_n = (2n+1)\frac{\pi}{\Theta}, \tag{3.5}$$

while the reflecting condition for $r = R$ reads

$$\frac{\partial u}{\partial r}(R, \theta) = 0 \text{ for all } \theta \in \left[-\frac{\Theta}{2}, \frac{\Theta}{2} \right]. \tag{3.6}$$

Using the uniqueness of Fourier decomposition and the boundary condition (3.6), we obtain that

$$A_n = \frac{(-1)^{n+1} 8R^{2-\lambda_n}}{D\Theta\lambda_n^2 \left(\lambda_n^2 - 4\right)}. \tag{3.7}$$

By averaging formula (3.4) over an initial uniform distribution, the MFPT to one of the absorbing edges of the wedge is given by

$$\bar{u}(r) = \frac{1}{\Theta}\int_{\theta=0}^{\theta=\Theta} u(r,\theta)\,d\theta = \frac{r^2}{4D}\left(\frac{\tan(\Theta)}{\Theta} - 1\right) - \sum_{n=0}^{\infty} \frac{16R^{2-\lambda_n} r^{\lambda_n}}{D\Theta^2\lambda_n^3\left(\lambda_n^2 - 4\right)}, \tag{3.8}$$

where $\lambda_n = (2n+1)\frac{\pi}{\Theta}$. For Θ small, (3.8) can be approximated by

$$\bar{u}(r) = \frac{r^2}{4D}\left(\frac{\tan(\Theta)}{\Theta} - 1\right) - \frac{16\Theta R^2 (r/R)^{\pi/\Theta}}{D\pi^3\left((\pi/\Theta)^2 - 4\right)}. \tag{3.9}$$

3.2. Exit points distribution. To estimate the position at which a virus will attach preferentially to the microtubule, we determine the distribution of exit points, when the viral particle initially started at a certain radial distance from the nucleus. We recall that the probability density function (pdf) $p(\mathbf{r}, t|\mathbf{r_0})$ for finding a diffusing particle in a volume element $d\mathbf{r}$ at time t inside the wedge $\tilde{\Omega}$, conditioned on the initial position $\mathbf{r} = \mathbf{r_0}$, is a solution of the diffusion equation

$$\begin{aligned}\frac{\partial p(\mathbf{r}, t|\mathbf{r_0})}{\partial t} &= D\Delta p(\mathbf{r}, t|\mathbf{r_0}) \text{ for } \mathbf{r} \in \Omega^R,\\ p(\mathbf{r}, t|\mathbf{r_0}) &= 0 \text{ for } \mathbf{r} \in \partial\Omega_a^R,\\ \frac{\partial p(\mathbf{r}, t|\mathbf{r_0})}{\partial n} &= 0 \text{ for } \mathbf{r} \in \partial\Omega_r^R,\end{aligned}$$

where the initial condition is $p(\mathbf{r}, 0|\mathbf{r_0}) = \delta(\mathbf{r} - \mathbf{r_0})$. The distribution of exit points $\epsilon(\boldsymbol{y})$ is given by

$$\epsilon(\boldsymbol{y}) = \int_0^\infty j(\boldsymbol{y}, t)\,dt, \tag{3.10}$$

where the flux j is defined by

$$j(\boldsymbol{y}, t) = -D\frac{\partial p(\mathbf{r}, t)}{\partial \boldsymbol{n}}_{|\mathbf{r} = \boldsymbol{y}}.$$

If we denote $C(\mathbf{r_0}, \mathbf{r}) = \int_0^\infty p(\mathbf{r}, t|\mathbf{r_0})\,dt$, then C is a solution of

$$-D\Delta C(\mathbf{r_0}, \mathbf{r}) = \delta(\mathbf{r} - \mathbf{r_0}), \tag{3.11}$$

and we have

$$\epsilon(\mathbf{y}) = -D\frac{\partial C}{\partial n}(\mathbf{r_0}, \mathbf{y}) \text{ for } \mathbf{y} \in \Omega_a^R. \tag{3.12}$$

Consequently, to obtain the pdf of exit points ϵ, we use the Green function in the wedge domain Ω^R. By using a conformal transformation, we hereafter solve a simplified case of an open wedge (i.e., without a reflecting boundary at $r = R$). This computation could be compared with the general one that will be derived in the next section.

To compute the exit points distribution, we consider the solution of (3.11), obtained by the image method and a conformal transformation from the open wedge to the upper complex half-plane. The Green function, solution of (3.11) in the upper complex half-plane, is given by

$$C(z) = \frac{1}{2\pi D} \ln \frac{z - z_0}{z - z_0^*}, \tag{3.13}$$

where z_0^* is the complex conjugate of z_0. Using the conformal transformation $\omega = f(z) = z^{\frac{\pi}{\Theta}}$ [10] that maps the interior of the wedge of opening angle Θ to the upper half-plane, the Green function in the wedge is given by

$$C(z) = \frac{1}{2\pi D} \ln \left(\frac{z^{\frac{\pi}{\Theta}} - z_0^{\frac{\pi}{\Theta}}}{z^{\frac{\pi}{\Theta}} - (z_0^*)^{\frac{\pi}{\Theta}}} \right). \tag{3.14}$$

The flux to the line θ is given by

$$\begin{aligned}
\epsilon_\theta(r) = -\frac{D}{r}\frac{\partial C}{\partial \theta}\left(re^{i\theta}\right) &= \frac{1}{2\pi r} \frac{i\nu \left(re^{i\theta}\right)^\nu . (k_0 - k_0^*)}{\left(\left(re^{i\theta}\right)^\nu - k_0\right)\left(\left(re^{i\theta}\right)^\nu - k_0^*\right)} \\
&= \frac{1}{2\pi r} \frac{-2\nu \left(re^{i\theta}\right)^\nu r_0^\nu \sin(\nu\theta_0)}{\left(re^{i\theta}\right)^{2\nu} + r_0^{2\nu} - 2\left(re^{i\theta}\right)^\nu r_0^\nu \cos(\nu\theta_0)},
\end{aligned}$$

where $\nu = \frac{\pi}{\Theta}$, $k_0 = z_0^\nu = \left(r_0 e^{i\theta_0}\right)^\nu$. Finally, the exit point distribution for $\theta = \Theta$ is given by

$$\epsilon_\Theta(r) = \frac{r_0}{\Theta} \frac{(rr_0)^{(\nu-1)} \sin(\nu\theta_0)}{r^{2\nu} + r_0^{2\nu} + 2(rr_0)^\nu \cos(\nu\theta_0)}, \tag{3.15}$$

while for $\theta = 0$ it is given by

$$\epsilon_0(r) = \frac{r_0}{\Theta} \frac{(rr_0)^{(\nu-1)} \sin(\nu\theta_0)}{r^{2\nu} + r_0^{2\nu} - 2(rr_0)^\nu \cos(\nu\theta_0)}. \tag{3.16}$$

A MATLAB check guarantees that

$$\int_0^\infty \{\epsilon_\Theta(r) + \epsilon_0(r)\} dr = 1. \tag{3.17}$$

This simple computation is instructive and shall be compared to the full one given in section 3.3.

3.3. Exit pdf in a pie wedge. To compute the exit points distribution in a pie wedge with a reflecting boundary at $r = R$, we search for an explicit solution of the diffusion equation in polar coordinates inside the pie wedge. We first consider the general diffusion equation

$$\begin{aligned}
\frac{\partial p}{\partial t}(\boldsymbol{x}, t|\boldsymbol{y}) &= D\left(\frac{\partial^2 p}{\partial r^2} + \frac{1}{r}\frac{\partial p}{\partial r} + \frac{1}{r^2}\frac{\partial^2 p}{\partial \theta^2}\right)(\boldsymbol{x}, t|\boldsymbol{y}), \\
p(\boldsymbol{x}, 0|\boldsymbol{y}) &= \delta(\boldsymbol{x} - \boldsymbol{y}),
\end{aligned} \tag{3.18}$$

where the boundary conditions are given in (3.1). We may often use the change of variable for all $n \in \mathbf{N}^*$:

$$k = \frac{n\pi}{\Theta}.$$

The initial condition is given by

$$p(\boldsymbol{x}, 0|\boldsymbol{y}) = p(r, \theta, 0|r_0, \theta_0) = \frac{2}{\Theta r_0}\delta(r - r_0)\sum_k \sin(k\theta)\sin(k\theta_0)$$

for $\theta < \theta_0$ (if $\theta > \theta_0$, θ_0 must be replaced by $\Theta - \theta_0$). To compute the solution of (3.18), we consider the Laplace transform $\hat{p}$ of the probability p,

$$\begin{aligned} s\hat{p}(r, \theta, s|r_0, \theta_0) - \frac{2}{\Theta r_0}\delta(r - r_0)\sum_k \sin(k\theta)\sin(k\theta_0) \\ = D\left(\frac{\partial^2 \hat{p}}{\partial r^2} + \frac{1}{r}\frac{\partial \hat{p}}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \hat{p}}{\partial \theta^2}\right)(r, \theta, s|r_0, \theta_0). \end{aligned}$$

Using the separation of variables, we have

$$\hat{p}(r, \theta, s|r_0, \theta_0) = \sum_k R_k(r, s)\sin(k\theta)\sin(k\theta_0).$$

Using the change of variable, $x(s) = r\sqrt{\frac{s}{D}}$ and $x_0(s) = r_0\sqrt{\frac{s}{D}}$, we get for all k that

$$\begin{aligned} R_k^{''}(x(s), s) + \frac{1}{x(s)}R_k^{'}(x(s), s) - \left(1 + \frac{k^2}{x(s)^2}\right)R_k(x(s), s) \\ = -\frac{2}{\Theta D x_0(s)}\delta(x(s) - x_0(s)). \end{aligned} \tag{3.19}$$

$R_k(x(s), s)$ is a superposition of modified Bessel functions of order k: $I_k(x(s))$ and $K_k(x(s))$. Thus, for $x(s) \neq x_0(s)$ we obtain that

$$R_k(x(s), s) = A_k I_k(x(s)) + B_k K_k(x(s)),$$

where A_k and B_k are real constants. Since K_k diverges as $x(s) \to 0$, the interior solution for $(x(s) < x_0(s))$ depends only on I_k. We denote by D_k the exterior solution for $(x(s) > x_0(s))$. We use the general notation $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$; thus

$$R_k(x(s), s) = A_k I_k(x(s) \wedge x_0(s)) D_k(x(s) \vee x_0(s)).$$

To determine $D_k = a_k I_k + b_k K_k$, we use the reflecting condition at $x(s) = x_+(s) = R\sqrt{\frac{s}{D}}$ and we get that

$$A_k I_k(x_0(s)).\left(a_k I_k^{'}(x_+(s)) + b_k K_k^{'}(x_+(s))\right) = 0.$$

We choose

$$a_k = -K_k^{'}(x_+(s)) \text{ and } b_k = I_k^{'}(x_+(s)).$$

Thus

$$R_k\left(x\left(s\right),s\right)=A_kI_k\left(x\left(s\right)\wedge x_0\left(s\right)\right)\left(I_k^{'}\left(x_+\left(s\right)\right)K_k-K_k^{'}\left(x_+\left(s\right)\right)I_k\right)\left(x\left(s\right)\vee x_0\left(s\right)\right).$$

The constants A_k are determined by integrating (3.19) over an infinitesimal interval that includes r_0. Using the continuity of R_k, we get

$$\left(R_k\right)_{x(s)>x_0(s)}^{'}|_{x(s)=x_0(s)}-\left(R_k\right)_{x(s)<x_0(s)}^{'}|_{x(s)=x_0(s)}=-\frac{2}{\Theta Dx_0\left(s\right)},$$

that is,

$$A_k\left(I_k\left(I_k^{'}\left(x_+\left(s\right)\right)K_k^{'}-K_k^{'}\left(x_+\left(s\right)\right)I_k^{'}\right)-I_k^{'}\left(I_k^{'}\left(x_+\left(s\right)\right)K_k-K_k^{'}\left(x_+\left(s\right)\right)I_k\right)\right)\left(x_0\left(s\right)\right)$$
$$=-\frac{2}{\Theta Dx_0\left(s\right)}.$$

After some simplifications, we get

$$A_kI_k^{'}\left(x_+\left(s\right)\right)\left(I_kK_k^{'}-I_k^{'}K_k\right)\left(x_0\left(s\right)\right)=-\frac{2}{\Theta Dx_0\left(s\right)}.$$

Using the recurrent relation between modified Bessel functions (see [11] or [12, p. 489]),

$$I_k^{'}\left(x_0\left(s\right)\right)=\left(I_{k-1}-\frac{k}{x_0\left(s\right)}I_k\right)\left(x_0\left(s\right)\right)\text{ and }K_k^{'}\left(x_0\left(s\right)\right)=\left(-K_{k-1}-\frac{k}{x_0\left(s\right)}K_k\right)\left(x_0\left(s\right)\right),$$

we get

$$A_kI_k^{'}\left(x_+\left(s\right)\right)\left(I_k\left(-K_{k-1}-\frac{k}{x_0\left(s\right)}K_k\right)-\left(I_{k-1}-\frac{k}{x_0\left(s\right)}I_k\right)K_k\right)\left(x_0\left(s\right)\right)=-\frac{2}{\Theta Dx_0\left(s\right)},$$

that is

$$A_kI_k^{'}\left(x_+\left(s\right)\right)\left(I_kK_{k-1}+I_{k-1}K_k\right)\left(x_0\left(s\right)\right)=\frac{2}{\Theta Dx_0\left(s\right)}.$$

Finally, using this relation and the following Wronskian relation [12, p. 489]:

$$\left(I_kK_{k-1}+I_{k-1}K_k\right)\left(x_0\left(s\right)\right)=\frac{1}{x_0\left(s\right)},$$

we obtain that

$$A_k=\frac{2}{\Theta DI_k^{'}\left(x_+\left(s\right)\right)}.$$

Thus

$$R_k\left(x\left(s\right),s\right)$$
$$=\frac{2}{\Theta DI_k^{'}\left(x_+\left(s\right)\right)}I_k\left(x\left(s\right)\wedge x_0\left(s\right)\right)\left(I_k^{'}\left(x_+\left(s\right)\right)K_k-K_k^{'}\left(x_+\left(s\right)\right)I_k\right)\left(x\left(s\right)\vee x_0\left(s\right)\right).$$

We can now express the solution $\hat{p}$ for $\theta<\theta_0$ by

$$\hat{p}\left(r,\theta,s\right)=\frac{2}{\Theta D}\sum_k\frac{I_k\left(x\left(s\right)\wedge x_0\left(s\right)\right)\left(I_k^{'}\left(x_+\left(s\right)\right)K_k-K_k^{'}\left(x_+\left(s\right)\right)I_k\right)\left(x\left(s\right)\vee x_0\left(s\right)\right)}{I_k^{'}\left(x_+\left(s\right)\right)}\sin\left(k\theta\right)\sin\left(k\theta_0\right).$$

The exit point distribution $\epsilon^0(r)$ is given by

$$\epsilon^0(r) = -\left(\frac{D}{r}\frac{\partial}{\partial\theta}\left(\int_0^\infty p(r,\theta,t)\,dt\right)\right)(\theta = 0). \tag{3.20}$$

To obtain an analytical expression for expression (3.20), we use the Laplace relation

$$\mathcal{L}\left(\int_0^t f(u)\,du\right) = \frac{F(z)}{z},$$

where $F = \mathcal{L}(f)$ is the Laplace transform of the function f. We have

$$\int_0^t p(r,\theta,u)\,du = \mathcal{L}^{-1}\left(\frac{\hat{p}(r,\theta,s)}{s}\right)$$

$$= \mathcal{L}^{-1}\left(\frac{2}{\Theta D}\sum_k \sin(k\theta)\sin(k\theta_0)\frac{I_k(x(s)\wedge x_0(s))\left(I_k'(x_+(s))K_k - K_k'(x_+(s))I_k\right)(x(s)\vee x_0(s))}{sI_k'(x_+(s))}\right).$$

The computation of the integral

$$\begin{aligned}&I(r,\theta,t)\\ &= \frac{1}{\Theta\pi Di}\sum_k \sin(k\theta)\sin(k\theta_0)\int_{-i\infty}^{+i\infty}\frac{I_k(x(s)\wedge x_0(s))(I_k'(x_+(s))K_k - K_k'(x_+(s))I_k)(x(s)\vee x_0(s))}{sI_k'(x_+(s))}e^{st}ds\end{aligned} \tag{3.21}$$

uses the residue theorem, and the details are given in the appendix. We have

$$I(r,\theta,t) = \int_0^t p(r,\theta,u)\,du = \frac{2}{\Theta D}\left(S_1(r,\theta,t) + S_2(r,\theta,t)\right),$$

where

$$S_1(r,\theta,t) = \sum_k \sin(k\theta)\sin(k\theta_0)\frac{r^k\left(r_0^{2k} + R^{2k}\right)}{2kR^{2k}r_0^k},$$

$$S_2(r,\theta,t) = -2\sum_k \sin(k\theta)\sin(k\theta_0)\sum_{j=1}^\infty e^{-D\alpha_{j,k}^2 t}\frac{J_k(r\alpha_{j,k})J_k(r_0\alpha_{j,k})}{\left(R^2\alpha_{j,k}^2 - k^2\right)J_k^2(R\alpha_{j,k})},$$

and J_k are the k-order Bessel functions and $\alpha_{j,k}$ are the roots of the equation:

$$J_k'(R\alpha) = 0.$$

Consequently, for $r < r_0$, using (3.20), we get the following exit distribution (for $\Theta = 0$):

$$\epsilon^0(r) = \frac{2}{\Theta}\frac{\partial}{r\partial\theta}\left(\lim_{t\to\infty}\left(S_1(r,\theta,t) + S_2(r,\theta,t)\right)\right)_{\theta=0}.$$

Because

$$\lim_{t\to\infty} S_1(r,\theta,t) = S_1(r,\theta) \text{ and } \lim_{t\to\infty} S_2(r,\theta,t) = 0,$$

we finally obtain that

$$\epsilon^0(r) = \frac{1}{\Theta}\sum_k \sin(k\theta_0)\frac{r^{k-1}\left(r_0^{2k} + R^{2k}\right)}{R^{2k}r_0^k}, \tag{3.22}$$

and, for $r > r_0$, a similar computation leads to

$$\epsilon^0(r) = \frac{1}{\Theta}\sum_k \sin(k\theta_0)\frac{r_0^k\left(r^{2k}+R^{2k}\right)}{R^{2k}r^{k+1}}. \tag{3.23}$$

These expressions can be further simplified. Indeed, we rewrite them as follows (for $r < r_0$):

$$\epsilon^0(r) = \frac{1}{\Theta r}\sum_k \sin(k\theta_0)\left(\frac{r}{r_0}\right)^k\left(1+\left(\frac{r_0}{R}\right)^{2k}\right).$$

Thus,

$$\epsilon^0(r) = \frac{1}{\Theta r}\Im m\left(\sum_{n\geq 1} e^{in\nu\theta_0}\left(\frac{r}{r_0}\right)^{n\nu}\left(1+\left(\frac{r_0}{R}\right)^{2n\nu}\right)\right),$$

where $\Im m$ denotes the imaginary part of the expression. We obtain two geometrical series that can be summed. We get

$$\epsilon^0(r) = \frac{1}{\Theta r}\Im m\left(\frac{e^{i\nu\theta_0}\left(\frac{r}{r_0}\right)^\nu}{1-e^{i\nu\theta_0}\left(\frac{r}{r_0}\right)^\nu}+\frac{e^{i\nu\theta_0}\left(\frac{r}{r_0}\right)^\nu\left(\frac{r_0}{R}\right)^{2\nu}}{1-e^{i\nu\theta_0}\left(\frac{r}{r_0}\right)^\nu\left(\frac{r_0}{R}\right)^{2\nu}}\right),$$

that is,

$$\epsilon^0(r) = \frac{1}{\Theta r}\Im m\left(e^{i\nu\theta_0}\left(\frac{\left(\frac{r}{r_0}\right)^\nu}{1-e^{i\nu\theta_0}\left(\frac{r}{r_0}\right)^\nu}+\frac{\left(\frac{rr_0}{R^2}\right)^\nu}{1-e^{i\nu\theta_0}\left(\frac{rr_0}{R^2}\right)^\nu}\right)\right).$$

After some rearrangements, we obtain the following exit point distribution on $\theta = 0$, conditioned on the initial position (r_0, θ_0):

$$\begin{aligned}\epsilon^0(r) &= \epsilon^0(r|r_0,\theta_0)\\ &= \frac{1}{\Theta r}\left(\frac{(rr_0)^\nu\sin(\nu\theta_0)}{r^{2\nu}+r_0^{2\nu}-2(rr_0)^\nu\cos(\nu\theta_0)}+\frac{\left(rr_0R^2\right)^\nu\sin(\nu\theta_0)}{(rr_0)^{2\nu}+R^{4\nu}-2\left(rr_0R^2\right)^\nu\cos(\nu\theta_0)}\right),\end{aligned} \tag{3.24}$$

for $0 \leq r \leq R$. Similarly, for $\theta = \Theta$, we obtain

$$\begin{aligned}\epsilon^\Theta(r) &= \epsilon^\Theta(r|r_0,\theta_0)\\ &= \frac{1}{\Theta r}\left(\frac{(rr_0)^\nu\sin(\nu\theta_0)}{r^{2\nu}+r_0^{2\nu}+2(rr_0)^\nu\cos(\nu\theta_0)}+\frac{\left(rr_0R^2\right)^\nu\sin(\nu\theta_0)}{(rr_0)^{2\nu}+R^{4\nu}+2\left(rr_0R^2\right)^\nu\cos(\nu\theta_0)}\right).\end{aligned} \tag{3.25}$$

We notice that by letting R tend to ∞, we recover the expressions computed in the open wedge case ((3.15) and (3.16)).

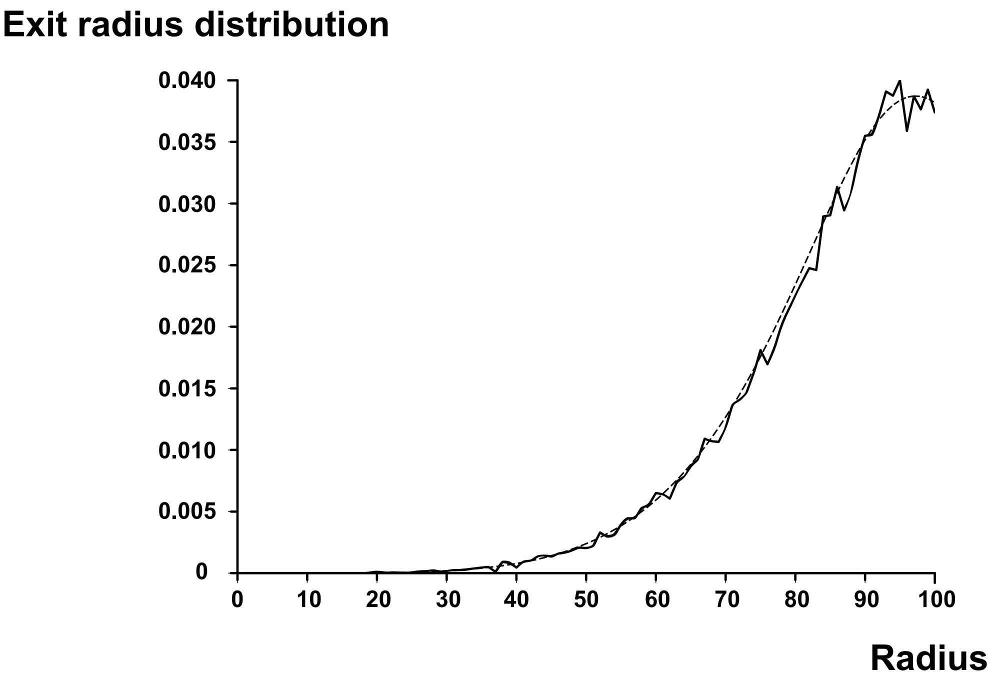


FIG. 3.1. *Mean exit points distribution. The theoretical distribution (dashed line) is tested against the empirical one (solid line) obtained by running a simulation of* 20,000 *Brownian particles, starting on the wedge bisectrix* ($\theta_0 = \frac{\Theta}{2}$ *at* $r_0 = R = 100$ *for* $\Theta = \frac{\pi}{6}$)*. Because the starting point is located on the bisectrix,* $\epsilon^0(x) = \epsilon^\Theta(x)$*, and thus the analytical curve is given by* $\epsilon(r) = \epsilon^0(r) + \epsilon^\Theta(r) = \frac{2}{\Theta r}\left(\frac{(rr_0)^{(\nu)}}{r^{2\nu}+r_0^{2\nu}} + \frac{(rr_0R^2)^{(\nu)}}{(rr_0)^{2\nu}+R^{4\nu}}\right)$*. In that case, the maximum of the function* $\epsilon(r)$ *is achieved at* $r = r_0 e^{\frac{1}{2\nu}\ln\left(\frac{\nu-1}{\nu+1}\right)}$.

3.4. The mean exit radius. To determine the mean exit distribution radius $\overline{\epsilon}(r|r_0)$ for a viral particle starting initially at position r_0, θ_0, where θ_0 is uniformly distributed between 0 and Θ, we consider $\epsilon(r|r_0,\theta_0) = \epsilon^0(r|r_0,\theta_0) + \epsilon^\Theta(r|r_0,\theta_0)$ and estimate the integral

$$\overline{\epsilon}(r|r_0) = \frac{1}{\Theta}\int_{\Theta_0=0}^{\Theta} \epsilon(r|r_0,\theta_0)\, d\theta_0. \tag{3.26}$$

Integrating expressions (3.24) and (3.25), we get

$$\overline{\epsilon}(r|r_0) = \frac{2}{\Theta\pi r}\left(\ln\left(\frac{r^\nu + r_0^\nu}{|r^\nu - r_0^\nu|}\right) + \ln\left(\frac{R^{2\nu} + (rr_0)^\nu}{R^{2\nu} - (rr_0)^\nu}\right)\right).$$

We define the mean exit point as $r_m(r_0) = \mathbf{E}(r|r_0)$ conditioned on the initial radius r_0. Thus,

$$r_m(r_0) = \mathbf{E}(r|r_0) = \int_0^R r\overline{\epsilon}(r|r_0)\, dr. \tag{3.27}$$

Using the expansion $\ln(1+x) = \sum_{n\geq 1}(-1)^{n+1}\frac{x^n}{n}$ for $x < 1$, we obtain by a direct integration that

$$r_m\left(r_0\right) = \frac{8}{\pi^2}\left(r_0\left(\sum_{n=0}^{\infty}\frac{1}{\left(2n+1\right)^2}\left(\frac{1}{1-\frac{1}{(2n+1)^2\left(\frac{\pi}{\Theta}\right)^2}}\right)\right)\right.$$
$$\left. - R\left(\sum_{n=0}^{\infty}\frac{\left(\frac{r_0}{R}\right)^{(2n+1)\frac{\pi}{\Theta}}\frac{\pi}{\Theta}}{\left(2n+1\right)\left(\left(\left(2n+1\right)\frac{\pi}{\Theta}\right)^2-1\right)}\right)\right). \tag{3.28}$$

Using the expansion in the first part,

$$\frac{1}{1-\frac{1}{(2n+1)^2\left(\frac{\pi}{\Theta}\right)^2}} = \sum_{p=0}^{\infty}\left(\frac{\Theta}{\left(2n+1\right)\pi}\right)^{2p}, \tag{3.29}$$

and the approximation $\Theta \ll 1$, by using the value of the Riemann ζ-function, $\zeta\left(2\right) = \frac{\pi^2}{6}$ and $\zeta\left(4\right) = \frac{\pi^4}{90}$, we obtain that

$$r_m\left(r_0\right) \approx r_0\left(1+\frac{\Theta^2}{12}\right) - \frac{8R}{\pi^2}\left(\frac{r_0}{R}\right)^{\pi/\Theta}\frac{\pi/\Theta}{\left(\pi/\Theta\right)^2-1}. \tag{3.30}$$

For Θ small, the second term in the right-hand side of (3.30) is exponentially small.

4. Approximation of a virus motion by an effective Markovian stochastic equation.

We replace the successive steps of viral dynamics with an effective stochastic equation containing a constant steady state drift.

4.1. Methodology.

Virus motion described in subsection 2.2 consists of a succession of drift and diffusing periods. We start with the stochastic equation

$$\dot{\mathbf{X}} = -B\frac{\mathbf{r}}{|\mathbf{r}|} + \sqrt{2D}\dot{\mathbf{w}}, \tag{4.1}$$

where $\boldsymbol{r}$ is the radial component of $\mathbf{X}$ and B is the amplitude of the drift. The MFPT of the process (4.1) to the nucleus, which is located at $r = \delta$, when the initial position is located on the cell surface $r = R$, is solution of

$$\begin{aligned} D\left(\frac{d^2t}{dr^2}+\frac{1}{r}\frac{dt}{dr}\right)\left(r,\theta\right) - B\frac{dt}{dr}\left(r,\theta\right) &= -1 \text{ for } \left(r,\theta\right)\in\Omega,\\ t\left(r,\theta\right) &= 0 \text{ for } r=\delta,\\ \frac{dt}{dr}\left(r,\theta\right) &= 0 \text{ for } r=R. \end{aligned}$$

A similar equation can be written in the domain $\tilde{\Omega}$ with reflective boundary conditions of the wedge. Both processes in the full domain or in $\tilde{\Omega}$ lead to the same MFPT. The solution $t(B,r)$ is given by

$$t\left(B,r\right) = C - \int_r^R\left(\int_v^R\frac{ue^{-\alpha(u-v)}}{Dv}du\right)dv, \tag{4.2}$$

where $\alpha = \frac{B}{D}$ and

$$t\left(B,R\right) = C = \int_\delta^R\left(\int_v^R\frac{ue^{-\alpha(u-v)}}{Dv}du\right)dv. \tag{4.3}$$

For a fixed radius R, the derivative of the function $t(B,R)$ with respect to B is strictly negative, which shows that $B \to t(B,R)$ is strictly decreasing. To determine the value of the amplitude B, we match the mean time $t(B,R)$ with the MFPT to reach the nucleus within the iterative procedure, as described in subsection 2.2: at time zero, the virus starts at position $r = R = R_0$ and reaches the edge boundary in mean time $\bar{u}(R_0)$ and at mean position $r_m(R_0)$. The viral particle is then transported toward the nucleus over a distance d_m during time t_m. Either the particle reaches the nucleus before time t_m and then the algorithm is terminated, or in a second step it starts at position $R_1 = r_m(R_0) - d_m$. The process iterates until the particle reaches the nucleus. We consider the mean number of fundamental steps (diffusion step and directed motion along a microtubule step) that the virus needs to reach the nucleus to be equal to $n \geq 0$. Thus the mean time to reach the nucleus computed by (4.2) has to be equal to the mean time $\tau = \sum_{k=0}^{n-1} \bar{u}(R_k) + nt_m + \langle t_r \rangle$ of the iterative trajectory. In a first approximation, we neglect the mean residual time $\langle t_r \rangle$ and we thus get the equality

$$t(B,R) = \tau = \sum_{k=0}^{n-1} \bar{u}(R_k) + nt_m, \tag{4.4}$$

$$R_{k+1} = r_m(R_k) - d_m, \tag{4.5}$$

$$R_0 = R. \tag{4.6}$$

For a fixed radius R, equation (4.4) has a unique solution B, which can be found in practice by any standard numerical method.

Remark. The MFPT of a particle in which the trajectory consists of alternating drift (traveling along microtubules) and diffusion periods can be either higher or lower than the MFPT of a pure Brownian particle. Indeed, when $B < 0$, the drift effect is less efficient than pure diffusion. For example, for $\Theta = \frac{\pi}{6}$, $R = 100\mu m$, and $\delta = \frac{R}{4} = 25\mu m$, a large diffusion constant $D = 10\mu m^2 s^{-1}$ with the dynamical parameters $t_m = 1s$ and $d_m = 1\mu m$ leads to a negative mean drift

$$B \approx -0.14\mu m s^{-1}. \tag{4.7}$$

On the other hand, for a small diffusion constant $D = 1\mu m^2 s^{-1}$, an efficient microtubule transport obtained for $t_m = 1s$ and $d_m = 5\mu m$ leads to a mean positive drift

$$B \approx 0.13\mu m s^{-1}. \tag{4.8}$$

4.2. Explicit expression of the drift in the limit of $\Theta \ll 1$.

When the number of microtubules is large enough, the condition $\Theta \ll 1$ is satisfied. Moreover, because a virus entering a cell surface has a deterministic motion, we can assume that the initial position satisfies $R_0 < R$ so that we can neglect any boundary effects and use the open wedge approximation, which consists of using formula (3.30) without the boundary layer term. Actually, this approximation is not that restrictive because after the first iteration process (movement along the microtubule followed by the particle release), the boundary layer term is negligible compared to the other term.

To obtain an explicit expression for the amplitude B, we consider the successive approximations

$$r_m(R_0) \approx R_0 \left(1 + \frac{\Theta^2}{12}\right) \tag{4.9}$$

and

$$\begin{aligned}
&R_0 = R_0;\\
&R_1 \simeq R_0\left(1+\frac{\Theta^2}{12}\right) - d_m;\\
&R_2 \simeq R_0\left(1+\frac{\Theta^2}{12}\right)^2 - d_m\left(1+\left(1+\frac{\Theta^2}{12}\right)\right);\\
&\quad\vdots\\
&R_i \simeq R_0\left(1+\frac{\Theta^2}{12}\right)^i - d_m\left(\sum_{k=0}^{i-1}\left(1+\frac{\Theta^2}{12}\right)^k\right);
\end{aligned}$$

that is,

$$R_i \simeq \left(R_0 - \frac{12d_m}{\Theta^2}\right)\left(1+\frac{\Theta^2}{12}\right)^i + \frac{12d_m}{\Theta^2}. \tag{4.10}$$

Thus the particle reaches the nucleus after n iteration steps which approximatively satisfies $R_n = \delta$,

$$n \simeq \frac{\ln\left(\frac{1-\frac{\delta\Theta^2}{12d_m}}{1-\frac{R_0\Theta^2}{12d_m}}\right)}{\ln\left(1+\frac{\Theta^2}{12}\right)} \approx \frac{R_0-\delta}{d_m} + o\left(1\right). \tag{4.11}$$

If T_n denotes the mean time a viral particle takes to reach the nucleus, then using formula (3.9), we obtain

$$T_n \simeq n.t_m + \frac{\left(\frac{\tan(\Theta)}{\Theta}-1\right)}{4D}\sum_{i=0}^{n-1} R_i^2, \tag{4.12}$$

that is,

$$\begin{aligned}
t \simeq n.t_m + &\frac{\left(\frac{\tan(\Theta)}{\Theta}-1\right)}{4D}\\
&\times\sum_{i=0}^{n-1}\left(\left(\frac{12d_m}{\Theta^2}\right)^2 + 2\left(\frac{12d_m}{\Theta^2}\right)\left(R_0-\frac{12d_m}{\Theta^2}\right)\left(1+\frac{\Theta^2}{12}\right)^i\right.\\
&\qquad\left. + \left(R_0-\frac{12d_m}{\Theta^2}\right)^2\left(1+\frac{\Theta^2}{12}\right)^{2i}\right),
\end{aligned}$$

$$\begin{aligned}
T_n \simeq nt_m + &\frac{\left(\frac{\tan(\Theta)}{\Theta}-1\right)}{4D}\\
&\times\left(n\left(\frac{12d_m}{\Theta^2}\right)^2 - \left(\frac{24d_m}{\Theta^2}\right)\left(R_0-\frac{12d_m}{\Theta^2}\right)\frac{1-\left(1+\frac{\Theta^2}{12}\right)^n}{\frac{\Theta^2}{12}}\right.\\
&\qquad\left. + \left(R_0-\frac{12d_m}{\Theta^2}\right)^2\frac{1-\left(1+\frac{\Theta^2}{12}\right)^{2n}}{1-\left(1+\frac{\Theta^2}{12}\right)^2}\right).
\end{aligned}$$

For $\Theta \ll 1$, a Taylor expansion gives that

$$T_n \simeq \left(\frac{R_0-\delta}{d_m}\right) t_m + \frac{t_m\left(R_0-\delta\right)}{24 d_m}\left(1+\frac{R_0+\delta}{d_m}\right)\Theta^2 + \frac{\left(R_0-\delta\right)}{72D}\left(d_m + 3\left(R_0+\delta\right) + \frac{2\left(R_0^2+R_0\delta+\delta^2\right)}{d_m}\right)\Theta^4 + o\left(\Theta^4\right).$$

In small diffusion limit $D \ll 1$, $\Theta \ll 1$, the velocity is $B \simeq \frac{R_0-\delta}{T_n}$, and consequently we obtain for $R_0 \approx R$ a second order approximation,

$$B \approx \frac{\frac{d_m}{t_m}}{1+\left(1+\frac{R+\delta}{d_m}\right)\frac{\Theta^2}{24}+O\left(\Theta^4\right)}, \tag{4.13}$$

where d_m, t_m are the mean distance and the mean time a virus stays on the microtubule, R (resp., δ) is the radius of the cell (resp., nucleus) and $\Theta = \frac{2\pi}{N}$, where N is the total number of microtubules.

4.3. Justification of the MFPT criteria. To justify the use of the MFPT criteria to estimate the steady state drift, we run numerical simulations of 1,000 viruses inside a two-dimensional domain Ω ($\delta < r < R$) with intermittent dynamics, alternating between epochs of free diffusion and directed motion along microtubules, and compare the steady state distribution with the one obtained by solving the Fokker–Planck equation for viruses whose trajectories are described by the effective stochastic equation (2.2) with our computed constant drift

$$\mathbf{b}\left(\mathbf{X}\right) = -\frac{\frac{d_m}{t_m}}{1+\left(1+\frac{R+\delta}{d_m}\right)\frac{\Theta^2}{24}}\frac{\mathbf{r}}{|\mathbf{r}|} = -B\frac{\mathbf{r}}{|\mathbf{r}|}. \tag{4.14}$$

We imposed reflecting boundary conditions at the nuclear and the external membrane. The theoretical normalized steady state distribution ρ satisfies

$$D\Delta\rho - \nabla.[\mathbf{b}\rho] = 0 \text{ in } \Omega,$$
$$\frac{d\rho}{dr}\left(R\right) = \frac{d\rho}{dr}\left(\delta\right) = 0,$$

and the solution ρ is given by

$$\rho(r) = \frac{e^{-\frac{Br}{D}}}{\int_\delta^R e^{-\frac{Br}{D}} 2\pi r dr} = \frac{e^{-\frac{Br}{D}}}{2\pi\frac{D}{B}(\delta e^{-\frac{B\delta}{D}} - Re^{-\frac{BR}{D}} + \frac{D}{B}(e^{-\frac{B\delta}{D}} - e^{-\frac{BR}{D}}))}. \tag{4.15}$$

The result of both distributions is presented in Figure 4.1, where we can observe that both curves match very nicely. This result shows that the criteria we have used is at least enough to recover the distribution. For the simulations, we consider that the directed run of the virus along a microtubule (loaded by dynein) lasts $t_m = 1s$ and covers a mean distance $d_m = 0.7\mu m$ [13]. The diffusion constant is $D = 1.3\mu m^2 s^{-1}$, as observed for the adeno-associated virus [14]. The two curves in Figure 4.1 fit very nicely except at the neighborhood of the nuclear membrane, where the simulation of the empirical distribution is plagued with a possible boundary layer. Another source of discrepancy comes from the difference of behavior of viruses far from and close to the nucleus: viruses far from the nucleus do not bind as often as those located in its neighborhood. Consequently, a constant effective drift cannot account for the radial geometry near the nucleus. A theory for radius-dependent effective drift has been derived in [15].

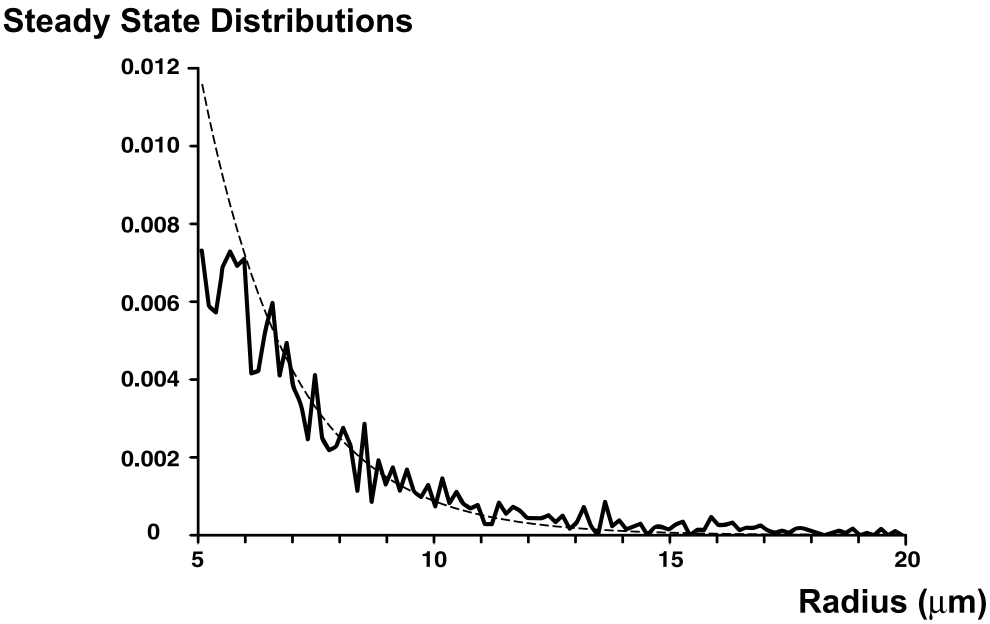


FIG. 4.1. *Steady state distributions. We show the empirical steady state distribution for* 1,000 *viral trajectories with an intermittent dynamic (solid line). The theoretical distribution of viruses whose trajectories are described by the stochastic equation* (2.2) *is shown by the dashed line. Geometrical parameters are* $R = 20\mu m$, $\delta = 5\mu m$, *and* $\Theta = \frac{\pi}{24}$.

5. Conclusion. For the limit of a cell containing an excess of microtubules, we have presented here a model to describe the motion of biological particles such as viruses, vesicles, and many others moving inside the cell cytoplasm by a complex combination of Brownian motion and deterministic drift. Our procedure consists mainly of approximating an alternative switching mode between diffusion and deterministic drift epochs by a steady state stochastic equation; it also consists of estimating the amplitude of the effective drift and is based on the criteria that the MFPTs to the nucleus computed in both cases are equal. In that case, this amplitude accounts for the directed transport along microtubules, the cell geometry, and the binding constants. The model has, however, several limitations. First, we do not take into account directly the backward movement of the virus along the microtubules [16, 17], which can affect the mean time and the amplitude of the drift. Second, the present computations are given only for two-dimensional cell geometry. It can still be applied to many in vitro culture cells; however, it is not clear how to generalize our approach to a three-dimensional cell geometry. For example, to study the trafficking inside cylindrical axons or dendrites of neuronal cells, a different approach should include these geometrical features. However, despite these real difficulties, the present model may be used to analyze plasmid transport in a host cell, at the molecular level, which is one of the fundamental limitations of gene delivery [18, 19, 20, 21].

Appendix. In this appendix, we provide an explicit computation of integral (3.21) using the method of the residues. This method was previously used in a similar context in [12, p. 386]. We denote by $\left(p_j^k\right)_{j\geq 0}$ the poles of the function

$$\Phi : s \to \frac{I_k\left(x\left(s\right) \wedge x_0\left(s\right)\right)\left(I_k^{'}\left(x_+\left(s\right)\right)K_k - K_k^{'}\left(x_+\left(s\right)\right)I_k\right)\left(x\left(s\right) \vee x_0\left(s\right)\right)}{sI_k^{'}\left(x_+\left(s\right)\right)}e^{st},$$

where ($x(s) = r\sqrt{\frac{s}{D}}$, $x_0(s) = r_0\sqrt{\frac{s}{D}}$, and $x_+(s) = R\sqrt{\frac{s}{D}}$). The associated residues are $\left(r_j^k\right)_{j\geq 0}$. We now compute the residues explicitly.

To identify the poles, we recall the relation between the k-order Bessel function J_k (that is true for z such that $-\pi < \arg(z) < \frac{\pi}{2}$) and the modified Bessel functions I_k [11, p. 375]:

$$I_k(z) = e^{-\frac{1}{2}k\pi i} J_k\left(z e^{\frac{1}{2}\pi i}\right). \tag{5.1}$$

All roots $\alpha_{j,k}$ of the equations

$$J_k^{'}(R\alpha) = 0$$

are real, simple, and strictly positive [11, p. 370] because k is real and

$$k \leq \alpha_{1,k} < \alpha_{2,k} \ldots .$$

Thus,

$$I_k^{'}(-iR\alpha_{j,k}) = 0.$$

Finally, the poles of Φ are simple, given by $p_0^k = 0$ and that for all $j \geq 1$, $p_j^k = -D\alpha_{j,k}^2$. Consequently the associated residues are given for each k for all $j \geq 0$ by

$$r_j^k = \lim_{s\to p_j^k} \left(s - p_j^k\right)\Phi(s). \tag{5.2}$$

Then using the residues, integral (3.21) is given by

$$I(r,\theta,t) = \frac{1}{\Theta\pi Di}\sum_k \sin(k\theta)\sin(k\theta_0)(2\pi i)\sum_{j\geq 0} r_j^k = \frac{2}{\Theta D}\sum_k \sin(k\theta)\sin(k\theta_0)\sum_{j\geq 0} r_j^k.$$

We now compute the residues r_j^k. The residue r_0^k is associated with the pole $p_0^k = 0$ and given by

$$r_0^k = \lim_{s\to 0} s\Phi(s).$$

Using the following identities on the modified Bessel functions [12, p. 489],

$$I_k^{'}(z) = I_{k+1}(z) + \frac{k}{z}I_k(z) \text{ and } K_k^{'}(z) = -K_{k-1}(z) - \frac{k}{z}K_k(z),$$

and substituting the derivatives $I_k^{'}$ and $K_k^{'}$ in the expression of Φ, we get

$$\begin{aligned} r_0^k = \lim_{s\to 0} & \frac{I_k(x(s)\wedge x_0(s))}{\left(I_{k+1} + \frac{k}{x_+(s)}I_k\right)(x_+(s))} \\ & \times\left(\left(\left(I_{k+1} + \frac{k}{x_+(s)}I_k\right)(x_+(s))K_k\right)\right. \\ & \left. + \left(\left(K_{k-1} + \frac{k}{x_+(s)}K_k\right)(x_+(s))I_k\right)\right)(x(s)\vee x_0(s)). \end{aligned}$$

Taking into account only the dominant terms, we get

$$r_0^k = \lim_{s \to 0} \frac{I_k\left(x\left(s\right) \wedge x_0\left(s\right)\right)\left(I_k\left(x_+\left(s\right)\right)K_k + K_k\left(x_+\left(s\right)\right)I_k\right)\left(x\left(s\right) \vee x_0\left(s\right)\right)}{I_k\left(x_+\left(s\right)\right)}.$$

To further compute this limit, we use the Taylor expansions of I_k and K_k [11, p. 375] expressed in terms of the Γ function:

$$I_k\left(z\right) \approx \frac{\left(\frac{1}{2}z\right)^k}{\Gamma\left(k+1\right)} \text{ and } K_k\left(z\right) \approx \frac{1}{2}\Gamma\left(k\right)\left(\frac{1}{2}z\right)^{-k}.$$

For $r < r_0$, we get

$$r_0^k = \lim_{s \to 0} \frac{\frac{\left(\frac{1}{2}\left(x\left(s\right)\right)\right)^k}{\Gamma(k+1)}\left(\frac{\left(\frac{1}{2}\left(x_+\left(s\right)\right)\right)^k}{\Gamma(k+1)}\frac{1}{2}\Gamma\left(k\right)\left(\frac{1}{2}\left(x_0\left(s\right)\right)\right)^{-k} + \frac{1}{2}\Gamma\left(k\right)\left(\frac{1}{2}\left(x_+\left(s\right)\right)\right)^{-k}\frac{\left(\frac{1}{2}\left(x_0\left(s\right)\right)\right)^k}{\Gamma(k+1)}\right)}{\frac{\left(\frac{1}{2}\left(x_+\left(s\right)\right)\right)^k}{\Gamma(k+1)}}.$$

Finally, using the relation $\Gamma\left(k+1\right) = k\Gamma\left(k\right)$ and the expressions of $x(s)$, $x_0(s)$, and $x_+(s)$, we get

$$r_0^k = \frac{r^k\left(r_0^{2k} + R^{2k}\right)}{2kR^{2k}r_0^k}.$$

The computation of the other residues $\left(r_j^k\right)_{j \geq 1}$ is slightly different,

$$r_j^k = \lim_{s \to p_j^k}\left(s - p_j^k\right)\Phi(s),$$

where $p_j^k = -D\alpha_{j,k}^2$. Using the Wronskian relation [12, p. 489],

$$I_k\left(z\right)K_k'\left(z\right) - K_k\left(z\right)I_k'\left(z\right) = -\frac{1}{z},$$

we now substitute

$$K_k^{'}\left(z\right) = \frac{-\frac{1}{z} + K_k\left(z\right)I_k^{'}\left(z\right)}{I_k\left(z\right)}.$$

In the expression of Φ, we get

$$r_j^k = \lim_{s \to p_j^k}\frac{\left(s - p_j^k\right)e^{st}}{s}\frac{I_k\left(x\left(s\right)\right)\left(I_k^{'}\left(x_+\left(s\right)\right)K_k - \left(\frac{-\frac{1}{x_+(s)} + K_k I_k^{'}}{I_k}\right)\left(x_+\left(s\right)\right)I_k\right)\left(x_0\left(s\right)\right)}{I_k^{'}\left(x_+\left(s\right)\right)}.$$

Because

$$\lim_{s \to p_j^k} I_k^{'}\left(x_+\left(s\right)\right) = I_k^{'}\left(x_+\left(p_j^k\right)\right) = 0,$$

we obtain the expression for the residues:

$$r_j^k = \frac{e^{p_j^k t}}{p_j^k}\frac{I_k\left(x\left(p_j^k\right)\right)I_k\left(x_0\left(p_j^k\right)\right)}{I_k\left(x_+\left(p_j^k\right)\right)x_+\left(p_j^k\right)}\lim_{s \to p_j^k}\frac{\left(s - p_j^k\right)}{I_k^{'}\left(x_+\left(s\right)\right)}.$$

Finally, since

$$\lim_{s\to p_j^k} \frac{\left(s-p_j^k\right)}{I_k^{'}\left(x_+(s)\right)} = \frac{2\sqrt{Dp_j^k}}{R} \lim_{s\to p_j^k} \frac{x_+(s)-x_+\left(p_j^k\right)}{I_k^{'}\left(x_+(s)\right)-I_k^{'}\left(x_+\left(p_j^k\right)\right)} = \frac{2\sqrt{Dp_j^k}}{RI_k^{''}\left(x_+\left(p_j^k\right)\right)},$$

we obtain

$$r_j^k = \frac{e^{p_j^k t}}{p_j^k} \frac{I_k\left(x\left(p_j^k\right)\right) I_k\left(x_0\left(p_j^k\right)\right)}{I_k\left(x_+\left(p_j^k\right)\right) x_+\left(p_j^k\right)} \frac{2\sqrt{Dp_j^k}}{RI_k^{''}\left(x_+\left(p_j^k\right)\right)}.$$

To simplify this expression, we use that I_k satisfies the differential equation [11, p. 374]

$$I_k^{''}(z) + \frac{1}{z} I_k^{'}(z) - \left(1 + \frac{k^2}{z^2}\right) I_k(z) = 0.$$

Thus for $z = x_+\left(p_j^k\right)$,

$$I_k^{''}\left(x_+\left(p_j^k\right)\right) = \frac{p_j^k R^2 + Dk^2}{p_j^k R^2} I_k\left(x_+\left(p_j^k\right)\right).$$

We get

$$r_j^k = \frac{2De^{p_j^k t}}{R^2 p_j^k + Dk^2} \frac{I_k\left(x\left(p_j^k\right)\right) I_k\left(x_0\left(p_j^k\right)\right)}{I_k^2\left(x_+\left(p_j^k\right)\right)},$$

and finally, using (5.1), we get

$$r_j^k = \frac{2e^{-D\alpha_{j,k}^2 t}}{-R^2\alpha_{j,k}^2 + k^2} \frac{J_k\left(r\alpha_{j,k}\right) J_k\left(r_0\alpha_{j,k}\right)}{J_k^2\left(R\alpha_{j,k}\right)}.$$

Integral (3.21) is given by

$$I(r,\theta,t) = \frac{2}{\Theta D} \sum_k \sin(k\theta)\sin(k\theta_0) \sum_{j\geq 0} r_j^k = \frac{2}{\Theta D}\left(S_1(r,\theta,t) + S_2(r,\theta,t)\right), \tag{5.3}$$

where

$$S_1(r,\theta,t) = \sum_k \sin(k\theta)\sin(k\theta_0) \frac{r^k\left(r_0^{2k} + R^{2k}\right)}{2kR^{2k}r_0^k},$$

$$S_2(r,\theta,t) = -2\sum_k \sin(k\theta)\sin(k\theta_0) \sum_{j=1}^{\infty} e^{-D\alpha_{j,k}^2 t} \frac{J_k\left(r\alpha_{j,k}\right) J_k\left(r_0\alpha_{j,k}\right)}{\left(R^2\alpha_{j,k}^2 - k^2\right) J_k^2\left(R\alpha_{j,k}\right)}.$$

REFERENCES

[1] C. M. Wiethoff and C. R. Middaugh, *Barriers to non-viral gene delivery*, J. Pharmaceutical Sci., 92 (2003), pp. 203–217.

[2] D. Dauty and A. S. Verkman, *Actin cytoskeleton as the principal determinant of size-dependent DNA mobility in cytoplasm: A new barrier for non-viral gene delivery*, J. Biol. Chem., 280 (2005), pp. 7823–7828.

[3] A. T. Dinh, T. Theofanous, and S. Mitragotri, *A model for intracellular trafficking of adenoviral vectors*, Biophys. J., 89 (2005), pp. 1574–1588.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed., Garland, New York, 2002.

[5] D. Holcman, *Modeling trafficking of a virus and a DNA particle in the cell cytoplasm*, J. Statist. Phys., 127 (2007), pp. 471–494.

[6] Z. Schuss, *Theory and Applications of Stochastic Differential Equations*, John Wiley & Sons, New York, 1981.

[7] N. Hirokawa, *Kinesin and dynein superfamily proteins and the mechanism of organelle transport*, Science, 279 (1998), pp. 519–526.

[8] R. Mallick, *Cytoplasmic dynein functions as a gear in response to load*, Nature, 427 (2004), pp. 649–652.

[9] S. Redner, *A Guide to First Passage Processes*, Cambridge University Press, Cambridge, UK, 2001.

[10] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 3, John Wiley & Sons, New York, 1977.

[11] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[12] H. S. Carslaw and J. C. Jaegger, *Conduction of Heat in Solids*, Oxford University Press, Oxford, UK, 1959.

[13] S. J. King and T. A. Schroer, *Dynactin increases the processivity of the cytoplasmic dynein motor*, Nat. Cell Biol., 2 (2000), pp. 20–24.

[14] G. Seisenberger, M. Ried, T. Endress, H. Büning, M. Hallek, and C. Bräuchle, *Real-time single-molecule imaging of the infection pathway of an adeno-associated virus*, Science, 294 (2001), pp. 1929–1932.

[15] T. Lagache and D. Holcman, *Quantifying intermittent transport in cell cytoplasm*, Phys. Rev. E, submitted.

[16] D. Katinka, N. Claus-Henning, and B. Sodeik, *Viral stop-and-go along microtubules: Taking a ride with dynein and kinesins*, Trends in Microbiology, 13 (2005), pp. 320–327.

[17] S. P. Gross, M. A. Welte, S. M. Block, and E. F. Wieschaus, *Dynein-mediated cargo transport in vivo: A switch controls travel distance*, J. Cell Biol., 5 (2000), pp. 945–955.

[18] G. R. Whittaker, *Virus nuclear import*, Adv. Drug Delivery Rev., 55 (2003), pp. 733–747.

[19] D. A. Dean, R. C. Geiger, and R. Zhou, *Intracellular trafficking of nucleic acids*, Expert Opinion Drug Delivery, 1 (2004), pp. 127–140.

[20] E. M. Campbell and T. J. Hope, *Gene therapy progress and prospects: Viral trafficking during infection*, Gene Therapy, 12 (2005), pp. 1353–1359.

[21] D. Luo and W. M. Saltzman, *Synthetic DNA delivery systems*, Nature Biotechnology, 18 (1999), pp. 33–37.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1168–1201

OPTIMAL LIQUIDATION BY A LARGE INVESTOR*

AJAY SUBRAMANIAN†

Abstract. We develop a partial equilibrium model to investigate the problem of optimal liquidation over a finite or infinite time horizon for an investor with large holdings in a risky asset. The imperfect liquidity in the market for the asset leads to a nonlinear path dependent feedback on the underlying asset price process due to the large investor's trades and his holdings in the asset. We use probabilistic techniques to prove verification and existence results for optimal liquidation policies for the utility-maximizing investor under broad assumptions. In particular, our results imply the existence of optimal policies if the investor has power utility functions. We provide analytical expressions for the optimal policy when the large investor has logarithmic preferences. We use these results to characterize the "liquidity discount," which is a measure of the liquidity risk of the large investor's position in the risky asset.

Key words. liquidation, large investors, stochastic control, duality

AMS subject classifications. 91B28, 93E20, 46N10

DOI. 10.1137/050643714

1. Introduction.

Traditional finance theory is based on the competitive market paradigm which assumes that markets are perfectly elastic, that is, that all investors are price-takers who can buy or sell as many shares of a security as they want without affecting its price and purchases/sales have immediate execution. The absence of these conditions is sometimes labeled *liquidity risk*. Liquidity risk is particularly important for thinly traded securities or in scenarios in which investors with comparatively large positions in a security wish to liquidate their holdings in a short period of time. Examples of such investors are insurance firms who face regulatory minimum capital requirements, financial institutions who face "value at risk" constraints, mutual or hedge fund managers who must make large trades in a short period of time in the face of redemptions by investors, and financially distressed firms who must liquidate large asset holdings to meet contractual debt obligations.[1]

Motivated by the above considerations, several studies attempt to provide characterizations of liquidity risk through the investigation of the optimal liquidation problem for an investor in an illiquid underlying asset market (see, for example, Bertsimas and Lo [3], Almgren and Chriss [1], Duffie and Ziegler [8], and Subramanian and Jarrow [30]). These studies use different frameworks to examine the optimal liquidation policy of a large investor with holdings in a single illiquid asset or a portfolio of such assets where trades by the investor may affect the underlying asset price processes. The liquidity risk is characterized in terms of the value (or utility) the investor derives from liquidating her portfolio over a finite time horizon.

In this article, we contribute to this aspect of the literature by examining the optimal liquidation problem over a finite or infinite time horizon for an investor with large holdings in a single risky asset where the investor's trades and holdings in the

*Received by the editors October 28, 2005; accepted for publication (in revised form) September 27, 2007; published electronically March 28, 2008.
http://www.siam.org/journals/siap/68-4/64371.html

†Department of Risk Management and Insurance, J. Mack Robinson College of Business, Georgia State University, 35 Broad Street, University Plaza, Atlanta, GA 30303 (insasu@langate.gsu.edu).

[1] The importance of liquidity risk is highlighted by the turmoil in global capital markets in the late 1990's when Long Term Capital Management (LTCM) was forced to unwind its large positions in several securities in a very short period of time in the aftermath of the Asian currency crisis.

asset may have a *nonlinear and path dependent* feedback on the underlying asset price process. We consider the situation where there are no execution lags for trade orders and the investor trades continuously, as in the continuous auction models of Kyle [21] and Back [2]. The investor's liquidation policy is characterized by a *liquidation rate* process, which is a mathematically tractable representation of the effect of insider trading restrictions that prevent the investor from making large trades on a single date. The absence of execution lags makes this a viable continuous time extension of a more realistic model where investors trade only at discrete, but "closely spaced," instants of time. There could be a nonlinear path dependent feedback on the underlying asset price process due to the investor's holdings in the asset as well as her liquidation rate. In particular, this general formulation allows for both permanent and temporary price impacts. The underlying asset price process is continuous and, similar to the models of Kyle [21], Back [2], and Cuoco and Cvitanic [5], the investor's liquidation policy affects only the expected return of the asset. As in these studies, the volatility of the asset is due to the presence of "small" or "liquidity" traders.

We adopt a *partial equilibrium* approach where the impact of the large investor's trades on the asset price is exogenously specified as opposed to the general equilibrium framework adopted in Kyle [21], where the impact is derived endogenously through equilibrium considerations. This facilitates the investigation of a much broader class of models of the effect of large investors on financial markets.

We use probabilistic "martingale" techniques to prove verification and existence results for optimal liquidation policies for the investor. Our analysis leads to an explicit characterization of sufficient conditions on the investor's preferences and the feedback function describing the impact of the large investor's trades on the asset price, under which optimal liquidation policies exist. The identification of such conditions that ensure the existence of an optimal policy and, therefore, a partial equilibrium, is important from an economic standpoint since it establishes the *viability* of the class of partial equilibrium models we propose as an appropriate description of large investor behavior. We use the results of our analysis to derive explicit analytical expressions for the investor's optimal liquidation policy and his value function when he has logarithmic preferences. We then characterize the *liquidity discount*, which is a measure proposed by Subramanian and Jarrow [30] to quantify the liquidity risk of a large investor's position in a risky asset. As they discuss, the liquidity discount could be used to modify value at risk measures to also incorporate liquidity risk.

From a mathematical standpoint, the probabilistic methodology we employ is similar to that adopted in earlier papers that examine optimal consumption/investment or pricing/hedging problems in incomplete markets within frameworks where dynamic programming techniques may be difficult to apply.[2] In particular, we show the existence of optimal liquidation policies by establishing the duality between the investor's liquidation problem and a dual control problem.

Apart from the fact that the problem we consider is different from those examined by previous studies in this area, there are some important aspects of our framework that make the application of the duality methodology significantly different from the analyses in these studies. Since we examine the optimal liquidation problem of an investor with holdings in a single risky asset, the only feasible trades are *sales* of

[2] See, for example, He and Pearson [14], Cvitanic and Karatzas [6], Cvitanic and Ma [7], Karatzas and Kou [16], El Karoui, Peng, and Quenez [11], Buckdahn and Hu [4], Cuoco and Cvitanic [5], El Karoui and Jeanblanc [10], Kramkov and Schachermayer [20], Ma and Yong [22], Follmer and Leukert [12], and Mnif and Pham [25]. See Pritsker [27], He and Mamaysky [15], and Polimenis [26] for frameworks with large traders where dynamic programming techniques are employed.

the asset. As we will see later, this feature causes the set of feasible dual processes to be *unbounded*. This leads to nontrivial complications in the demonstration of verification and existence results for solutions to the dual control problem and hence the primal one. Further, the investor derives utility from the periodic liquidation of her holdings in the asset, but her liquidation rate affects the underlying asset price process. Finally, the "wealth process" of the investor that is the market value of her holdings in the asset is always nonnegative. Therefore, our study also contributes to the existing literature from a mathematical standpoint by illustrating the application of the convex duality methodology to a stochastic control problem where the typical assumptions and features of the frameworks examined by prior studies do not hold.

We analyze the optimal liquidation problem of a large investor over a finite as well as an infinite time horizon. The finite time horizon formulation is particularly relevant in the context of the motivating examples discussed earlier in which large investors such as insurance firms, financial institutions, mutual and hedge funds, and financially distressed firms have to liquidate large asset holdings over short periods of time. Insurance firms have to liquidate large holdings in risky assets to make liability payments in the aftermath of catastrophic events such as the September 11th terrorist attack and Hurricane Katrina. Given the time taken to file and process liability claims, the liquidation horizon for such firms would be expected to be of the order of 6 months to a year. For financial services firms who face minimum capital requirements or "value at risk" constraints, the relevant liquidation horizon could be much shorter (10 days to a month).[3] Mutual funds and hedge funds that face redemptions by investors, especially in bear markets, typically have to liquidate large holdings over a period of one to two months. Finally, financially distressed firms often have to liquidate large asset holdings over short periods of time (3–6 months) to meet contractual debt obligations.

In the infinite time horizon version of the model, the investor does not self-impose a possibly suboptimal liquidation horizon, but optimally liquidates his asset holdings until they fall to zero. The theoretical advantages of the infinite time horizon formulation are, however, mitigated to some extent by two important observations. First, as discussed above, the choice of liquidation horizon may not always be at the discretion of the investor and could be a consequence of regulatory requirements or market considerations. Second, it follows from standard convergence arguments that one can choose a sufficiently large finite liquidation horizon such that the investor's optimal value function in the finite time horizon model is arbitrarily close to the optimal value function in the infinite time horizon model. In our view, therefore, the finite time horizon and infinite time horizon analyses are complementary.

The plan for the paper is as follows. Section 2 presents the model in which a large investor liquidates his holdings over a finite time horizon. In section 3, we formulate the large investor's dual control problem. In section 4, we prove verification and existence theorems for optimal liquidation policies. In section 5, we explicitly derive the optimal policies for an investor with logarithmic preferences and characterize the liquidity discount. In section 6, we analyze two extensions of the model: the scenario in which the investor could incur costs from holding nonzero wealth in the asset at the terminal date, and the scenario in which the liquidation horizon is infinite and the investor liquidates his shares until they fall to zero. Section 7 concludes the paper. Detailed proofs are in the appendix.

2. The model. We consider a time horizon $[0, T]$ and a probability space (Ω, F, P) with a complete and augmented filtration $\{F_t\}$ generated by a Brownian

[3] In fact, standard "value at risk" measures presume a liquidation horizon of ten days.

motion B_t with $F_T \equiv F$. The number of shares of the asset that the large investor holds at any date is represented by the nonnegative F_t-adapted process $N(.)$. The large investor's intertemporal liquidation process is absolutely continuous; that is, the large investor liquidates at the rate $n(.)$ so that $n(t)\,dt$ shares of the asset are liquidated over the time interval $[t, t+dt]$. We examine the scenario where there are no execution lags and transaction costs are insignificant (see Almgren and Chriss [1] or Subramanian and Jarrow [30] for models that incorporate execution lags and transaction costs). The process $n(.)$ may be random but is assumed to be $\{F_t\}$-progressively measurable, reflecting the assumption that the large investor's choice trade at a certain time depends on his information at that time and cannot anticipate the future. Thus,

$$N(t) = N(0) - \int_0^t n(s)\,ds.$$

The presence of insider trading restrictions and imperfect liquidity in financial markets prevents investors from making large trades on a single date. The assumption of an absolutely continuous liquidation process is a mathematically tractable representation of the effect of such restrictions. Hence, the investor may only liquidate his large position in the risky asset over an extended time horizon. Further, the fact that we examine the scenario where execution lags are insignificant implies that this is a viable continuous time extension of a more realistic model where the investor trades at discrete, but closely spaced, instants of time.

The asset price process $S(.)$ satisfies the following stochastic differential equation:

$$dS(t,\omega) = S(t,\omega)(\overline{\mu}(t,\omega) + \overline{\overline{\mu}}(W(t,\omega), c(t,\omega), t, \omega))\,dt + S(t,\omega)\sigma(t,\omega)\,dB_t, \tag{2.1}$$

where $W(t,\omega) = N(t,\omega)S(t,\omega)$ represents the *market value* of the large investor's holdings in the asset and $c(t,\omega) = n(t,\omega)S(t,\omega)\,dt$ is the large investor's cash flow from liquidating $n(t,\omega)$ shares at time t. Equation (2.1) implies that the asset price process $S(.)$ may be affected by the large investor's liquidation rate $n(.)$ as well as his wealth in the asset. The feedback function $\overline{\overline{\mu}}$ may depend explicitly on time and the probability space parameter ω and is therefore, in general, path dependent. Therefore, the large investor's past liquidation policy may affect his present and future cash flows from liquidation. This formulation allows us to incorporate both permanent and transient price impacts of the large investor's trades. The restrictions on $\overline{\mu}, \overline{\overline{\mu}}, \sigma$ that ensure the existence of a unique strong solution to (2.1) are specified later. We refer to the process $c(.)$ as the *liquidation rate* process and the process $W(.)$ as the *wealth process.*

ASSUMPTION 1. *We assume that* $\overline{\mu}(t,\omega)$ *is uniformly bounded and* F_t*-progressively measurable and* $\sigma(.)$ *and* $\sigma^{-1}(.)$ *are uniformly bounded and* F_t*-progressively measurable. In the following, we shall occasionally drop the explicit dependence on the probability space parameter* ω *wherever there is no danger of confusion. From* (2.1)*, we see that the process* $W(.)$ *evolves as follows:*

$$dW(t) = W(t)[\overline{\mu}(t) + \overline{\overline{\mu}}(W(t), c(t), t)]\,dt + W(t)\sigma(t)\,dB(t) - c(t)\,dt. \tag{2.2}$$

ASSUMPTION 2. *A feasible liquidation strategy is a nonnegative process* $(W(.), c(.))$ *satisfying*

$$\begin{aligned} &\int_0^T |W(t)(\overline{\mu}(t) + \overline{\overline{\mu}}(W(t), c(t), t))\mid dt + \int_0^T \left|(W(t)\sigma(t))^2\right| dt < \infty \ \textit{a.s.},\\ &\int_0^T c(t)\,dt < \infty \ \textit{a.s.} \end{aligned} \tag{2.3}$$

The conditions in (2.3) *basically ensure that the process* $W(.)$ *described by* (2.2) *is well defined. Alternatively, we say that the terminal wealth-liquidation rate process* $(W, c(.))$ *is feasible if* $W(T) = W$*, where* $W(.)$ *is the wealth process corresponding to the liquidation rate process* $c(.)$ *and* $(W(.), c(.))$ *is feasible. We denote the set of feasible liquidation strategies* $(W(.), c(.))$ *by* Θ.

ASSUMPTION 3. *The large investor has a time-additive utility function for intertemporal cash flows and a utility of terminal wealth defined by*

$$U(c, W) = E\left[\int_0^T u_1(c(t), t))\, dt\right] + E[u_2(W, T)]. \tag{2.4}$$

Remark 1. We allow for the possibility that the investor may derive nonzero utility from terminal wealth for the sake of generality. As we discuss in Remark 4 in section 4, all our results hold even when the investor derives no utility from remaining holdings in the asset at date T so that he liquidates his entire position over the interval $[0, T)$. The explicit time dependence of u_1, u_2 allows for the incorporation of costs borne by the investor from delaying liquidation of his holdings. In section 6.2, we analyze the scenario in which the time horizon $T = \infty$.

The objective function (2.4) is well defined for liquidation rate–terminal wealth processes satisfying

$$E\left[\int_0^T u_1(c(t), t)^-\, dt\right] < \infty; \quad E[u_2(W, T)^-] < \infty. \tag{2.5}$$

We assume that the large investor's liquidation rate process and the terminal wealth are *admissible* if they satisfy the above condition.

Note. If $u_1(c, t) \geq 0, u_2(W, T) \geq 0$ for all (c, t) and W, then (2.5) is trivially satisfied.

The investor's objective function (2.4), which implies that he has a concave, Von Neumann–Morgenstern utility function over his intertemporal liquidation rate $c(.)$ and terminal wealth $W(.)$, is justified as follows.

First, recall that the investor's cash flow from liquidating shares at date t is $c(t)\, dt$. This cash flow could be used for current consumption as well as to finance future consumption. If the cash flow is used for current consumption, then the investor's utility payoff is $u_1(c(t), t)$. If a portion of the cash flow is used to finance future consumption (by possibly trading in other securities), then, like the terminal utility function$u_2(W, T)$, the function $u_1(c(t), t)$ should be viewed as an "indirect" utility function over cash flows as in Chapter 6 of Mas-Colell, Whinston, and Green [23].

Second, as discussed in Mas-Colell, Whinston, and Green [23], the objective function (2.4) could more generally be viewed as corresponding to the scenario in which the investor has convex preferences over *monetary lotteries*; that is, the investor's preferences are directly defined over cash flow streams.

Third, the concavity of the function u_1 also incorporates the possible presence of "direct" liquidation/transaction costs that are convex in the size of the trade (recall that the model incorporates "indirect" costs through the effects of trades on the asset price process—see (2.1)). In the absence of friction due to illiquidity, the payoff to a risk-neutral investor would be $c(t)\, dt$. In an illiquid market, the payoff would be $c(t)\, dt$ less liquidation/transaction costs that are convex in the size of the trade, which implies that the *net* payoff to the investor is concave. Hence, even if the investor were risk-neutral, his payoffs from liquidation would be concave.

Fourth, the objective function (2.4) incorporates precautionary motives of the investor to smooth liquidation proceeds over time and across states in order to lower the direct and indirect costs he faces due to the illiquidity of the underlying asset.

A possible alternative objective function is one in which the investor maximizes an increasing, concave function of his discounted "cumulative" liquidation proceeds; that is, the investor maximizes

$$Eu\left[\int_0^T e^{-lt}c(t)\,dt\right],$$

where l is a discount rate and u is increasing and concave. The investor, therefore, derives the same utility from liquidation proceeds that have the same discounted present value; that is, the *timing* of intertemporal liquidation cash flows along a particular sample path does not matter. Adapting the arguments in Fudenberg, Holmstrom, and Milgrom [13] to our setting, such an objective function is, in general, justified from a decision-theoretic standpoint when the investor has unlimited access to credit markets from which he can borrow against future liquidation proceeds. As discussed earlier, however, liquidity risk is particularly relevant precisely when the investor is credit- or liquidity-constrained and does not have easy access to outside credit. Further, in contrast with (2.4), the above objective function does not incorporate the fact that the *timing* of intertemporal proceeds from liquidation often matters to liquidity-constrained investors such as insurance firms faced with large liability claims, mutual or hedge funds faced with rapid redemptions by investors, and financially distressed firms who must meet contractual debt obligations. Finally, the objective function above precludes the possibility of precautionary motives for the investor to smooth liquidation proceeds over time that are potentially important, especially in highly illiquid markets with large costs arising from the price impact of trades.

ASSUMPTION 4. *In the following, the symbol $u(.)$ generically refers to the utility functions u_1 and u_2 above. The utility function $u(.,t)$ is increasing, strictly concave, and continuously differentiable on $(0,\infty)$ for all $t \in [0,T]$. It satisfies $\lim_{c\uparrow\infty} u(c,t) = \infty$ for all $t \in [0,T]$. Moreover, it satisfies*

$$\lim_{c\downarrow 0} u_c(c,t) = \infty \text{ and } \lim_{c\uparrow\infty} u_c(c,t) = 0, \tag{2.6}$$

and there exist constants $\delta \in (0,1)$ and $\gamma \in (0,\infty)$ such that

$$\delta u_c(c,t) \geq u_c(\gamma c,t) \quad \forall (c,t) \in (0,\infty)\times[0,T]. \tag{2.7}$$

Finally, $u(c,.)$ is continuous and decreasing on $[0,T]$ for all $c \geq 0$. The above conditions ensure that the function $u_c(.,t)$ has a continuous and strictly decreasing inverse $f(.,t)$ mapping $(0,\infty)$ onto itself which also satisfies the property

$$\forall\delta \in (0,\infty), \exists\,\gamma \in (0,\infty)$$

such that

$$f(\delta y,t) \leq \gamma f(y,t) \quad \forall (y,t) \in (0,\infty)\times[0,T]. \tag{2.8}$$

Remark 2. In section 6, we analyze the scenario in which the investor could incur costs from any remaining wealth in the asset at date T. In this case, the function $u_2(.,T)$ could be a general, concave, possibly *nonmonotonic* function taking both positive and negative values.

We define the function

$$h(W,c,t,\omega)=W\overline{\overline{\mu}}(W,c,t,\omega) \text{ for } (W,c)\in R_+\times R_+. \tag{2.9}$$

Assumption 5. *We assume that $h(.)$ is concave and upper semicontinuous in (W,c) for each $(t,\omega)$and uniformly bounded above. We also assume that $h(W,.,t,\omega)$ is monotonically decreasing for each (W,t,ω) and $h(0,0,t,\omega)=0$ so that there is no feedback on the asset price process if the large investor's asset holdings and liquidation rate are zero.*

From (2.2), we see that Assumption 5 implies that the expected increase in the wealth process $W(.)$ of the large investor decreases with the liquidation rate. The concavity of the function h implies that liquidating at a higher rate *relative* to his holdings has a disproportionately greater impact on the investor's wealth process. For mathematical reasons that will become clear later, we extend the definition of $h(.)$ to *negative* values of W. More precisely, we define $h(W',c,t,\omega)$ for $W'<0$ to be *any* nonpositive, concave, upper semicontinuous extension of $h(W,c,t,\omega)$ for $W>0$ (such an extension always exists by recalling that $h(0,0,t,\omega)=0$). Therefore,

$$\begin{gathered} h(.,.,t,\omega): R\times R_+\to R \text{ is concave and upper semicontinuous with} \\ h(W,c,t,\omega)\le 0 \text{ for } W\le 0. \end{gathered} \tag{2.10}$$

For convenience of notation, we also denote this extension by $h(.)$.

We now introduce some additional notation that we will need subsequently and some technical assumptions. For $(\mu_1,\nu)\in R^2$, define

$$\overrightarrow{h}(\mu_1,\nu,t,\omega)=\sup_{(W,c)\in R^\times R_+} h(W,c,t,\omega)-\mu_1 W-\nu c. \tag{2.11}$$

Let

$$M_t(\omega)=\left\{(\mu_1,\nu)\in R^2:\overrightarrow{h}(\mu_1,\nu,t,\omega)<\infty\right\}. \tag{2.12}$$

$M_t(\omega)$ is the *effective domain* of $\overrightarrow{h}$. It is well known that $\overrightarrow{h}$ is a nonnegative, convex, lower semicontinuous function and $M_t(\omega)$ is a nonempty, closed convex set with $0\in M_t(\omega)$ for all (t,ω) (see, e.g., Ekeland and Temam [9]). We assume that $h(.,.,t,\omega)$ is such that the sets $M_t(\omega)$ are *uniformly bounded* in the first coordinate and *uniformly bounded below* in the second coordinate. A sufficient, but not necessary, condition for this to hold is that $h(W,c,t,\omega)$ be uniformly Lipschitz in (W,c) (see, for example, El Karoui, Peng and Quenez [11]). Denote by M the family of two-dimensional progressively measurable processes $(\mu_1(t,\omega),\nu(t,\omega))$ such that $\chi(t,\omega)\in M_t(\omega)$ for all $(t,\omega)\in[0,T]\times\Omega$, $\nu(.)$ is *uniformly bounded*, and

$$E\left[\int_0^T \|\chi(t,\omega)\|^2\,dt\right]<\infty. \tag{2.13}$$

For later use, we define the sets

$$N_t(\omega)=R\times M_t(\omega) \tag{2.14}$$

and denote by N the family of three-dimensional progressively measurable processes $(\mu_0(t,\omega),\mu_1(t,\omega),\nu(t,\omega))$ such that $(\mu_0(t,\omega),\mu_1(t,\omega),\nu(t,\omega))\in N_t(\omega)$ for all $(t,\omega)\in$

$[0,T] \times \Omega$, $\nu(.)$ is uniformly bounded, and $E[\int_0^T [\|\mu_0(t,\omega)\|^2 + \|\mu_1(t,\omega)\|^2]\, dt] < \infty$. For subsequent notational convenience, we define the function $\overrightarrow{g}$ by

$$\overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \nu(t,\omega), t, \omega) = \overrightarrow{h}(\mu_1(t,\omega), \nu(t,\omega), t, \omega). \tag{2.15}$$

Assumption 6. *The function $\overrightarrow{h}$ is continuous and bounded on its effective domain.*

It follows from Theorem 12.2 in Rockafellar [28] that

$$h(W, c, t, \omega) = \inf_{((\mu_1,\nu)\in M_t(\omega))} \left[\overrightarrow{h}(\mu_1, \nu, t, \omega) + W\mu_1 + c\nu\right]. \tag{2.16}$$

In section 5, we present an example of a feedback function h that satisfies all the assumptions above, and we explicitly solve the optimal liquidation problem for an investor with logarithmic preferences. We now state an approximation lemma whose proof we omit for the sake of brevity since it follows using standard arguments.

Lemma 2.1. *For any $(W, c) \in \Theta$, there exists a sequence $\{\chi^n() \equiv (\mu_1(), \nu^n()) \in M\}$ and an F_t-progressively measurable process$\nu^*(.)$ such that a.e.*

$$\begin{aligned} &h(W(t), c(t), t) = \lim\nolimits_{n\to\infty} \left[\overrightarrow{h}(\chi^n(t), t) + W(t)\mu_1(t) + c(t)\nu^n(t)\right], \\ &\nu^*(t,\omega) = \lim\nolimits_{n\to\infty} \nu^n(t,\omega), \end{aligned} \tag{2.17}$$

and

$$\int_0^T |\nu^*(t,\omega)c(t,\omega)|\, dt < \infty \quad a.s.$$

3. Feasible liquidation policies. We now establish the duality between the primal problem ((2.4) and (2.5)) of maximizing the large investor's expected utility with a *dual problem* for which it is easier to verify optimal policies and to show their existence under broad assumptions. The space of dual processes $\chi(.)$ is

$$\{\chi(t) \equiv (\mu_0(t), \mu_1(t), \nu(t)), \chi \in N\},$$

where N is defined in the paragraph following (2.14).

Remark 3. The space of dual processes is *unbounded*, unlike in Cuoco and Cvitanic [5]. This makes our framework and analysis significantly different. Mathematically, this feature complicates the demonstration of verification and existence results of solutions to the dual control problem.

For an arbitrary process $\chi \in N$, define the exponential local martingale

$$\xi_\chi(t) = \exp\left(\int_0^t \kappa_\chi(s)\, dB(s) - \frac{1}{2}\int_0^t |\kappa_\chi(s)|^2\, ds\right) \tag{3.1}$$

and the discount factor

$$\beta_\chi(t) = \exp\left(-\int_0^t \mu_0(s)\, ds\right), \tag{3.2}$$

where

$$\kappa_\chi(t) = -\sigma(t)^{-1}(\overline{\mu}(t) + \mu_1(t) - \mu_0(t)), \tag{3.3}$$

and let

$$\pi_\chi(t) = \beta_\chi(t)\xi_\chi(t). \tag{3.4a}$$

Note. Although $\pi_\chi(.)$ does not depend on $\nu(.)$, we retain the notation above to simplify subsequent exposition. Throughout the paper, the notation $\pi_\chi(.)$ should be taken to represent $\pi_{(\mu_0(.),\mu_1(.))}(.)$ to indicate that it depends only on $(\mu_0(.), \mu_1(.))$. From (3.4a), it follows that

$$\pi_\chi(t) = 1 - \int_0^t \pi_\chi(s)\mu_0(s)\,ds - \int_0^t \pi_\chi(s)\sigma(s)^{-1}(\overline{\mu}(s) + \mu_1(s) - \mu_0(s))\,dB_s. \tag{3.4b}$$

The following proposition establishes the fundamental relationship between *feasible* liquidation policies and the space of dual processes.

Proposition 3.1. *If $(W(.), c(.))$ is a feasible liquidation policy and W is the terminal wealth, then*

$$E\left[\int_0^T \pi_\chi(t)(1-\nu(t))c(t)\,dt + \pi_\chi(T)W\right] \leq W_0 + E\left[\int_0^T \pi_\chi(t)\overrightarrow{g}(\chi(t),t)\,dt\right] \tag{3.5}$$

for all $\chi \in N$ with $\nu(.) \leq 1$.

Proof. The proof is in the appendix. □

Equation (3.5) is a *necessary condition* for feasibility of the policy $(c(.), W(.))$ and therefore represents a constraint on the arguments of the large investor's optimization problem (2.4).

4. Verification and existence of optimal liquidation policies.

In this section, we first prove a verification theorem for the optimal liquidation policy for the large investor in terms of the solution to a dual optimization problem. We then prove the existence of optimal liquidation policies for the large investor under fairly broad assumptions by proving the existence of a solution to the dual problem satisfying the conditions of the verification theorem. We thereby provide broad sufficient conditions for the existence of optimal liquidation policies.

Suppose the large investor's optimal liquidation/terminal wealth policy exists and is denoted by (c^*, W^*). By the result of Proposition 3.1, there should exist a Lagrange multiplier $\lambda^* > 0$ such that $(c^*, W^*, \lambda^*, \chi^*)$ is a saddle point of the map

$$\begin{aligned}L(c, W, \lambda, \chi) = U(c, W) - \lambda \\ \times \left(E\left[\int_0^T \pi_\chi(t)((1-\nu(t))c(t) - \overrightarrow{g}(\chi(t),t))\,dt + \pi_\chi(T)W\right] - W_0\right),\end{aligned} \tag{4.1}$$

where we maximize with respect to c, W and minimize with respect to (λ, χ). Let

$$\begin{aligned}\overline{u_1}(y,t) &= \sup_{c\geq 0}[u_1(c,t) - yc] = u_1(f_1(y,t),t) - yf_1(y,t),\\ \overline{u_2}(y,t) &= \sup_{W\geq 0}[u_2(W,T) - yW] = u_2(f_2(y,T),T) - yf_2(y,T).\end{aligned} \tag{4.2}$$

We have the following well-known lemma (see Karatzas et al. [17, p. 707]).

Lemma 4.1. *The function $\overline{u}(.,t) : (0,\infty) \to R$ is strictly decreasing and strictly convex for all $t \in [0,T]$, with $\frac{\partial}{\partial y}\overline{u}(y,t) = -f(y,t)$. Moreover, $\overline{u}(0+,t) = u(\infty,t)$ and $\overline{u}(\infty,t) = u(0+,t)$, where the symbol u refers generically to the functions u_1 and u_2.*

If we maximize (4.1) with respect to c, we obtain

$$\begin{aligned}(4.3)\quad & J(\lambda,\chi)\\ &= E\left[\int_0^T \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t),t)\,dt + \lambda\int_0^T \pi_\chi(t)\overrightarrow{g}(\chi(t),t)\,dt + \overline{u_2}(\lambda\pi_\chi(T),T)\right] + \lambda W_0.\end{aligned}$$

We see that the functional above is well defined for all $\chi \in N$ if and only if

$$E\left[\int_0^T \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t),t)^- dt\right] < \infty \quad \text{and} \quad E\left[\overline{u_2}(\lambda\pi_\chi(T),T)^-\right] < \infty \text{ for all } \chi \in N.$$

By Lemma 4.1 and the continuity of $\overline{u_1}(0,t)$ for $t \in [0,T]$, which follows from Assumption 4, $\overline{u_1}(.,t)$ is uniformly bounded below for all $t \in [0,T]$ and $\overline{u_2}(.,T)$ is bounded below. Therefore, both the inequalities above hold and $J : (0,\infty) \times N \to R \cup \{\infty\}$ is well defined. We note that $J(\lambda,\chi) = \infty$ if $1-\nu(t) \leq 0$ on a set of positive measure in $[0,T]\times\Omega$ (since $\overline{u_1}(y,t) = \infty$ for $y \leq 0$). Therefore, it suffices to consider the problem

$$(4.4)\qquad \inf_{(\lambda,\chi)\in(0,\infty)\times N'} J(\lambda,\chi),$$

where

$$N' = \{\chi(.) \in N : \nu(t) \leq 1 \text{ on } [0,T]\times\Omega\}.$$

We can then prove the following *verification theorem* for the optimal liquidation policy.

PROPOSITION 4.1. *If* $(\lambda^*,\chi^*) \in (0,\infty)\times N'$ *solves problem* (4.4) *and*

$$(4.5)\qquad \begin{aligned} &E\left[\int_0^T \pi_{\chi^*}(t) f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt + \pi_{\chi^*}(T) f_2(\lambda^*\pi_{\chi^*}(T),T)\right] < \infty,\\ &E\left[\int_0^T \pi_{\chi^*}(t)\,dt\right] < \infty, E[\pi_{\chi^*}(T)] < \infty,\end{aligned}$$

then the policy

$$(4.6)\qquad \begin{aligned} c_{\chi^*}(t) &= f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t),\\ W_{\chi^*} &= f_2(\lambda^*\pi_{\chi^*}(T),T)\end{aligned}$$

is optimal, and the optimal wealth process is given by

$$\begin{aligned}(4.7)\quad & W_{\chi^*}(t)\\ &= \pi_{\chi^*}(t)^{-1} E\left[\int_t^T \pi_{\chi^*}(s)((1-\nu^*(s))c_{\chi^*}(s) - \overrightarrow{g}(\chi^*(s),s))\,ds + \pi_{\chi^*}(T)W_{\chi^*}\,|F_t\right].\end{aligned}$$

Proof. The proof is in the appendix. □

We now use the previous results to show that optimal liquidation policies for the large investor exist under fairly broad assumptions on the large investor's preferences and the feedback function $h(.)$. Before proceeding, we need to introduce some notation. Let $(\mu_0(.),\mu_1(.))$ be a fixed, square integrable, progressively measurable process such that the set of processes $\nu(.)$ (denoted by$D_{(\mu_0(.),\mu_1(.))}$), for which

$(\mu_0(.), \mu_1(.), \nu(.)) \in N'$, is nonempty. The (t, ω)-section of the set $D_{(\mu_0(.),\mu_1(.))}$ is denoted by $D_{(\mu_0(.),\mu_1(.))}(t, \omega)$, i.e.,

$$D_{(\mu_0(.),\mu_1(.))}(t, \omega) = \left\{\nu(t, \omega) : \nu(.) \in D_{(\mu_0(.),\mu_1(.))}\right\}.$$

By Assumption 5, the set $D_{(\mu_0(.),\mu_1(.))}(t, \omega)$ is closed, convex, uniformly bounded below, and uniformly bounded above (by the definition of the set N').

Lemma 4.2. *For a fixed, square integrable, progressively measurable process $(\mu_0(.), \mu_1(.))$ and a fixed $\lambda \in (0, \infty)$, such that $D_{(\mu_0(.),\mu_1(.))}$ is nonempty, if*

$$\inf_{\nu(.) \in D_{(\mu_0(.),\mu_1(.))}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.))$$

exists, then it is uniquely attained. Moreover, $\inf_{\nu(.) \in D_{(\mu_0(.),\mu_1(.))}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.))$ is a convex functional of $\pi_\chi \equiv \pi_{(\mu_0(.),\mu_1(.))}$.

Proof. The proof is in the appendix. □

Assumption 7. *$D_{(\mu_0(.),\mu_1(.))}(t, \omega) = [-\Gamma_{(t,\omega)}, 1]$ for all square integrable, progressively measurable $(\mu_0(.), \mu_1(.))$ for which $D_{(\mu_0(.),\mu_1(.))}$ is nonempty. Here $0 \leq \Gamma_{(t,\omega)} < \infty$ is a fixed constant for each (t, ω).*

It is easy to see that a sufficient, but not necessary, condition for this to hold is that $h(.)$ be separable, i.e., that it have the form

$$h(\theta, c, t, \omega) = g_1(\theta, t, \omega) + g_2(c, t, \omega),$$

where $g_1(., t, \omega)$ and $g_2(., t, \omega)$ are concave, upper semicontinuous, and uniformly bounded above.

Proposition 4.2. *In the following, the symbol u generically refers to the functions u_1 and u_2. An optimal liquidation policy for the large investor exists if, in addition to Assumptions 1–7,*

(a) *$u(\infty, t) = \infty$ for all $t \in [0, T]$ and $0 \leq u(c, t) \leq k(1 + c^{1-b})$ on $(0, \infty) \times [0, T]$ for some $k \geq 0, 0 < b < 1$;*

(b) *$c \to cu'(c, t)$ is increasing on $(0, \infty)$ for each $t \in [0, T]$ so that $x \to \overline{u}(\exp(x), t)$ is convex on R;*

(c) *for all, there exists $\chi \in N'$ such that $J(\lambda, \chi) < \infty$;*

(d) *the functions $\overline{u}(., t) : R_+ \to R$ and $\overrightarrow{g}(., t, \omega) : N_t(\omega) \to R$ are strictly convex and continuously differentiable on the interior of their domains for each (t, ω); and*

(e) *$cu_c(c, t) \leq a + (1 - b)u(c, t)$ on $(0, \infty) \times [0, T]$ for some $a, b \geq 0$.*

Note. Assumption (c) above actually follows from assumption (a). (see Remark 11.9 in Karatzas et al. [17]).

Proof. The proof is in the appendix. □

It is easily checked that the assumptions of the above proposition hold if $u(c, t) = \beta(t)c^\gamma, 0 < \gamma < 1$, for some nonnegative uniformly bounded measurable function $\beta(.)$, where $u(.)$ refers generically to the utility functions $u_1(.)$ and $u_2(.)$. Therefore, the result of Proposition 4.2 implies that an optimal policy exists for a large investor with power utility functions.

Remark 4. All our results hold if the investor derives no utility from wealth at the terminal date, that is, $u_2 \equiv 0$. Under the hypotheses of Proposition 4.2, the investor's optimal liquidation policy is given by (4.6) with the terminal wealth W_{χ^*} equal to zero. Hence, the investor liquidates his entire position in the asset prior to date T. In section 6, we analyze the scenario in which the terminal payoff function u_2 could be nonmonotonic and take on positive and negative values.

5. Optimal liquidation with logarithmic utility and the liquidity discount. We now use the results of the previous section to explicitly derive the optimal liquidation policy for the large investor in the situation where his utility functions are logarithmic. Specifically,

$$\begin{aligned} u_1(c,t) &= \exp(-\rho t)\log(c), \\ u_2(W,T) &= \exp(-\rho T)\log(W). \end{aligned} \tag{5.1}$$

From (4.2), we see that

$$\begin{aligned} \overline{u_1}(y,t) &= -\exp(-\rho t)(1+\rho t+\log(y)), \\ \overline{u_2}(y,T) &= -\exp(-\rho T)(1+\rho T+\log(y)). \end{aligned} \tag{5.2}$$

We also assume that

$$h(W,c,t,\omega) = -\alpha(c,t,\omega), \tag{5.3}$$

where $\alpha(.,t,\omega)$ is a differentiable (with a bounded derivative), nonnegative, convex function with $\alpha(0)=0$ so that $h(.,.,t,\omega)$ is concave, uniformly Lipschitz, and nonpositive. Therefore,

$$\overrightarrow{g}(\chi(t),t,\omega) = \overrightarrow{h}(\chi(t),t,\omega) = \sup_{c(t)\geq 0}[-\alpha(c(t),t,\omega)-\nu(t)c(t)].$$

We assume that the function $\alpha(.)$ is such that $\overrightarrow{g}(\chi(t),t)$ is zero on its effective domain. It is easily seen that an example of such a function is the linear function $\alpha(c,t,\omega) = \Gamma(t,\omega)c$, where $\Gamma(t,\omega)>0$.

Equation (5.3) implies that $\mu_1(t,\omega)\equiv 0$. We also note that $\nu(t,\omega)\in[-\Gamma(t,\omega),\infty)$, where $\Gamma(t,\omega)>0$. Therefore, the effective domain is given by

$$N_t(\omega) = R\times\{0\}\times[-\Gamma(t,\omega),\infty).$$

The dual problem (4.4) can now be expressed as

$$\begin{aligned} &\inf_{(\lambda,\chi)\in(0,\infty)\times N'} E\Bigg[-\int_0^T \exp(-\rho t)(1+\rho t+\log(1-\nu(t))+\log(\lambda\pi_\chi(t)))\,dt \\ &\qquad\qquad - \exp(-\rho T)(1+\rho T+\log(\lambda\pi_\chi(T)))\Bigg] + \lambda W_0 \\ &\quad = (T(1-\rho)-1)\exp(-\rho T) - 2\frac{1-\exp(-\rho t)}{\rho} \\ &\qquad + \inf_{\lambda>0}\left[\lambda W_0 - \frac{1-(1-\rho)\exp(-\rho T)}{\rho}\log(\lambda)\right] \\ &\qquad + \inf_{\nu} E\left[-\int_0^T \exp(-\rho t)\log(1-\nu(t))\,dt\right] \\ &\qquad + \inf_{\mu_0,\mu_1} E\Bigg[\int_0^T \exp(-\rho t)\left(\int_0^t\left(\mu_0(s)+\frac{1}{2}\left|\kappa_\chi(s)^2\right|\right)ds\right)dt \\ &\qquad\qquad + \exp(-\rho T)\left(\int_0^T\left(\mu_0(s)+\frac{1}{2}\left|\kappa_\chi(s)\right|^2\right)ds\right)\Bigg]. \end{aligned}$$

From the above, we obtain

$$\lambda^* = \frac{1-(1-\rho)e^{-\rho T}}{\rho W_0}, \tag{5.4}$$

$$\nu^*(t,\omega) = -\Gamma(t,\omega), \tag{5.5}$$

$$\mu_1^*(t,\omega) = 0, \tag{5.6}$$

$$\begin{aligned} \mu_0^*(t,\omega) &= \arg \min_{\mu_0(t)\in R} \left(\mu_0(t) + \frac{1}{2}\sigma(t,\omega)^{-2}(\overline{\mu}(t,\omega) - \mu_0(t,\omega))^2\right) \\ &= \overline{\mu}(t,\omega) - \sigma(t,\omega)^2. \end{aligned} \tag{5.7}$$

Since $\sigma(.)$ is uniformly bounded, we see that $\mu_0^*(.)$ is uniformly bounded, and consequently we easily see that condition (4.5) of Proposition 4.1 is satisfied. By Proposition 4.1, (4.7), and (5.4)–(5.7), the optimal liquidation rate and wealth processes are given by

$$\begin{aligned} c_{\chi^*}(t) &= W_0 \frac{\rho e^{-\rho T}}{(1+\Gamma(t,\omega))(1-(1-\rho)e^{-\rho T})} \pi_{\chi^*}(t)^{-1}, \\ W_{\chi^*}(t) &= W_0 \pi_{\chi^*}(t)^{-1} \frac{\rho e^{-\rho T}(T+1-t)}{(1-(1-\rho)e^{-\rho T})}. \end{aligned} \tag{5.8}$$

In the above, the process $\pi_{\chi^*}(.)$ may be derived from (3.1)–(3.4). From (5.8), we see that

$$\frac{c_{\chi^*}(t)}{W_{\chi^*}(t)} = \frac{T+1-t}{1+\Gamma(t,\omega)}. \tag{5.9}$$

Since $\Gamma(t,\omega) > 0$ and is uniformly bounded, (5.9) implies that $\frac{c_{\chi^*}(t)}{W_{\chi^*}(t)}$ is also uniformly bounded. Hence, the feedback on the stock price drift $\overline{\overline{\mu}}(W_{\chi^*}, c_{\chi^*}, t, \omega) = \frac{h(W_{\chi^*}, c_{\chi^*}, t,\omega)}{W_{\chi^*}} = -\frac{\Gamma(t,\omega)c_{\chi^*}}{W_{\chi^*}}$ is also uniformly bounded. If $\Gamma(t,\omega)$ is deterministic, then (5.9) implies that the investor liquidates a deterministic proportion of his holdings at each date; that is, the proportion liquidated does not depend on the stock price process. Moreover, if $\Gamma(t,\omega)$ is a constant, then (5.9) implies that the investor liquidates a decreasing proportion of his holdings in the risky asset over time.

We can use (5.4)–(5.8) to derive the optimal value function of the large investor, which is

$$\begin{aligned} U &= E\left[\int_0^T e^{-\rho t}\log(c_{\chi^*}(t))\,dt + e^{-\rho T}\log(W_{\chi^*}(T))\right] \\ &= E\left[\int_0^T e^{-\rho t}\left\{\log(N(0)S(0)) + \int_0^t \left(\overline{\mu}(s) - \frac{1}{2}\sigma(s)^2\right) ds \right.\right. \\ &\qquad \left. + \log\left(\frac{\rho e^{-\rho T}}{(1+\Gamma(t,\omega))(1-(1-\rho)e^{-\rho T})}\right)\right\} dt \\ &\qquad + e^{-\rho T}\left\{\log(S(0)) + \log(N(0)) + \int_0^T \left(\overline{\mu}(s) - \frac{1}{2}\sigma(s)^2\right) ds \right. \\ &\qquad \left.\left. + \log\left(\frac{\rho e^{-\rho T}}{(1-(1-\rho)e^{-\rho T})}\right)\right\}\right]. \end{aligned} \tag{5.10}$$

The liquidity discount. Subramanian and Jarrow [30] characterize the liquidity risk of a large investor's position in a risky asset in terms of the *liquidity discount.* The liquidity discount is defined in terms of the value functions of the large investor's optimal liquidation problem, and an "equivalent" hypothetical price-taker's optimal liquidation problem. The "equivalent" hypothetical price-taker has the same preferences and holdings in the risky asset, but his trades have no effect on the stock price. Specifically, the liquidity discount is defined as follows:

$$\text{liquidity discount} \;=\; S(0) - S(0)\prime, \tag{5.11}$$

where

$$U(S(0), N(0)) = U_{price\text{-}taker}(S(0)\prime, N(0)). \tag{5.12}$$

As discussed in Subramanian and Jarrow [30], the liquidity discount is the difference between the actual asset price and the asset price that would provide an equivalent price-taker with the same optimal expected utility. In (5.12), we explicitly indicate the dependence of the value functions on the initial stock price and holdings in the risky asset. The equivalent price-taker's value function is given by (5.10) with $\Gamma(t,\omega) \equiv 0$. Therefore, from (5.10) and (5.12), $S(0)\prime$ solves the equation

$$\begin{aligned}&\frac{1-(1-\rho)e^{-\rho T}}{\rho}\log(S(0)) + E\left[\int_0^T e^{-\rho t}\log\left(\frac{\rho e^{-\rho T}}{(1+\Gamma(t,\omega))(1-(1-\rho)e^{-\rho T})}\right)dt\right]\\ &\quad = \frac{1-(1-\rho)e^{-\rho T}}{\rho}\log(S(0)') + E\left[\int_0^T e^{-\rho t}\log\left(\frac{\rho e^{-\rho T}}{(1-(1-\rho)e^{-\rho T})}\right)dt\right].\end{aligned} \tag{5.13}$$

Hence,

$$S(0) = S(0)'\exp\left[\frac{\rho}{1-(1-\rho)e^{-\rho T}}E\left[\int_0^T e^{-\rho t}\log(1+\Gamma(t,\omega))\,dt\right]\right]. \tag{5.14}$$

Finally, from the definition (5.11), we have

$$\begin{aligned}&\text{liquidity discount}\\ &\quad = S(0)\left(1-\exp\left[-\frac{\rho}{1-(1-\rho)e^{-\rho T}}E\left[\int_0^T e^{-\rho t}\log(1+\Gamma(t,\omega))\,dt\right]\right]\right).\end{aligned} \tag{5.15}$$

6. Extensions.

In this section, we analyze two extensions of the model developed in the previous sections.[4]

6.1. Optimal liquidation with costs of residual terminal wealth.

Thus far, our analysis has focused on the case where the investor derives utility $u_2(W,T)$ from wealth W at date T, where $u_2(.,T)$ is increasing and concave (all the results presented thus far hold if u_2 is identically equal to zero—see Remark 4). Given that the investor wishes to liquidate his asset holdings, it is interesting to consider the scenario in which the investor actually bears *costs* from retaining nonzero wealth in the asset at the terminal date T. The interaction of these costs with the positive

[4] We thank an anonymous referee for suggesting these extensions.

(indirect) utility that the investor gains from the fact that residual wealth could be used to finance future consumption could lead to *nonmonotonicity* in the function $u_2(.,T)$. An example of such a function is

$$u_2(W,T) = A(W) - B(W), \tag{6.1}$$

where $A(.)$ is an increasing concave function and $B(.)$ is a nonnegative, increasing, and convex function.

We now modify our previous analysis by considering the class of functions $u_2(.,T)$ that are concave and possibly nonmonotonic and could take positive and negative values. It follows from the concavity of $u_2(.,T)$ that it is *either* monotonically decreasing *or* strongly unimodal (hump-shaped) and has a single local (therefore, global) maximum in $(0,\infty)$. A general hump-shaped function $u_2(.,T)$ (see the functional form (6.1)) has the economically intuitive features that, for low wealth levels, the benefits outweigh the costs so that the function increases, while for wealth levels above a threshold, the costs outweigh the benefits so that the function decreases. We make the following additional technical assumptions on $u_2(.,T)$ that modify Assumption 4.

ASSUMPTION 4. *The function $u_2(.,T)$ is strictly concave and continuously differentiable on $(0,\infty)$. If it is nonmonotonic, it satisfies*

$$\lim_{c\downarrow 0}(u_2)_c(c,T) = \infty. \tag{6.2}$$

Let $\hat{c}$ be the point at which $u_2(c,T)$ attains its unique local (therefore, global) maximum for $c \in [0,\infty)$. There exist constants $\delta \in (0,1)$ and $\gamma \in (0,\infty)$ such that

$$\delta(u_2)_c(c,T) \geq (u_2)_c(\gamma c,T) \quad \forall c \in (0,\hat{c}). \tag{6.3}$$

The above conditions ensure that the function $(u_2)_c(.,T)$ has a continuous and strictly decreasing inverse $f_2(.,T)$, which also satisfies the property

$$\forall \delta \in (0,\infty), \exists \gamma \in (0,\infty) \tag{6.4}$$

such that

$$f_2(\delta y,T) \leq \gamma f_2(y,T) \quad \forall\ y \in (0,\infty).$$

The function u_1 satisfies Assumption 4 *and Assumptions* 1–3, 5, *and* 6 *remain unaltered.*

Proposition 3.1, which characterizes the feasible liquidation policies of the investor, does not depend on the function u_2 and, therefore, continues to hold. Since $u_2(.,T)$ is concave, (4.1) defines the Lagrangian $L(c,W,\lambda,\chi)$ associated with the investor's optimal liquidation problem. In particular, if (c^*,W^*) is the investor's optimal liquidation/terminal wealth policy, there exists a Lagrange multiplier $\lambda^* > 0$ such that $(c^*,W^*,\lambda^*,\chi^*)$ is a saddle point of the map $L(c,W,\lambda,\chi)$ defined in (4.1).

We define the conjugate function $\overline{u_2}(y,T)$ as in (4.2). The following lemma describes the properties of this function that we will need in what follows.

LEMMA 6.1. *We have*

$$\overline{u_2}(y,T) = u_2(l_2(y,T),T) - yl_2(y,T) \text{ for } y \in [0,\infty), \tag{6.5}$$

where

$$l_2(y,T) = 0 \quad \forall y \in [0,\infty) \tag{6.6}$$

if $u_2(.,T)$ is monotonically decreasing, and

$$l_2(y,T) = (u_2')^{-1}(y,T) = f_2(y,T) \in [0,\hat{c}] \quad \forall y \in [0,\infty) \tag{6.7}$$

if $u_2(.,T)$ is nonmonotonic. In the above, $\hat{c}$ is as defined in Assumption 4.

Proof. The proof is in the appendix. □

To simplify the notation in the subsequent discussion, we combine the scenario in which $u_2(.,T)$ is strictly decreasing with the scenario in which it is nonmonotonic by setting $\hat{c} = 0$ in the former case. We can now proceed as in section 4 to consider the dual problem (4.4) and arrive at the following verification theorem for the optimal liquidation policy of the investor.

Proposition 6.1. *If $(\lambda^*, \chi^*) \in (0,\infty) \times N'$ solves problem* (4.4) *and*

$$\begin{aligned} &E\left[\int_0^T \pi_{\chi^*}(t) f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt + \pi_{\chi^*}(T) l_2(\lambda^*\pi_{\chi^*}(T),T)\right] < \infty,\\ &E\left[\int_0^T \pi_{\chi^*}(t)\,dt\right] < \infty, E[\pi_{\chi^*}(T)] < \infty, \end{aligned} \tag{6.8}$$

then the policy

$$\begin{aligned} c_{\chi^*}(t) &= f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t),\\ W_{\chi^*} &= l_2(\lambda^*\pi_{\chi^*}(T),T) \end{aligned} \tag{6.9}$$

is optimal, and the optimal wealth process is given by

$$\begin{aligned} &(6.10)\ W_{\chi^*}(t) =\\ &\pi_{\chi^*}(t)^{-1} E\left[\int_t^T \pi_{\chi^*}(s)((1-\nu^*(s))c_{\chi^*}(s) - \overrightarrow{g}(\chi^*(s),s))\,ds + \pi_{\chi^*}(T) W_{\chi^*}\,|F_t\right]. \end{aligned}$$

Proof. The proof is in the appendix. □

Lemma 6.1 and Proposition 6.1 ((6.6), (6.7), and (6.9)) lead to the following result.

Corollary 6.1. (a) *If the function $u_2(.,T)$ is decreasing, then the optimal terminal wealth is zero; that is, it is optimal for the investor to liquidate his entire holdings prior to the date T.*

(b) *If the function $u_2(.,T)$ is nonmonotonic, then the optimal terminal wealth $W_{\chi^*} \in [0,\hat{c}]$; that is, the terminal wealth lies in the increasing region of the function $u_2(.,T)$.*

The above results follow from the duality between the investor's primal problem (2.4) and the dual problem (4.4). For a given pair of dual control parameters (λ,χ), the liquidation rate process and the terminal wealth solve separate optimization problems. Proposition 4.1 (which depends on the martingale representation theorem for Brownian motion) shows that there exist an optimal liquidation policy that gives rise to the liquidation rate and terminal wealth that solve these two separate optimization problems. Because of the separability of the investor's problem of optimizing his terminal wealth level, it is suboptimal for the investor to retain nonzero terminal wealth in the region in which $u_2(.,T)$ is decreasing. It then follows that the optimal terminal wealth is zero when the function $u_2(.,T)$ is decreasing and lies in the increasing region of the function when it is nonmonotonic.

Corollary 6.1 immediately implies that the large investor's optimal liquidation policy is completely determined by the *increasing* part of the function $u_2(.,T)$. In other words, two concave terminal payoff functions with identical increasing parts are associated with the same optimal liquidation policies. This result is not transparent from an examination of the primal problem (2.4) but is immediately apparent from the dual problem (4.4). The duality approach, therefore, facilitates the characterization of optimal liquidation policies under more general terminal payoff functions.

Note. In the context of example (6.1), the increasing part of the function $u_2(.,T)$ is, in general, affected by *both* the "benefit" function $A(.)$ and the "cost" function $B(.)$. Hence, the investor's optimal liquidation policy is influenced by both these functions.

We now derive an existence result for optimal liquidation policies. We define the set $D_{(\mu_0(.),\mu_1(.))}$ as in the discussion following Proposition 4.1. An examination of the proof of Lemma 4.2 shows that it continues to hold.

PROPOSITION 6.2. *Suppose that conditions* (a), (b), (d), *and* (e) *of Proposition* 4.2 *hold for the functions* u_1 *and* $\overline{u_1}$. *An optimal liquidation policy for the large investor exists if the following conditions hold:*

(a) $0 \leq u_2(W,T) \leq k(1+W^{1-b})$ *for* $W \in [0,\hat{c}]$ *for some* $k \geq 0, 0 < b < 1$.

(b) $W \rightarrow W(u_2)'(W,T)$ *is increasing for* $W \in [0,\hat{c}]$ *so that* $x \rightarrow \overline{u_2}(\exp(x),T)$ *is convex on* R.

(c) *For all* $\lambda \in (0,\infty)$, *there exists* $\chi \in N'$ *such that* $J(\lambda,\chi) < \infty$.

(d) *The function* $\overline{u_2}(.,T)$ *is strictly convex and continuously differentiable on* R_+, *and the function* $\overrightarrow{g}(.,t,\omega) : N_t(\omega) \rightarrow R$ *is strictly convex and continuously differentiable on the interior of its domain for each* (t,ω).

(e) $W(u_2)_W(W,T) \leq a + (1-b)u_2(W,T)$ *for* $W \in [0,\hat{c}]$ *and some* $a, b \geq 0$.

Proof. The proof is in the appendix. □

6.2. Infinite liquidation horizon. In this section, we analyze the scenario in which the investor's liquidation horizon is $[0,\infty)$. Our presentation is brief because it closely follows the discussion in sections 3 and 4. A feasible liquidation strategy is a nonnegative process $(W(.),c(.))$ satisfying (2.3) with the time horizon $T=\infty$. A liquidation rate process $c(.)$ is admissible if it satisfies the first condition in (2.5) with $T=\infty$. The investor's objective is to choose an admissible liquidation strategy to maximize

$$U(c) = E\left[\int_0^\infty u_1(c(t),t))\,dt\right]. \tag{6.11}$$

The investor, therefore, optimally liquidates his holdings in the asset until they fall to zero.

We impose Assumptions 1, 4, 5, and 6 as in section 2 with $T=\infty$ and define the space of dual processes as in section 3. The following proposition is the analogue of Proposition 3.1.

PROPOSITION 6.3. *If* $(W(.),c(.))$ *is a feasible liquidation policy, then*

$$E\left[\int_0^\infty \pi_\chi(t)(1-\nu(t))c(t)\,dt\right] \leq W_0 + E\left[\int_0^\infty \pi_\chi(t)\overrightarrow{g}(\chi(t),t)\,dt\right] \tag{6.12}$$

for all $\chi \in N$ *with* $\nu(.) \leq 1$.

Proof. The proof proceeds exactly as the proof of Proposition 3.1 with $T=\infty$. □

The Lagrangian for the investor's optimal liquidation problem is

$$L(c,\lambda,\chi) = U(c) - \lambda\left(E\left[\int_0^\infty \pi_\chi(t)((1-\nu(t))c(t) - \overrightarrow{g}(\chi(t),t))\,dt\right] - W_0\right). \tag{6.13}$$

If an optimal liquidation policy $c^*(.)$ exists, then there exists a Lagrange multiplier $\lambda^* > 0$ and a dual process χ^* such that (c^*, λ^*, χ^*) is a saddle point of the above map where we maximize with respect to c and minimize with respect to (λ, χ). We define the conjugate function $\overline{u_1}$ as in (4.2). The investor's dual problem is defined by (4.4), where

$$J(\lambda, \chi) = E\left[\int_0^\infty \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t), t)\, dt + \lambda \int_0^\infty \pi_\chi(t) \overrightarrow{g}(\chi(t), t)\, dt\right] + \lambda W_0. \tag{6.14}$$

The following proposition is the analogue of the verification Proposition 4.1.

PROPOSITION 6.4. *If* $(\lambda^*, \chi^*) \in (0, \infty) \times N'$ *solves the problem* (4.4) *and*

$$E\left[\int_0^\infty \pi_{\chi^*}(t) f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t), t)\, dt < \infty, \quad E\left[\int_0^\infty \pi_{\chi^*}(t)\, dt\right]\right] < \infty, \tag{6.15}$$

then the policy

$$c_{\chi^*}(t) = f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t), t) \tag{6.16}$$

is optimal, and the optimal wealth process is given by

$$W_{\chi^*}(t) = \pi_{\chi^*}(t)^{-1} E\left[\int_t^\infty \pi_{\chi^*}(s)((1-\nu^*(s))c_{\chi^*}(s) - \overrightarrow{g}(\chi^*(s), s))\, ds \,|F_t\right]. \tag{6.17}$$

Proof. The proof proceeds exactly as the proof of Proposition 4.1 setting $T = \infty, u_2 \equiv 0, f_2 \equiv 0$, and $W^* = 0$. □

The analysis now proceeds along the lines of the analysis following Proposition 4.1. Lemma 4.2 holds in this setting, and we impose Assumption 7. The existence result Proposition 4.2 holds as stated in this setting with $T = \infty, u_2 \equiv 0, f_2 \equiv 0$, and $W^* = 0$.

7. Conclusions. We study the optimal liquidation problem for an investor with a large holding in a risky asset where there may be a nonlinear path dependent feedback on the asset price process due to his liquidation policy. Under broad and general assumptions on the large investor's preferences, and the parameters of the asset price process, we establish the duality between the investor's optimal liquidation problem and a dual optimization problem. We prove verification and existence results for optimal liquidation policies for the large investor. In particular, our results imply the existence of optimal policies if the investor has power utility functions. We explicitly derive the optimal policy when the large investor's utility functions are logarithmic. We use our results to characterize the *liquidity discount*, which is a measure of the *liquidity risk* of the investor's position in the risky asset. The derivation of conditions under which optimal liquidation policies for the large investor exist is important from an economic standpoint since it identifies a class of *viable* partial equilibrium models of large investor behavior. In future research, it would be interesting and important to construct a general equilibrium model of large investor behavior.

Appendix.

Proof of Proposition 3.1. Recall that, by definition (2.15), $\overrightarrow{g}(\chi(t), t) = \overrightarrow{h}(\mu_1(t), \nu(t), t)$. Using (2.2) and Itǒ's lemma,

$$\begin{aligned}\pi_\chi(t)W(t) &+ \int_0^t \pi_\chi(s)c(s)\, ds - \int_0^t \pi_\chi(s)[W(s)\sigma(s) + W(s)\kappa_\chi(s)]\, dB(s) \\ &= W_0 + \int_0^t \pi_\chi(s)(h(W(s), c(s), s) - W(s)\mu_1(s))\, ds.\end{aligned}$$

Since $\nu(.) \leq 1$ by hypothesis, $\int_0^T c(s)\,ds < \infty$ by the definition of the feasibility of the liquidation policy, and $\pi_\chi(.)$ has continuous paths a.s., we see that $\int_0^t \pi_\chi(s)\nu(s)c(s)\,ds$ exists a.s. for each $t \in [0,T]$. Subtracting the term $\int_0^t \pi_\chi(s)\nu(s)c(s)\,ds$ from both sides, we see that for all $\chi \in N$,

(A.1)

$$\begin{aligned}\pi_\chi(t)W(t) &+ \int_0^t \pi_\chi(s)(1-\nu(s))c(s)\,ds - \int_0^t \pi_\chi(s)[W(s)\sigma(s) + W(s)\kappa_\chi(s)]\,dB(s)\\ &= W_0 + \int_0^t \pi_\chi(s)(h(W(s),c(s),s) - W(s)\mu_1(s) - c(s)\nu(s))\,ds\\ &\leq W_0 + \int_0^t \pi_\chi(s)\overrightarrow{h}(\chi(s),s)\,ds\end{aligned}$$

from the definition (2.11) of $\overrightarrow{h}(.)$. For each positive integer n, let

$$\begin{aligned}\tau_n = T \wedge \inf\Bigg\{t \in [0,T] : \int_0^t &[|\pi_\chi(s)(W(s)\sigma(s) + W(s)\kappa_\chi(s))|^2\\ &+ \pi_\chi(s)(1-\nu(s))c(s)]\,ds \geq n\Bigg\}.\end{aligned}$$

Since the stochastic integral in (A.1) is a martingale on $[0,\tau_n]$, taking expectations gives

(A.2)

$$E[\pi_\chi(\tau_n)W(\tau_n)] + E\left[\int_0^{\tau_n} \pi_\chi(t)(1-\nu(t))c(t)\,dt\right] \leq W_0 + E\left[\int_0^{\tau_n} \pi_\chi(t)\overrightarrow{h}(\chi(t),t)\,dt\right].$$

Since $W(\tau_n) \geq 0$ a.s. (by the definition of feasibility of the liquidation strategy), $\tau_n \uparrow T$, $\overrightarrow{h}(.)$ is nonnegative and bounded on its effective domain, and $\nu(.) \leq 1$, we see from (A.2) and the monotone convergence theorem that

$$\lim_{n\to\infty} E\left[\int_0^{\tau_n} \pi_\chi(t)(1-\nu(t))c(t)\,dt\right] = E\left[\int_0^T \pi_\chi(t)(1-\nu(t))c(t)\,dt\right].$$

An application of the monotone convergence theorem and Fatou's lemma implies that

$$\begin{aligned}E&\left[\int_0^T \pi_\chi(t)(1-\nu(t))c(t)\,dt + \pi_\chi(T)W\right]\\ &\leq W_0 + E\left[\int_0^T \pi_\chi(t)\overrightarrow{h}(\chi(t),t)\,dt\right] \text{ for all } \chi \in N \text{ with } \nu(.) \leq 1.\end{aligned}$$

This completes the proof. □

Proof of Proposition 4.1. Let the liquidation process be defined by $c^* = c_{\chi^*}$ and the terminal wealth level by $W^* = W_{\chi^*}$.

Step 1. *Optimality of* $(c^*(.), W^*)$ *given feasibility.* We shall first show that the policy $(c^*(.), W^*)$ is optimal *provided it is feasible.* Since $u_1(1,t) - y \leq \sup_{c>0}[u_1(c,t) - yc] = u_1(f_1(y,t),t) - yf_1(y,t)$, we have

$$E\left[\int_0^T u_1(c^*(t),t)^- dt\right] \leq \int_0^T u_1(1,t)^- dt + \lambda^* E\left[\int_0^T (1-\nu^*(t))\pi_{\chi^*}(t)\,dt\right] < \infty.$$

The last inequality follows from the fact that $E[\int_0^T (1-\nu^*(t))\pi_{\chi^*}(t)\,dt] < \infty$ as $\nu^*(.)$ is uniformly bounded and $E[\int_0^T \pi_{\chi^*}(t)\,dt] < \infty$ by assumption. Therefore, the process $c^*(.)$(and similarly the random variable W^*) satisfies (2.5). We shall now prove that $U(c^*, W^*) \geq U(c, W)$ for all *feasible* liquidation and terminal wealth processes $(c(.), W)$. By (2.7), we see that (4.5) implies that

$$\text{(A.3)}\quad E\left[\int_0^T \pi_{\chi^*}(t) f_1(\lambda(1-\nu^*(t))\pi_{\chi^*}(t), t)\,dt\right] + E\left[\pi_{\chi^*}(T) f_2(\lambda\pi_{\chi^*}(T), T)\right] < \infty$$

holds for all $\lambda \in (0, \infty)$. Since λ^* is optimal by assumption, we have

$$\begin{aligned}
0 &= \lim_{\varepsilon\to 0}\frac{J(\lambda^*+\varepsilon, \chi^*) - J(\lambda^*, \chi^*)}{\varepsilon} \\
&= E\left[\int_0^T \lim_{\varepsilon\to 0}\frac{\overline{u}_1((\lambda^*+\varepsilon)(1-\nu^*(t))\pi_{\chi^*}(t), t) - \overline{u_1}(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t), t)}{\varepsilon}\right. \\
&\qquad \left. + \int_0^T \pi_{\chi^*}(t)\overrightarrow{g}(\chi^*(t), t)\,dt + \frac{\overline{u}_2((\lambda^*+\varepsilon)\pi_{\chi^*}(T), T) - \overline{u_2}(\lambda^*\pi_{\chi^*}(T), T)}{\varepsilon} + W_0\right] \\
&= W_0 - E\left[\int_0^T \pi_{\chi^*}(t)[(1-\nu^*(t))c^*(t) - \overrightarrow{g}(\chi^*(t), t)]\,dt + \pi_{\chi^*}(T)W^*\right],
\end{aligned}$$

where the second and third equalities above follow from the dominated convergence theorem and (A.3). It follows that

$$W_0 = E\left[\int_0^T \pi_{\chi^*}(t)[(1-\nu^*(t))c^*(t) - \overrightarrow{g}(\chi^*(t), t)]\,dt + \pi_{\chi^*}(T)W^*\right].$$

Let $(c(.), W)$ be any feasible liquidation and terminal wealth process. Since

$$\begin{aligned}
&u_1(f_1(y,t),t) - u_1(c,t) \geq y[f_1(y,t) - c] \quad \forall c > 0, y > 0, \\
&u_2(f_2(y,T),T) - u_2(W,T) \geq y[f_2(y,T) - W] \quad \forall W > 0, y > 0,
\end{aligned}$$

(3.5) leads to

$$\begin{aligned}
U(c^*, W^*) - U(c, W) &= E\left[\int_0^T (u_1(c^*(t), t) - u_1(c(t), t))\,dt + u_2(W^*, T) - u_2(W, T)\right] \\
&\geq \lambda^* E\left[\int_0^T \pi_{\chi^*}(t)(1-\nu^*(t))(c^*(t) - c(t))\,dt\right] + \lambda^* E[\pi_{\chi^*}(T)(W^* - W)] \geq 0.
\end{aligned}$$

Therefore, $(c^*(.), W^*)$ is optimal *provided it is feasible.*

Step 2: *Feasibility of* $(c^*(.), W^*)$. From (4.7), we see that

$$\pi_{\chi^*}(t)W_{\chi^*}(t) + \int_0^t \pi_{\chi^*}(s)((1-\nu^*(s))c^*(s) - \overrightarrow{g}(\chi^*(s), s))\,ds$$

is a P-martingale. By the martingale representation theorem, there exists a predictable process $\phi(.)$ with $\int_0^T \phi(t)^2 dt < \infty$ a.s. such that

$$\pi_{\chi^*}(t)W_{\chi^*}(t) + \int_0^t \pi_{\chi^*}(s)((1-\nu^*(s))c^*(s) - \overrightarrow{g}(\chi^*(s), s))\,ds = W_0 + \int_0^t \phi(s)\,dB(s).$$

Define the pair (α, θ) by

$$\pi_{\chi^*}(t)(\sigma(t)\theta(t) + W_{\chi^*}(t)\kappa_{\chi^*}(t)) = \phi(t),$$
$$\alpha(t) = W_{\chi^*}(t) - \theta(t).$$

By Itŏ's lemma, we easily see that

$$W_{\chi^*}(t) = W_0 + \int_0^t [\alpha(s)\mu_0^*(s) + \theta(s)(\overline{\mu}(s) + \mu_1^*(s)) + c^*(s)\nu^*(s) + \overrightarrow{g}(\chi^*(s), s)]\, ds$$
$$+ \int_0^t \theta(s)\sigma(s)\, dB(s) - \int_0^t c^*(s)\, ds.$$

Therefore, in order to prove that $(c^*(.), W^*)$ is feasible, we need to show, from (2.2) and the definition of the feedback function $h(.)$, that

$$\alpha(.) \equiv 0,$$
$$h(\theta(t), c^*(t), t) = \overrightarrow{g}(\chi^*(t), t) + \theta(t)\mu_1^*(t) + c^*(t)\nu^*(t),$$

and $W_{\chi^*}(t) \geq 0$ a.s. for $t \in [0, T]$ with $W_{\chi^*}(T) = W^*$. We shall now prove these assertions. By the result of Lemma 2.1, there exist processes $(\mu_1(.), \nu^{n'}(.)) \in M$ such that

$$h(\theta(t), c^*(t), t) = \lim_{n' \to \infty} \left[\overrightarrow{h}(\mu_1(t), \nu^{n'}(t), t) + \theta(t)\mu_1(t) + c^*(t)\nu^{n'}(t)\right] \tag{A.4}$$

and a process $\nu(.)$ with $\int_0^T \nu(t)c^*(t)\, dt < \infty$ a.s. such that $\nu^{n'}(.) \uparrow \nu(.)$ a.e. Define a process $\mu_0^n(.)$ by

$$\mu_0^n(t) = (\mu_0^*(t) - n)1_{\alpha(t)>0} + (\mu_0^*(t) + n)1_{\alpha(t)<0}. \tag{A.5}$$

Define the two-dimensional sequence of processes $\chi^{n,n'}(.)$ by

$$\chi^{n,n'}(t) = (\mu_0^n(t), \mu_1(t), \nu^{n'}(t)). \tag{A.6}$$

By Assumption 6 and (2.15), (2.16), (A.4), and (A.5), we easily see that

$$-\infty 1_{\alpha(t)\neq 0} + h(\theta(t), c^*(t), t)1_{\alpha(t)=0} = \lim_{n,n' \to \infty} \overrightarrow{h}(\mu_1(t), \nu^{n'}(t), t) + \alpha(t)\mu_0^n(t)$$
$$+ \theta(t)\mu_1(t) + c^*(t)\nu^{n'}(t) = \lim_{n,n' \to \infty} \overrightarrow{g}(\chi^{n,n'}(t), t) + \alpha(t)\mu_0^n(t) + \theta(t)\mu_1(t) + c^*(t)\nu^{n'}(t)$$

since $\overrightarrow{h}(\mu_1(t), \nu^{n'}(t), t) = \overrightarrow{g}(\chi^{n,n'}(t), t)$ by (2.15). Define the processes

$$\zeta^n(t) = \int_0^t (\mu_0^n(s) - \mu_0^*(s))\, ds$$
$$+ \int_0^t \sigma(s)^{-1}(\mu_1(s) - \mu_1^*(s) - (\mu_0^n(s) - \mu_0^*(s)))(dB(s) - \kappa_{\chi^*}(s)\, ds)$$

and the sequence of stopping times (for each n)

$$\tau_m^n = T \wedge \inf \left\{ \begin{array}{l} t \in [0, T] : |\zeta^n(t)| + |\pi_{\chi^*}(t)| + |W_{\chi^*}(t)| \geq m, \quad \text{or} \\ \displaystyle\int_0^t \left[|\sigma(s)\theta(s) + W_{\chi^*}(s)\kappa_{\chi^*}(s)|^2 + |\ \nu(s)c^*(s)\ |\right] ds \geq m \end{array} \right\}.$$

Then $\tau_m^n \uparrow T$ a.s. for each n. Denoting

$$\chi_{\varepsilon,m}^{n,n'}(t) = \chi^*(t) + \varepsilon[\chi^{n,n'}(t) - \chi^*(t)]1_{t\le\tau_m^n}$$

for $\varepsilon \in (0,1)$, we have $\chi_{\varepsilon,m}^{n,n'} \in N$ (since the sets N_t are convex). We now see that

(A.7)

$$\begin{aligned}
&\lim_{\varepsilon\downarrow 0} \frac{J(\lambda^*, \chi_{\varepsilon,m}^{n,n'}) - J(\lambda^*, \chi^*)}{\varepsilon} \\
&\quad = \lim_{\varepsilon\downarrow 0} E\left[\int_0^T \frac{\overline{u_1}(\lambda^*(1-\nu_{\varepsilon,m}^{n'}(t))\pi_{\chi_{\varepsilon,m}^{n,n'}}(t),t) - \overline{u_1}(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)}{\varepsilon}\,dt\right] \\
&\qquad\qquad - \lambda^*\int_0^T \pi_{\chi^*}(t)\frac{\overrightarrow{g}(\chi^*(t),t)}{\varepsilon}\left(1 - \frac{\pi_{\chi_{\varepsilon,m}^{n,n'}}(t)}{\pi_{\chi^*}(t)}\right)dt \\
&\qquad\qquad + \lambda^*\int_0^T \pi_{\chi_{\varepsilon,m}^{n,n'}}(t)\frac{\overrightarrow{g}(\chi_{\varepsilon,m}^{n,n'}(t),t) - \overrightarrow{g}(\chi^*(t),t)}{\varepsilon}\,dt \\
&\qquad\qquad + \frac{\overline{u_2}(\lambda^*\pi_{\chi_{\varepsilon,m}^{n,n'}}(T),T) - \overline{u_2}(\lambda^*\pi_{\chi^*}(T),T)}{\varepsilon} \\
&\quad \le \lambda^* E\left[\int_0^T \zeta^n(t\wedge\tau_m^n)\pi_{\chi^*}(t)((1-\nu^*(t))c^*(t) - \overrightarrow{g}(\chi^*(t),t))\,dt\right. \\
&\qquad\qquad + \int_0^{\tau_m^n} \pi_{\chi^*}(t)(\nu^{n'}(t) - \nu^*(t))c^*(t)\,dt \\
&\qquad\qquad + \int_0^{\tau_m^n} \pi_{\chi^*}(t)(\overrightarrow{g}(\chi^{n,n'}(t),t) - \overrightarrow{g}(\chi^*(t),t))\,dt \\
&\qquad\qquad \left. + \zeta^n(\tau_m^n)\pi_{\chi^*}(T)W^*\right] \\
&\quad = E\left[\int_0^{\tau_m^n} \zeta^n(t)\pi_{\chi^*}(t)((1-\nu^*(t))c^*(t) - \overrightarrow{g}(\chi^*(t),t))\,dt + \zeta^n(\tau_m^n)\pi_{\chi^*}(\tau_m^n)W_{\chi^*}(\tau_m^n)\right. \\
&\qquad\qquad \left. + \int_0^{\tau_m^n} \pi_{\chi^*}(t)(\overrightarrow{g(}\chi^{n,n'}(t),t) + \nu^{n'}(t)c^*(t) - \overrightarrow{g}(\chi^*(t),t) - \nu^*(t)c^*(t))\,dt\right],
\end{aligned}$$

where the first inequality above follows from the convexity of $\overrightarrow{g}$ and the dominated convergence theorem and the last equality follows from the definition (4.7) of W_{χ^*}. The first term in the sequence of expressions above $\lim_{\varepsilon\downarrow 0}\frac{J(\lambda^*,\chi_{\varepsilon,m}^{n,n'})-J(\lambda^*,\chi^*)}{\varepsilon}$ is nonnegative for all n, n', m since $J(\lambda^*, \chi_{\varepsilon,m}^{n,n'})$ attains a minimum at $\varepsilon = 0$ as χ^* is optimal by hypothesis. By an application of Itŏ's lemma, we can easily show that

$$\begin{aligned}
&E\left[\int_0^{\tau_m^n} \zeta^n(t)\pi_{\chi^*}(t)[(1-\nu^*(t))c^*(t) - \overrightarrow{g}(\chi^*(t),t)]\,dt + \zeta^n(\tau_m^n)\pi_{\chi^*}(\tau_m^n)W_{\chi^*}(\tau_m^n)\right] \\
&\qquad = E\left[\int_0^{\tau_m^n} \pi_{\chi^*}(t)[\alpha(t)(\mu_0^n(t) - \mu_0^*(t)) + \theta(t)(\mu_1(t) - \mu_1^*(t))]\,dt\right].
\end{aligned}$$

Substituting the above into (A.7), we see that

(A.8)

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0}\frac{J(\lambda^*,\chi^{n,n'}_{\varepsilon,m})-J(\lambda^*,\chi^*)}{\varepsilon}\\ &\quad\le E\left[\int_0^{\tau^n_m}\pi_{\chi^*}(t)(\alpha(t)(\mu^n_0(t)-\mu^*_0(t))+\theta(t)(\mu_1(t)-\mu^*_1(t))+c^*(t)(\nu^{n'}(t)-\nu^*(t))\right.\\ &\qquad\qquad\left.+\ \overrightarrow{g}(\chi^{n,n'}(t),t)-\overrightarrow{g}(\chi^*(t),t))\,dt\right].\end{aligned}$$

The left-hand side of the equation above is nonnegative by hypothesis for all n, n', m. If Λ denotes Lebesgue measure on $[0,T]$, and if $\Lambda\times P((t,\omega):\alpha(t,\omega)\neq 0)>0$, then (A.5) and (A.6) imply that

$$\lim_{n\to\infty}E\left[\int_0^T\pi_{\chi^*}(t)\alpha(t)(\mu^n_0(t)-\mu^*_0(t))\,dt\right]=-\infty.$$

Since $\overrightarrow{g}(\chi^{n,n'}(t),t)=\overrightarrow{h}(\mu_1(t),\nu^{n'}(t),t)$ does not depend on $\mu^n_0()$, we easily see that the right-hand side of (A.8) is strictly negative for sufficiently large n and m, which contradicts the fact that the left-hand side of (A.8) is nonnegative by hypothesis for every n, n' and m. Hence, $\alpha(t)=0$, $\Lambda\times P$ a.s. on $[0,T]\times\Omega$. Therefore, we have

(A.9)

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0}\frac{J(\lambda^*,\chi^{n,n'}_{\varepsilon,m})-J(\lambda^*,\chi^*)}{\varepsilon}\\ &\le E\left[\int_0^{\tau^n_m}\pi_{\chi^*}(t)(\theta(t)(\mu_1(t)-\mu^*_1(t))+c^*(t)(\nu^{n'}(t)-\nu^*(t))\right.\\ &\qquad\left.+\ \overrightarrow{g}(\chi^{n,n'}(t),t)-\overrightarrow{g}(\chi^*(t),t))\,dt\right]\\ &=E\left[\int_0^{\tau^n_m}\pi_{\chi^*}(t)(\theta(t)(\mu_1(t)-\mu^*_1(t))+c^*(t)(\nu^{n'}(t)-\nu^*(t))\right.\\ &\qquad\left.+\ \overrightarrow{h}(\mu_1(t),\nu^{n'}(t),t)-\overrightarrow{h}(\mu^*_1(t),\nu^*(t),t))\,dt\right]\\ &=E\left[\int_0^{\tau^n_m}\pi_{\chi^*}(t)[(-\theta(t)\mu^*_1(t))-c^*(t)(\nu^*(t))+h(\theta(t),c^*(t),t)-\overrightarrow{h}(\mu^*_1(t),\nu^*(t),t)\right.\\ &\qquad\left.+\ \overrightarrow{h}(\mu_1(t),\nu^{n'}(t),t)+\theta(t)\mu_1(t)+c^*(t)\nu^{n'}(t)-h(\theta(t),c^*(t),t)]\,dt\right].\end{aligned}$$

Since

$$h(\theta(t),c^*(t),t)=\lim_{n'\to\infty}[\overrightarrow{h}(\mu_1(t),\nu^{n'}(t),t)+\theta(t)\mu_1(t)+c^*(t)\nu^{n'}(t)]$$

and

$$\overrightarrow{h}(\mu^*_1(t),\nu^*(t),t)\ge h(\theta(t),c^*(t),t)-\theta(t)\mu^*_1(t)-c^*(t)\nu^*(t)$$

and the first expression in (A.9) is nonnegative by hypothesis, we see that equality must in fact hold in the expression above by applying the dominated convergence theorem and using the fact that $h(.,t,\omega)$ is uniformly Lipschitz and $\overrightarrow{h}$ is uniformly bounded by assumption. Since $\alpha(t) = 0$, and $\overrightarrow{g}(\chi^*(t),t) = \overrightarrow{h}(\mu_1^*(t), \nu^*(t), t)$ by (2.15), we see that

$$h(\theta(t), c^*(t), t) = \overrightarrow{g}(\chi^*(t),t) + \theta(t)\mu_1^*(t) + c^*(t)\nu^*(t).$$

We have therefore shown that $\alpha(.) \equiv 0$ and

$$h(\theta(t), c^*(t), t) = \overrightarrow{h}(\mu_1^*(t), \nu^*(t), t) + \theta(t)\mu_1^*(t) + c^*(t)\nu^*(t).$$

Clearly, $W_{\chi^*}(T) = W$ by (4.7). Therefore, it remains only to show that $W_{\chi^*}(t) \geq 0$ a.s. for all $t \in [0,T]$. *We prove this by contradiction.*

Since $\alpha(s) = W_{\chi^*}(s) - \theta(s)$ by definition, we see that $\alpha(s) = 0$ implies that $\theta(s) = W_{\chi^*}(s)$. Therefore, we have

$$\begin{aligned} W_{\chi^*}(t) = W_0 &+ \int_0^t [W_{\chi^*}(s)\overline{\mu}(s) + h(W_{\chi^*}(s), c^*(s), s)]\, ds \\ &+ \int_0^t W_{\chi^*}(s)\sigma(s)\, dB(s) - \int_0^t c^*(s)\, ds. \end{aligned}$$

Suppose, *to the contrary*, that $P(W_{\chi^*}(t) < 0) > 0$ for some $t \in [0,T]$. Let $\tau_\varepsilon = \inf\{t : W_{\chi^*}(t) < -\varepsilon\} \wedge T$. Clearly, $P[\tau_\varepsilon < T] > 0$ for some $\varepsilon > 0$. Clearly, $W_{\chi^*}(\tau) \leq 0$ on $\tau_\varepsilon < T$. On $[\tau_\varepsilon, T]$, we have

$$\begin{aligned} W_{\chi^*}(t) = W_{\chi^*}(\tau_\varepsilon) &+ \int_{\tau_\varepsilon}^t [W_{\chi^*}(s)\overline{\mu}(s) + h(W_{\chi^*}(s), c^*(s), s)]\, ds \\ &+ \int_0^t W_{\chi^*}(s)\sigma(s)\, dB(s) - \int_0^t c^*(s)\, ds. \end{aligned}$$

Since $h(W,c) \leq 0$ for $W \leq 0$ from the definition (2.10) of $h(.)$ and Assumption 5, we can compare the above equation with the equation

$$\widehat{W}(t) = W_{\chi^*}(\tau_\varepsilon) + \int_{\tau_\varepsilon}^t [\widehat{W}(s)\overline{\mu}(s)]\, ds + \int_{\tau_\varepsilon}^t \widehat{W}(s)\sigma(s)\, dB(s).$$

By the comparison theorem for stochastic differential equations (see Proposition 2.18 in Karatzas and Shreve [19]), we see that on $[\tau_\varepsilon, T]$, $P[W_{\chi^*}(t) \leq \widehat{W}(t)] = 1$. Since $\widehat{W}(T) < 0$ on the set $\{\tau_\varepsilon < T\}$ as $W_{\chi^*}(\tau) \leq 0$ on$\tau_\varepsilon < T$, we have $P[W_{\chi^*}(T) \leq \widehat{W}(T) < 0] > 0$, which *contradicts* the fact that $W_{\chi^*}(T) = W \geq 0$ a.s. This contradiction allows us to conclude that $W_{\chi^*}(t) \geq 0$ a.s. This completes the proof. □

Proof of Lemma 4.2. By Lemma 4.1 and Assumption 4, $\overline{u_1}(.,t), \overline{u_2}(.,T)$ are both bounded below. For a fixed process $(\mu_0(.), \mu_1(.))$ and a fixed λ as in the statement of the lemma, we note that

$$\overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) + \lambda\pi_\chi(t,\omega)\overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)$$

is a strictly convex, lower semicontinuous function of $\eta \in D_{(\mu_0(.),\mu_1(.))}(t,\omega)$ that is uniformly bounded below. Since $D_{(\mu_0(.),\mu_1(.))}(t,\omega)$ is closed, convex, and bounded for each (t,ω), we see that

$$\inf_{\eta \in D_{(\mu_0(.),\mu_1(.))}(t,\omega)} \overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) + \lambda\pi_\chi(t,\omega)\overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)$$

exists and is uniquely attained by some $\eta^*_{(t,\omega)}$. Moreover, by measurable selection theorems of the Dubins–Savage type, the process $\nu^*(.)$ defined by $\nu^*(t,\omega) = \eta^*_{(t,\omega)}$ is bounded and progressively measurable. By an application of Fatou's lemma, we see that if

$$\begin{aligned}&\inf_{\nu(.)\in D_{(\mu_0(.),\mu_1(.))}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.))\\ &\quad = E\left[\int_0^T \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t), t)\,dt + \lambda\int_0^T \pi_\chi(t)\overrightarrow{g}(\chi(t),t)\,dt + \overline{u_2}(\lambda\pi_\chi(T),T)\right] + \lambda W_0\end{aligned}$$

exists, it is uniquely attained by the process $\nu^*(.)$. From (3.4b), it follows easily that $\{(\pi_\chi, \pi_\chi\mu_0, \pi_\chi\mu_1) : \chi \in N\}$ is a convex set and that $\pi_\chi$is uniquely determined by $(\mu_0(.), \mu_1(.))$. It clearly suffices to consider the case where the infimum above is attained. Let

$$\begin{aligned}&\inf_{\nu(.)\in D_{(\mu_0^1(.),\mu_1^1(.))}} J(\lambda, \mu_0^1(.), \mu_1^1(.), \nu(.)) = J(\lambda, \mu_0^1(.), \mu_1^1(.), \nu^1(.)),\\ &\inf_{\nu(.)\in D_{(\mu_0^2(.),\mu_1^2(.))}} J(\lambda, \mu_0^2(.), \mu_1^2(.), \nu(.)) = J(\lambda, \mu_0^2(.), \mu_1^2(.), \nu^2(.)).\end{aligned}$$

By the convexity of $\overline{u_1}(.), \overline{u_2}, \overrightarrow{g}(.)$, we have (with $0 < q < 1$)

$$\begin{aligned}&(1-q)J(\lambda, \mu_0^1(.), \mu_1^1(.), \nu^1(.)) + qJ(\lambda, \mu_0^2(.), \mu_1^2(.), \nu^2(.))\\ &\quad > E\Bigg[\int_0^T \overline{u_1}((1-q)\lambda(1-\nu^1(t))\pi_{\chi^1}(t) + q\lambda(1-\nu^2(t))\pi_{\chi^2}(t), t)\,dt\\ &\qquad + \lambda\int_0^T [(1-q)\pi_{\chi^1}(t)\overrightarrow{g}(\chi^1(t),t) + q\pi_{\chi^2}(t)\overrightarrow{g}(\chi^2(t),t)]\,dt\\ &\qquad + \overline{u_2}(\lambda(1-q)\pi_{\chi^1}(T) + \lambda q\pi_{\chi^2}(T), T)\Bigg] + \lambda W_0.\end{aligned}$$

If we define

$$\begin{aligned}\pi_{\chi^*}(t) &= (1-q)\pi_{\chi^1}(t) + q\pi_{\chi^2}(t),\\ (1-\nu^*(t)) &= \frac{(1-q)\pi_{\chi^1}(t)(1-\nu^1(t)) + q\pi_{\chi^2}(t)(1-\nu^2(t))}{\pi_{\chi^*}(t)},\\ \mu_0^*(t) &= \frac{(1-q)\pi_{\chi^1}(t)\mu_0^1(t) + q\pi_{\chi^2}(t)\mu_0^2(t)}{\pi_{\chi^*}(t)},\\ \mu_1^*(t) &= \frac{(1-q)\pi_{\chi^1}(t)\mu_1^1(t) + q\pi_{\chi^2}(t)\mu_1^2(t)}{\pi_{\chi^*}(t)},\end{aligned}$$

we see that

$$\begin{aligned}&(1-q)J(\lambda, \mu_0^1(.), \mu_1^1(.), \nu^1(.)) + qJ(\lambda, \mu_0^2(.), \mu_1^2(.), \nu^2(.))\\ &\quad > E\Bigg[\int_0^T \overline{u_1}(\lambda(1-\nu^*(t))\pi_{\chi^*}(t), t) + \int_0^T \lambda\pi_{\chi^*}(t)\overrightarrow{g}(\chi^*(t),t)\,dt + \overline{u_2}(\lambda\pi_{\chi^*}(T),T)\\ &\qquad + \lambda W_0\Bigg] \geq \inf_{\nu(.)} J(\lambda, \mu_0^*(.), \mu_1^*(.), \nu(.)).\end{aligned}$$

Note. In the inequalities above, we have used the fact that the set $\{(\pi_\chi, \pi_\chi\mu_0, \pi_\chi\mu_1)\colon \chi \in N'\}$ is a convex set, which follows from relation (3.4b), and that $\overline{u_1}(.), \overline{u_2}(.), \overrightarrow{g}(.)$ are convex functions of their arguments. It follows that $\inf_{\nu(.)\in D_{(\mu_0(.),\mu_1(.))}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.))$ is a convex functional of π_χ.

This completes the proof. □

Proof of Proposition 4.2. Since the proof of this proposition is rather involved, we will split it into several steps. We shall first prove that for each $\lambda \in (0, \infty)$, the problem

$$V(\lambda) = \inf_{\chi \in N'} J(\lambda, \chi) \tag{A.10}$$

has a solution. By hypothesis (c) of the proposition, $V(\lambda)$ exists for each λ.

Step 1. By the result of Lemma 4.2, if

$$\inf_{\nu(.)\in D_{(\mu_0(.),\mu_1(.))}} J(\lambda, \chi)$$

exists, it is attained uniquely by some bounded, progressively measurable process $\nu^*()$ for each fixed λ and square-integrable, progressively measurable process $(\mu_0(.), \mu_1(.))$ for which the set $D_{(\mu_0(.),\mu_1(.))}$ is nonempty.

Step 2. The problem (A.10) can be rewritten as

$$V(\lambda) = \inf_{(\mu_0(.),\mu_1(.))} \inf_{\nu(.)\in D_{\mu_0(.),\mu_1(.)}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.)).$$

Since $V(\lambda)$ exists by hypothesis (c) of the proposition, $\inf_{\nu(.)\in D_{\mu_0(.),\mu_1(.)}} \overline{J}(\lambda, \pi_\chi, \nu(.))$ exists and is attained uniquely by some bounded, progressively measurable process $\nu^*()$ for each fixed λ and square integrable, progressively measurable process $(\mu_0(.), \mu_1(.))$ for which the set $D_{(\mu_0(.),\mu_1(.))}$ is nonempty. By arguments analogous to those used in the proof of Theorem 3 in Cuoco and Cvitanic [5], the problem above can be rewritten as

$$V(\lambda) = \inf_{(\pi_\chi:\chi\in N')} \inf_{\nu(.)\in D_{\mu_0(.),\mu_1(.)}} \overline{J}(\lambda, \pi_\chi, \nu(.)),$$

where $\overline{J}(\lambda, \pi_\chi, \nu(.)) = J(\lambda, \mu_0(.), \mu_1(.), \nu(.))$.

Note. In the above, we have used the fact that π_χ (defined in (3.4a)) does not depend on $\nu(.)$.

Moreover, by arguments identical to those in the proof of Theorem 3 in Cuoco and Cvitanic [5], the set $\Pi = (\pi_\chi : \chi \in N')$ is a convex subset of $L^2([0,T] \times \Omega)$ and the functional $\overline{J}(\lambda, ., \nu(.))$ is convex on Π. By the result of Lemma 4.2, the functional

$$T(\lambda, .) = \inf_{\nu(.)} \overline{J}(\lambda, \dots, \nu(.)) \text{ is also convex on } \Pi.$$

We can now write

$$V(\lambda) = \inf_{(\pi_\chi:\chi\in N')} T(\lambda, \pi_\chi), \tag{A.11}$$

where $T(\lambda, .)$ is convex on Π.

Step 3. We shall now prove that in the problem (A.11) above, it is enough to consider the set of processes π_χ defined by processes $(\mu_0(.), \mu_1(.))$ for which $\|\mu_0\|_2$ is uniformly bounded by some constant Λ. By Assumption 4 and Lemma 4.1, $\overline{u_1}(c, t)$

is uniformly bounded below for all $(c,t) \in [0,\infty) \times [0,T]$. By hypothesis (c) of the proposition, for each $\lambda \in (0,\infty)$, there exists $\chi^* \in N'$ such that $J(\lambda,\chi^*) = \mathrm{A}(\lambda) < \infty$. We have

(A.12)

$$\begin{aligned}
T(\lambda,\pi_\chi) = &\inf_{\nu(.)\in D_{\mu_0(.),\mu_1(.)}} E\left[\int_0^T \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t),t)\,dt \right.\\
&\qquad\left. + \lambda\int_0^T \pi_\chi(t)\overrightarrow{g}(\chi(t),t)\,dt + \overline{u_2}(\lambda\pi_\chi(T),T)\right] + \lambda W_0\\
\geq{}& \Psi + \overline{u_2}\left(\lambda\left(\left[\exp\left(-E\int_0^T \mu_0(s)\,ds - \frac{1}{2}E\int_0^T \kappa_\chi(s)^2\,ds\right)\right]\right),T\right) + \lambda W_0\\
={}& \Psi + \overline{u_2}\left(\lambda\left(\left[\exp\left(-E\int_0^T \mu_0(s)\,ds - \frac{1}{2}E\right.\right.\right.\right.\\
&\qquad\left.\left.\left.\left. \times\int_0^T(\sigma(s)^{-1}(\overline{\mu}(s)+\mu_1(s)-\mu_0(s)))^2\,ds\right)\right]\right),T\right) + \lambda W_0\\
={}& \Psi + \overline{u_2}\left(\lambda\left(\left[\exp\left(-\frac{1}{2}E\int_0^T \sigma(s)^{-2}\mu_0(s)^2\,ds - E\right.\right.\right.\right.\\
&\qquad \int_0^T\left[\sigma(s)^{-1}\mu_0(s)(\overline{\mu}(s)+\mu_1(s)+\sigma(s))\right.\\
&\qquad\left.\left.\left.\left.\left. -\frac{1}{2}\sigma(s)^{-2}(\overline{\mu}(s)+\mu_1(s))^2\right]ds\right)\right]\right),T\right) + \lambda W_0\\
\geq{}& \Psi + \overline{u_2}\left(\lambda\left(\exp\left(-\frac{1}{2}E\int_0^T \sigma(s)^{-2}\mu_0(s)^2\,ds + ME\int_0^T |\mu_0(s)|\,ds\right)\right),T\right),
\end{aligned}$$

where Ψ is a lower bound on the integral $\int_0^T \overline{u_1}(\lambda(1-\nu(t))\pi_\chi(t),t)\,dt$. The first inequality in (A.12) follows from Jensen's inequality, condition (b) of the proposition, and the fact that $\overrightarrow{g}(.)$ is nonnegative and independent of $\mu_0(.)$ by (2.16). The last inequality follows from the fact that $\overline{u_2}(.)$ is decreasing. The constant M in the last inequality above is the uniform bound on $|\sigma(s)^{-1}(\overline{\mu}(s)+\mu_1(s))+1|$ which exists by Assumption 1.

Since $J(\lambda,\chi^*) = \mathrm{A}(\lambda) < \infty$ for some $\chi^* \in N'$, it is clearly enough in problem (A.11) to consider $\chi \in N'$ for which

$$\begin{aligned}
&\overline{u_2}\left(\lambda\left(\exp\left(-\frac{1}{2}E\int_0^T \sigma(s)^{-2}\mu_0(s)^2\,ds + ME\int_0^T |\mu_0(s)|\,ds\right)\right),T\right) \leq A(\lambda) \quad \text{or}\\
&\exp\left(-\frac{1}{2}E\int_0^T \sigma(s)^{-2}\mu_0(s)^2\,ds + ME\int_0^T |\mu_0(s)|\,ds\right) \geq \frac{1}{\lambda}\overline{u_2}^{-1}(A(\lambda),T)
\end{aligned}$$

(since $\overline{u_2}$ is decreasing), or

$$E\int_0^T \sigma(s)^{-2}\mu_0(s)^2\,ds - ME\int_0^T |\mu_0(s)|\,ds$$

is uniformly bounded above, which in turn implies that $\|\mu_0\|_2$ is uniformly bounded by some constant Λ since $\sigma(.)$ and $\sigma^{-1}(.)$ are uniformly bounded by assumption. Let us denote $\Pi_\Lambda = \{\pi_\chi : \chi \in N'; \|\mu_0\|_2 \leq \Lambda\}$. Then, we can rewrite problem (A.11) as

$$V(\lambda) = \inf_{(\pi_\chi : \pi_\chi \in \Pi_\Lambda)} T(\lambda, \pi_\chi). \tag{A.13}$$

Step 4. Let $\Pi_\Lambda^l = \Pi_\Lambda \cap \{\pi_\chi : \chi \in N'; \|\mu_0\|_\infty \leq l\}$ for each $l \in Z_{++}$; i.e., Π_Λ^l is the subset of Π_Λ consisting of those processes π_χ for which $|\mu_0|$ is uniformly bounded by l. Clearly, for each $\pi_\chi \in \Pi_\Lambda$, there exists a sequence $\pi_{\chi^n} \in \Pi_\Lambda^n$ such that

$$\lim_{n\to\infty} \pi_{\chi^n}(t,\omega) = \pi_\chi(t,\omega) \text{ a.e. on } [0,T] \times \Omega. \tag{A.14}$$

In this step, we prove that the problem

$$V^l(\lambda) = \inf_{(\pi_\chi : \pi_\chi \in \Pi_\Lambda^l)} T(\lambda, \pi_\chi)$$

has a solution; i.e., the minimum above *exists and is attained*. Since $\chi \in N' \Rightarrow |\mu_1|$ is uniformly bounded (Assumption 6) , we can use arguments identical to those used in the proof of Theorem 3 in Cuoco and Cvitanic [5] to show that Π_Λ^l is convex, uniformly bounded, and closed in $L^2([0,T] \times \Omega)$.

We now prove that $T(\lambda, .)$ is lower semicontinuous on Π_Λ^l. *Suppose this is not the case.* Then, there is an $\alpha > 0$, $\pi_\chi \in \Pi_\Lambda^l$, and a sequence $\{\pi_{\chi^n}\}$ converging to π_χ in $L^2([0,T] \times \Omega)$ such that $T(\lambda, \pi_{\chi^n}) \leq \alpha < T(\lambda, \pi_\chi)$. Since Π_Λ^l is convex, uniformly bounded, and closed in $L^2([0,T] \times \Omega)$, there exists a subsequence $\{\pi_{\overline{\chi^n}}\} \subset co(\{\pi_{\chi^n}\})$ such that $(\pi_{\overline{\chi^n}}, \overline{\mu_0^n}, \overline{\mu_1^n}) \to (\pi_\chi, \mu_0, \mu_1)$ a.e. and it follows from the convexity of $T(\lambda, .)$ that $T(\lambda, \pi_{\overline{\chi^n}}) \leq \alpha$ for all n. By the result of Step 1, there exist processes $\overline{\nu^n}$ and ν such that $(\overline{\mu_0^n}, \overline{\mu_1^n}, \overline{\nu^n}) \in N'$ and $(\mu_0, \mu_1, \nu) \in N'$ satisfying

$$T(\lambda, \pi_{\overline{\chi^n}}) = \overline{J}(\lambda, \pi_{\overline{\chi^n}}, \overline{\nu^n}),$$
$$T(\lambda, \pi_\chi) = \overline{J}(\lambda, \pi_\chi, \nu).$$

Moreover, it is proved in Lemma A.1 following this proposition that $\overline{\nu^n} \to \nu$ a.e. Due to the continuity of $\overline{u_1}(., t), \overline{u_2}$, and $\overrightarrow{g}(.)$ and the fact that they are uniformly bounded below, we can apply Fatou's lemma to conclude that

$$\alpha < T(\lambda, \pi_\chi) = \overline{J}(\lambda, \pi_\chi, \nu) \leq \liminf_{n\to\infty} \overline{J}(\lambda, \pi_{\overline{\chi^n}}, \overline{\nu^n}) = \liminf_{n\to\infty} T(\lambda, \pi_{\overline{\chi^n}}) \leq \alpha,$$

which is a contradiction. Therefore, $T(\lambda, .)$ *is lower semicontinuous on* Π_Λ^l. We have therefore shown that Π_Λ^l is *convex, uniformly bounded, and closed* in $L^2([0,T] \times \Omega)$ and that $T(\lambda, .)$ is *convex and lower semicontinuous* on Π_Λ^l. We can now apply the results of Proposition 2.1.2 in Ekeland and Temam [9] to conclude that

$$V^l(\lambda) = \inf_{(\pi_\chi : \pi_\chi \in \Pi_\Lambda^l)} T(\lambda, \pi_\chi) \tag{A.15}$$

exists and is attained. Moreover, by (A.13), (A.14), Lemma A.1, and Fatou's lemma,

$$\lim_{l\to\infty} V^l(\lambda) = V(\lambda) \text{ and } V^l(\lambda) \text{ decreases with } l.$$

Step 5. We now show that the minimum in the definition (A.10) of $V(\lambda)$ is attained. By the results of the previous steps, we have for each l $V^l(\lambda) = J(\lambda, \mu_0^l(.),$

$\mu_1^l(.), \nu^l(.))$ for some sequence of processes $\{(\mu_0^l(.), \mu_1^l(.), \nu^l(.))\}$ which is uniformly bounded in $L_3^2([0,T] \times \Omega)$; i.e., $\{\|\mu_0^l\|_2 + \|\mu_1^l\|_2 + \|\nu^l\|_2\}$ is uniformly bounded (using the result of Step 3).

Since the space $L_3^2([0,T] \times \Omega)$ is locally convex and meterizable, it follows that there is a subsequence, again denoted by $\{(\mu_0^l(.), \mu_1^l(.), \nu^l(.))\}$ for notational simplicity, converging weakly to some process $(\mu_0(.), \mu_1(.), \nu(.)) \in N'$. Therefore, there exists a further sequence $\{(\mu_0^{l'}(.), \mu_1^{l'}(.), \nu^{l'}(.))\} \subset co\{(\mu_0^l(.), \mu_1^l(.), \nu^l(.))\}$ that converges a.e. to $(\mu_0(.), \mu_1(.), \nu(.))$. By the continuity and convexity of $\overline{u_1}(., t), \overline{u_2}(., t)$ and $\overrightarrow{g}(., t, \omega)$ and the fact that they are uniformly bounded below, we can use Fatou's lemma to see that

$$J(\lambda, \mu_0(.), \mu_1(.), \nu(.)) \leq \liminf_{l\to\infty} J(\lambda, \mu_0^l(.), \mu_1^l(.), \nu^l(.)) = \liminf_{l\to\infty} V^l(\lambda) = V(\lambda).$$

Therefore, $J(\lambda, \mu_0(.), \mu_1(.), \nu(.)) = V(\lambda)$ and $(\mu_0(.), \mu_1(.), \nu(.))$ is an optimal solution.

Step 6. We now prove that the function $V(\lambda)$ is strictly convex and coercive for $\varepsilon(0, \infty)$, which would imply that

$$\inf_{\lambda \in (0,\infty)} V(\lambda)$$

exists and is attained. Let $\chi^1_, \chi^2$ be the optimal solutions derived in *Step 5* corresponding to λ_1, λ_2, respectively, and let $\lambda = (1-q)\lambda_1 + q\lambda_2$, where $0 < q < 1$. Therefore,

$$V(\lambda_i) = J(\lambda_i, \chi^i) = E\left[\int_0^T \overline{u_1}(\lambda_i(1-\nu^i(t))\pi_{\chi^i}(t), t)\,dt \right.$$
$$\left. + \lambda_i \int_0^T \pi_{\chi^i}(t) \overrightarrow{g}(\chi^i(t), t)\,dt + \overline{u_2}(\lambda_i \pi_{\chi^i}(T), T)\right] + \lambda_i W_0; i = 1, 2.$$

By the strict convexity of $\overline{u_1}, \overline{u_2}$ and the convexity of $\overrightarrow{g}$, we have

$$(1-q)V(\lambda_1) + qV(\lambda_2)$$
$$> E\left[\int_0^T \overline{u_1}((1-q)\lambda_1(1-\nu^1(t))\pi_{\chi^1}(t) + q\lambda_2(1-\nu^2(t))\pi_{\chi^2}(t), t) \right.$$
$$+ (1-q)\lambda_1 \int_0^T \pi_{\chi^1}(t)\overrightarrow{g}(\chi^1(t), t)\,dt + q\lambda_2 \int_0^T \pi_{\chi^2}(t)\overrightarrow{g}(\chi^2(t), t)\,dt$$
$$\left. + \overline{u_2}((1-q)\lambda_1\pi_{\chi^1}(T) + q\lambda_2\pi_{\chi^2}(T), T) + \lambda W_0\right].$$

If we define

$$\pi_{\chi^*}(t) = \frac{(1-q)\lambda_1\pi_{\chi^1}(t) + q\lambda_2\pi_{\chi^2}(t)}{\lambda},$$

$$(1-\nu^*(t)) = \frac{(1-q)\lambda_1\pi_{\chi^1}(t)(1-\nu^1(t)) + q\lambda_2\pi_{\chi^2}(t)(1-\nu^2(t))}{\lambda\pi_{\chi^*}(t)},$$

$$\mu_0^*(t) = \frac{(1-q)\lambda_1\pi_{\chi^1}(t)\mu_0^1(t) + q\lambda_2\pi_{\chi^2}(t)\mu_0^2(t)}{\lambda\pi_{\chi^*}(t)},$$

$$\mu_1^*(t) = \frac{(1-q)\lambda_1\pi_{\chi^1}(t)\mu_1^1(t) + q\lambda_2\pi_{\chi^2}(t)\mu_1^2(t)}{\lambda\pi_{\chi^*}(t)},$$

we easily see that

$$\begin{aligned}&(1-q)V(\lambda_1)+qV(\lambda_2)\\&\quad> E\left[\int_0^T \overline{u_1}(\lambda(1-\nu^*(t))\pi_{\chi^*}(t),t)\right.\\&\qquad\left.+\int_0^T \lambda\pi_{\chi^*}(t)\overrightarrow{g}(\chi^*(t),t)\,dt+\overline{u_2}(\lambda\pi_{\chi^*}(T),T)+\lambda W_0\right]\geq V(\lambda).\end{aligned}$$

Note. In the inequalities above, we have used the fact that the set $\{(\pi_\chi,\pi_\chi\mu_0,\pi_\chi\mu_1):\chi\in N'\}$ is a convex set, which follows from relation (3.4b).

This proves the strict convexity of $V(.)$. Since $W_0>0$ by assumption and $\overline{u_1},\overline{u_2},\overrightarrow{g}$ are uniformly bounded below,

$$\lim_{\lambda\to\infty} J(\lambda,\chi)=\infty \text{ uniformly in } \chi(.).$$

From (A.12),

$$\begin{aligned}J(\lambda,\chi)&\geq\overline{u_2}\left(\lambda\left(\exp\left(-\frac{1}{2}E\int_0^T\sigma(s)^{-2}\mu_0(s)^2\,ds+ME\int_0^T|\mu_0(s)|\,ds\right)\right),T\right)\\&\geq\overline{u_2}\left(\lambda\exp\left(\frac{1}{2}E\int_0^T M^2\sigma(s)^2\,ds\right),T\right)\geq\overline{u_2}\left(\lambda\exp\left(\frac{1}{2}M^2K^2T\right),T\right),\end{aligned}$$

where K is the uniform upper bound on $\sigma(.)$. By Lemma 4.1, we therefore have

$$\lim_{\lambda\to 0} J(\lambda,\chi)=\infty \textit{ uniformly} \text{ in } \chi(.).$$

Therefore, $\lim_{\lambda\to 0}V(\lambda)=\lim_{\lambda\to\infty}V(\lambda)=\infty$. Hence, $V(.)$ is strictly convex and coercive on $(0,\infty)$. This completes the proof of the existence of an optimal solution to the dual problem. Let the optimal solution be denoted by $(\lambda^*,\mu_0^*(.),\mu_1^*(.),\nu^*(.))$.

Step 7. It remains only to verify that condition (4.5) of the verification Proposition 4.1 is satisfied. By condition (e) of the proposition, we have

$$\begin{aligned}yf_1(y,t)&\leq a+(1-b)u_1(f_1(y,t),t),\\yf_2(y,T)&\leq a+(1-b)u_2(f_2(y,T),T).\end{aligned}$$

Therefore,

$$\begin{aligned}&E\left[\int_0^T\pi_{\chi^*}(t)f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt+\pi_{\chi^*}(T)f_2(\lambda^*\pi_{\chi^*}(T),T)\right]\\&\qquad\leq\frac{2a}{b\lambda^*}+\frac{1-b}{b\lambda^*}E\left[\int_0^T\overline{u_1}(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt+\overline{u_2}(\lambda^*\pi_{\chi^*}(T),T)\right]<\infty\end{aligned}$$

since $|J(\lambda^*,\chi^*)|<\infty$ from assumption (c) of the proposition. By an examination of the proof of Proposition 4.1, we see that the conditions $E[\int_0^T\pi_{\chi^*}(t)\,dt]<\infty$ and $E[\pi_{\chi^*}(T)]<\infty$ are required only to ensure that $E[\int_0^T u_1(c^*(t),t)^-dt]<\infty$ and $E[u_2(W^*,T)^-]<\infty$, which are trivially satisfied since $u_1(.)$ and $u_2(.)$ are assumed to be nonnegative by hypothesis (a) of the proposition. Therefore, by the result of Proposition 4.1, an optimal liquidation policy for the large investor exists. □

LEMMA A.1. *Assume the conditions of Proposition* 4.2. *In the notation of the proof of Proposition* 4.2, *if*

$$T(\lambda, \pi_{\chi^n}) = \inf_{\nu \in D_{(\mu_0^n(.),\mu_1^n(.))}} \overline{J}(\lambda, \pi_{\chi^n}, \nu) = \inf_{\nu \in D_{(\mu_0^n(.),\mu_1^n(.))}} J(\lambda, \mu_0^n(.), \mu_1^n(.), \nu(.))$$

and

$$\begin{aligned} &= J(\lambda, \mu_0^n(.), \mu_1^n(.), \nu^n(.)) < \infty \\ &T(\lambda, \pi_\chi) = \inf_{\nu \in D_{(\mu_{0(.)},\mu_{1(.)})}} \overline{J}(\lambda, \pi_\chi, \nu) = \inf_{\nu \in D_{\mu_0(.),\mu_1(.)}} J(\lambda, \mu_0(.), \mu_1(.), \nu(.)) \\ &= J(\lambda, \mu_0(.), \mu_1(.), \nu^*(.)) < \infty \end{aligned}$$

and $\mu_0^n(.) \to \mu_0(.)$ *a.e.,* $\mu_1^n(.) \to \mu_1(.)$ *a.e., and* $\pi_{\chi^n}(.) \to \pi_\chi(.)$ *a.e., then* $\nu^n(.) \to \nu^*(.)$ *a.e.*

Proof. By the arguments in the proof of Lemma 4.2 we have

$$\begin{aligned} \nu^n(t,\omega) = \arg\min\nolimits_{\eta \in D_{\mu_0^n(.),\mu_1^n(.)}(t,\omega)} [&\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega), t) \\ &+ \lambda \pi_{\chi^n}(t,\omega) \overrightarrow{g}(\mu_0^n(t,\omega), \mu_1^n(t,\omega), \eta, t, \omega)] \end{aligned}$$

and

$$\begin{aligned} \nu^*(t,\omega) = \arg\min\nolimits_{\eta \in D_{\mu_0(.),\mu_1(.)}(t,\omega)} [&\overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) \\ &+ \lambda \pi_\chi(t,\omega) \overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)]. \end{aligned}$$

By condition (a) of Proposition 4.2, the functions being minimized in both equations above are strictly convex functions of η. It follows that $\nu^n(t,\omega)$ and $\nu^*(t,\omega)$ are uniquely defined. By Assumption 7, the sets $D_{\mu_0^n(.),\mu_1^n(.)}(t,\omega)$ and $D_{\mu_0(.),\mu_1(.)}(t,\omega)$ are equal and have the form $[-\Gamma_{(t,\omega)}, 1]$, where $0 \leq \Gamma_{(t,\omega)} < \infty$. Since we cannot have $\nu^*(t,\omega) = 1$ (as $\overline{u_1}(0,t) = \infty$), we can distinguish two cases.

Case 1. $-\Gamma_{(t,\omega)} < \nu^*(t,\omega) < 1$; i.e., $\nu^*(t,\omega)$ is an interior minimum.

Since $\overline{u_1}(.)$ and $\overrightarrow{g}(.)$ are continuously differentiable functions by hypothesis (d) of Proposition 4.2,

$$\frac{\partial}{\partial \eta}[\overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) + \lambda \pi_\chi(t,\omega) \overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)]|_{\eta = \nu^*(t,\omega)} = 0.$$

Since the function inside the parentheses above is a strictly convex function of η, the first partial derivative above is strictly increasing as a function of η. Therefore, for $\nu_l < \nu^*(t,\omega) < \nu_h$, we have

$$\begin{aligned} &\frac{\partial}{\partial \eta}[\overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) + \lambda \pi_\chi(t,\omega) \overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)]|_{\eta = \nu_l} < 0 \\ &\quad < \frac{\partial}{\partial \eta}[\overline{u_1}(\lambda(1-\eta)\pi_\chi(t,\omega), t) + \lambda \pi_\chi(t,\omega) \overrightarrow{g}(\mu_0(t,\omega), \mu_1(t,\omega), \eta, t, \omega)]|_{\eta = \nu_h}. \end{aligned}$$

Since $\mu_0^n(t,\omega) \to \mu_0(t,\omega)$, $\mu_1^n(t,\omega) \to \mu_1(t,\omega)$, $\pi_{\chi^n}(t,\omega) \to \pi_\chi(t,\omega)$, and the first partial derivatives above are continuous functions of their arguments, it is easy to see that for n sufficiently large

$$\begin{aligned} &\frac{\partial}{\partial \eta}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega), t) + \lambda \pi_{\chi^n}(t,\omega) \overrightarrow{g}(\mu_0^n(t,\omega), \mu_1^n(t,\omega), \eta, t, \omega)] \,|_{\eta = \nu_l} < 0 \\ &\quad < \frac{\partial}{\partial \eta}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega), t) + \lambda \pi_{\chi^n}(t,\omega) \overrightarrow{g}(\mu_0^n(t,\omega), \mu_1^n(t,\omega), \eta, t, \omega)] \,|_{\eta = \nu_h}. \end{aligned}$$

It follows that

$$\nu_l < \nu^n(t,\omega) = \arg\min_{\eta \in D_{\mu_0^n(.),\mu_1^n(.)}(t,\omega)}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega),t) \\ + \lambda\pi_{\chi^n}(t,\omega)\overrightarrow{g}(\mu_0^n(t,\omega),\mu_1^n(t,\omega),\eta,t,\omega)] < \nu_h.$$

Since $\nu_h - \nu_l$ can be chosen to be arbitrarily small, we easily see that $\nu^n(t,\omega) \to \nu^*(t,\omega)$.

Case 2. $\nu^*(t,\omega) = -\Gamma_{(t,\omega)}$; i.e., $\nu^*(t,\omega)$ is an extreme point.

In this case, we have

$$\frac{\partial}{\partial\eta}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi}(t,\omega),t) + \lambda\pi_{\chi}(t,\omega)\overrightarrow{g}(\mu_0(t,\omega),\mu_1(t,\omega),\eta,t,\omega)]\mid_{\eta=\nu^*(t,\omega)} \geq 0.$$

Since the first partial derivative above is strictly increasing as a function of η, we have

$$\frac{\partial}{\partial\eta}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi}(t,\omega),t) + \lambda\pi_{\chi}(t,\omega)\overrightarrow{g}(\mu_0(t,\omega),\mu_1(t,\omega),\eta,t,\omega)]\mid_{\eta=\eta_h} > 0$$

for $\eta_h > \nu^*(t,\omega) = -\Gamma^*(t,\omega)$. It follows that for n sufficiently large

$$\frac{\partial}{\partial\eta}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega),t) + \lambda\pi_{\chi^n}(t,\omega)\overrightarrow{g}(\mu_0^n(t,\omega),\mu_1^n(t,\omega),\eta,t,\omega)]\mid_{\eta=\eta_h} > 0.$$

It follows that

$$-\Gamma(t,\omega) \leq \nu^n(t,\omega) = \arg\min_{\eta \in D_{\mu_0^n(.),\mu_1^n(.)}(t,\omega)}[\overline{u_1}(\lambda(1-\eta)\pi_{\chi^n}(t,\omega),t) \\ + \lambda\pi_{\chi^n}(t,\omega)\overrightarrow{g}(\mu_0^n(t,\omega),\mu_1^n(t,\omega),\eta,t,\omega)] < \eta_h.$$

Since η_h can be chosen arbitrarily close to $-\Gamma(t,\omega)$, it follows that

$$\nu^n(t,\omega) \to -\Gamma(t,\omega) = \nu^*(t,\omega).$$

This completes the proof of the lemma. □

Proof of Lemma 6.1. If $u_2(.,T)$ is monotonically decreasing, then the maximization problem on the right-hand side of the definition (4.2) of the function $\overline{u_2}$ is solved by $W = 0$ if $y \geq 0$. This implies (6.6). If $u_2(.,T)$ is nonmonotonic, Assumption 4 implies that, for $y \in [0,\infty)$, $f_2(y,T)$ must satisfy $u_2'(f_2(y,T),T) = y$. Since $u_2(.,T)$ is strictly increasing on $[0,\hat{c}]$, attains a maximum at $\hat{c}$ (so that its derivative is zero), and satisfies (6.2), it follows that (6.7) holds. □

Proof of Proposition 6.1. The proof proceeds along the lines of the proof of Proposition 4.1. Let the liquidation process be defined by $c^* = c_{\chi^*}$ and the terminal wealth level by $W^* = W_{\chi^*}$. The consumption process c^* satisfies (2.5) using exactly the same arguments used in Step 1 of the proof of Proposition 4.1. By (6.6), (6.7), and (6.9), the terminal wealth W^* is uniformly bounded so that it satisfies (2.5). The rest of the proof follows the proof of Proposition 4.1. □

Proof of Proposition 6.2. The proof proceeds exactly as the proof of Proposition 4.2 until Step 7, which establishes that condition (6.8) of the verification Proposition 6.1 is satisfied. By condition (e) of the proposition,

$$yl_2(y,T) \leq a + (1-b)u_2(l_2(y,T),T) \text{ for } y \geq 0.$$

Since $\lambda^*\pi_{\chi^*}(T) > 0$, we have

$$E\left[\int_0^T \pi_{\chi^*}(t)f_1(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt + \pi_{\chi^*}(T)l_2(\lambda^*\pi_{\chi^*}(T),T)\right] \\ \leq \frac{2a}{b\lambda^*} + \frac{1-b}{b\lambda^*}E\left[\int_0^T \overline{u_1}(\lambda^*(1-\nu^*(t))\pi_{\chi^*}(t),t)\,dt + \overline{u_2}(\lambda^*\pi_{\chi^*}(T),T)\right] < \infty,$$

since $|J(\lambda^*, \chi^*)| < \infty$ from assumption (c) of the proposition. The conditions $E[\int_0^T \pi_{\chi^*}(t)\, dt] < \infty$ and $E[\pi_{\chi^*}(T)] < \infty$ are required only to ensure that $E[\int_0^T u_1(c^*(t), t)^- dt] < \infty$ and $E[u_2(W^*, T)^-] < \infty$. These conditions are trivially satisfied since, by hypothesis (a) of the proposition and Lemma 6.1, $u_1(.)$ is nonnegative, $u_2(.)$ is nonnegative for $W \in [0, \hat{c}]$, and $W^* \in [0, \hat{c}]$. Therefore, by the result of Proposition 6.1, an optimal liquidation policy for the large investor exists. □

Acknowledgments. I thank Bill Morokoff (the associate editor), two anonymous referees, Robert Jarrow, and seminar participants at the 2001 INFORMS Applied Probability Society Conference (New York City) for valuable comments and suggestions.

REFERENCES

[1] R. Almgren and N. Chriss, *Optimal execution of portfolio transactions*, J. Risk, 3 (2000), pp. 5–39.

[2] K. Back, *Insider trading in continuous time*, Review of Financial Studies, 5 (1992), pp. 387–409.

[3] D. Bertsimas and A. Lo, *Optimal control of execution costs*, J. Financial Markets, 1 (1998), pp. 1–50.

[4] R. Buckdahn and Y. Hu, *Hedging contingent claims for a large investor in an incomplete market*, Adv. in Appl. Probab., 30 (1998), pp. 239–255.

[5] D. Cuoco and J. Cvitanic, *Optimal consumption choices for a large investor*, J. Econom. Dynam. Control, 22 (1998), pp. 401–436.

[6] J. Cvitanic and I. Karatzas, *Convex duality in constrained portfolio optimization*, Ann. Appl. Probab., 2 (1992), pp. 767–818.

[7] J. Cvitanic and J. Ma, *Hedging options for a large investor and forward-backward SDE's*, Ann. Appl. Probab., 3 (1996), pp. 652–681.

[8] D. Duffie and A. Ziegler, *Liquidation Risk*, Financial Analysts J., 59 (2003), pp. 42–51.

[9] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, Amsterdam, 1976.

[10] N. El Karoui and M. Jeanblanc, *Optimization of consumption with labor income*, Finance Stoch., 4 (1998), pp. 409–440.

[11] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–72.

[12] H. Follmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.

[13] D. Fudenberg, B. Holmstrom, and P. Milgrom, *Short-term contracts and long-term agency relationships*, J. Econom. Theory, 51 (1990), pp. 1–31.

[14] H. He and N. Pearson, *Consumption and portfolio policies with incomplete markets and short-selling constraints: The infinite-dimensional case*, J. Econom. Theory, 54 (1991), pp. 259–304.

[15] H. He and H. Mamaysky, *Dynamic trading policies with price impact*, J. Econom. Dynam. Control, 29 (2005), pp. 891–930.

[16] I. Karatzas and S. Kou, *On the pricing of contingent claims under constraints*, Ann. Appl. Probab., 6 (1996), pp. 321–369.

[17] I. Karatzas, J. P. Lehoczky, S. E. Shreve, and G.-L. Xu, *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim., 29 (1991), pp. 702–730.

[18] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[19] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1998.

[20] D. Kramkov and W. Schachermayer, *The asymptotic elasticity of utility functions and optimal investment in incomplete markets*, Ann. Appl. Probab., 9 (1999), pp. 904–950.

[21] A. Kyle, *Continuous auctions and insider trading*, Econometrica, 53 (1985), pp. 1315–1335.

[22] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, Berlin, 1999.

[23] A. Mas-Colell, M. Whinston, and J. Green, *Microeconomic Theory*, Oxford University Press, Oxford, UK, 1995.

[24] R. Merton, *Optimal consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413.

[25] M. Mnif and H. Pham, *Stochastic optimization under constraints*, Stochastic Process. Appl., 93 (2001), pp. 149–180.

[26] V. Polimenis, *A realistic model of market liquidity and depth*, J. Futures Markets, 25 (2005), pp. 443–464.

[27] M. Pritsker, *Large Investors: Implications for Equilibrium Asset Returns, Shock Absorption, and Liquidity*, working paper, Board of Governors of the Federal Reserve System, Washington, D.C., 2005.

[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1997.

[29] M. Schael, *A selection theorem for optimization problems*, Arch. Math., 25 (1974), pp. 219–224.

[30] A. Subramanian and R. Jarrow, *The liquidity discount*, Math. Finance, 11 (2001), pp. 447–474.

SIAM J. APPL. MATH.
Vol. 68, No. 4, pp. 1202

CORRIGENDUM: SUPPRESSION OF THE DIRICHLET EIGENVALUES OF A COATED BODY[*]

STEVE ROSENCRANS[†] AND XUEFENG WANG[†]

Abstract. In our paper [*SIAM J. Appl. Math.*, 66 (2006), pp. 1895–1916] there are several mistakes in signs in the statements of Theorems 3 and 6.

Key words. nanocomposite, Dirichlet eigenvalue, anisotropic heat conduction, thermal tensor, thermal management, insulation, reinforcement

AMS subject classifications. 35J05, 35J20, 80A20, 80M30, 80M40

DOI. 10.1137/070705404

In the upper bounds on $\lambda_1(\Omega)$ stated in Theorems 3 and 6, there should be "$-$" signs in front of π and $\overline{H}$.

On page 1905 all "$-$" signs in front of the principal curvatures k_1 and k_2 should be changed to "$+$" signs. This leads to some obvious minor changes in the rest of the proof of Theorem 3 and the statement and proof of Lemma 5. The speed of the curves of principal curvature are 1 only at q, and so the six equations following "speed equal to 1" hold only at q.

These corrections necessitate *only* the above-mentioned changes in the statements of Theorems 3 and 6. The statements of all the other theorems remain unchanged.

[*] Received by the editors October 15, 2007; accepted for publication October 22, 2007; published electronically March 28, 2008.
http://www.siam.org/journals/siap/68-4/70540.html

[†] Mathematics Department, Tulane University, New Orleans, LA 70118 (srosenc@tulane.edu, xdw@math.tulane.edu).

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1203–1222

SOLUTE TRANSPORT IN POROUS MEDIA. MEDIA WITH CAPILLARIES AS VOIDS[*]

GUILLERMO H. GOLDSZTEIN[†]

Abstract. We study solute transport in porous media with periodic microstructures consisting of interconnected thin channels. We discuss a local physical mechanism that occurs at the intersections of channels and promotes mixing of the solute with the solvent (i.e., the host liquid). We identify the parameter regime, where this mechanism is the dominant cause of dispersion, and obtain the effective or macroscopic transport equation that the concentration of solute satisfies when the medium is subjected to a time periodic applied pressure gradient. We conclude with illustrative examples.

Key words. porous media, solute transport, hydrodynamic dispersion, network approximation, macroscopic behavior, homogenization

AMS subject classifications. 76S05, 76M45, 76M50, 76R05, 76R50

DOI. 10.1137/070695228

1. Introduction.

A porous medium is a material that contains relatively small spaces filled with fluid (such as a gas, a liquid, or a mixture of different fluids) embedded in a solid matrix. These fluid filled spaces are called pores or voids. With the exception of metals, some dense rocks, and some plastics, virtually all solid materials are porous to varying degrees.

Solutes are materials that dissolve in liquids forming solutions. An example is salt (not at very large concentrations) in water. The host liquid, such as water in the mentioned example, is called the solvent. The transport of a solute in porous media depends on several factors, including the solvent and solute properties, the fluid velocity field within the porous medium, and the *microgeometry*, i.e., shape, size, and location of the solid part of the medium and the voids. The objective of this paper is to provide new tools for the study of the influence of these factors on solute transport.

Solute transport in liquid filled porous media plays a significant role in several phenomena of scientific and technological importance including the transport of contaminants in soils [17, 32], the transport of nutrients in bones [50, 45, 43, 44, 65, 39], the intrusion of salt in fresh water in soils near ocean coasts, movement of minerals (e.g., fertilizers) in soils, secondary recovery techniques in oil reservoirs (where the injected fluid dissolves the reservoir's oil), the use of tracers in petroleum engineering and hydrology research projects, etc. (see more about these and other examples in [14, 6, 9, 19, 30, 58]).

Several theoretical methods are used to study solute transport in porous media [23]. These include the use of numerical experiments on networks of channels with varying widths forming regular grids [2, 15, 16, 20, 27, 58, 59], percolation methods [2, 7, 8, 10, 47, 55, 56, 57, 59, 62], numerical experiments on media with fractal geometry [2, 16, 59, 64], assuming periodic media and calculating the effective transport equation by means of the method of moments [2, 11, 12, 13, 14, 31]

[*] Received by the editors June 25, 2007; accepted for publication (in revised form) December 11, 2007; published electronically April 2, 2008. This research was supported by the NSF.
http://www.siam.org/journals/siap/68-5/69522.html

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (ggold@math.gatech.edu).

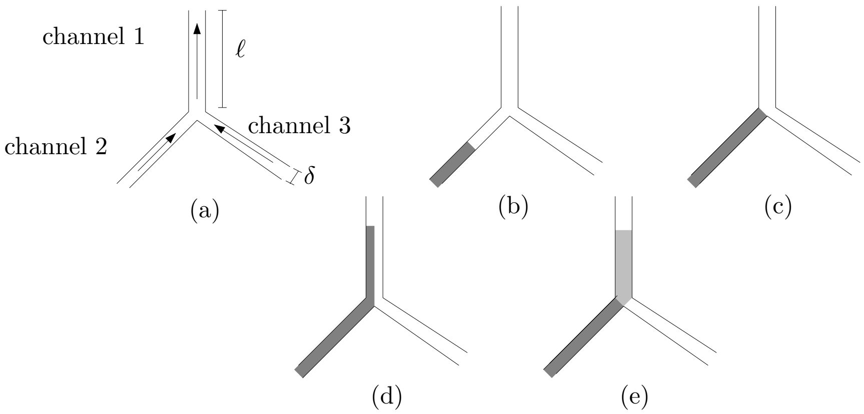


FIG. 1.1. (a) *Direction of the fluid velocity.* (b) *Initial distribution of solute concentration.* (c) *Solute concentration at the time when solute first reaches the intersection.* (d) *Solute concentration after solute reaches channel* 1 *if diffusion does not homogenize the solute concentration in slices perpendicular to the channels.* (e) *Solute concentration after solute reaches channel* 1 *if the channels are thin enough that diffusion homogenizes the solute concentration in slices perpendicular to the channels. We work in the regime of* (e).

or homogenization techniques [42, 48, 49, 51], and the calculation of effective transport equations by means of the method of moments on periodic networks of channels [2, 1, 29] (see also [38] for a study of diffusion in periodic networks with no flow). The most well-known early theoretical works are studies of solute transport in single straight tubes [63, 3]. We also mention the work on random networks of thin channels [52, 53, 54], the work on media with trapped fluid in dead-end pores [21], the early work using the method of moments [41], and the work on solute transport in a dilute suspension of spheres [46] and in parallel channels [18]. Further discussions can be found in [10, 36, 60, 5, 24, 25, 9, 26]. Experimentally, these phenomena have also been extensively studied (see summaries in [37, 30, 22]; see also [40, 28]).

The local phenomenon that motivates our work is simple and described next. Consider the three interconnected channels of Figure 1.1(a). We labeled the channels 1, 2, and 3. The arrows indicate the direction of the fluid velocity field within each channel. The channels are thin. More precisely, assume that the lengths of the channels are $O(\ell)$, their diameters are $O(\delta)$, the fluid velocities within the channels are $O(v)$, and these parameters satisfy $\delta^2/D \ll \ell/v \ll \ell^2/D$, where D is the diffusion coefficient of the solute in the solvent under consideration. In this parameter regime, the concentration of solute in each channel is homogeneous in slices perpendicular to the channel and is convected with the average fluid velocity within the channel (see [63, 3] and our section 2.1).

In Figure 1.1(b) we display the initial solute concentration, i.e., at time $t = 0$. The darker the regions, the larger the solute concentration. Only channel 2 has solute at $t = 0$. Let t_1 be the time when solute first reaches the intersection of the channels. The solute concentration at time $t = t_1$ is shown in Figure 1.1(c). Let $t_2 > t_1$. In the absence of diffusion, fluid from channel 2 with solute and fluid from channel 3 without solute would be convected next to each other along channel 1 during the time interval (t_1, t_2), and the distribution of solute concentration at $t = t_2$ would look as displayed in Figure 1.1(d). This is in contradiction with the fact that solute concentration is homogeneous in slices perpendicular to the channels. In fact, as discussed in section 2.1, diffusion homogenizes the concentration of solute in slices perpendicular to the channels, and thus the distribution of solute at $t = t_2$ is as illustrated in Figure 1.1(e). During the time interval (t_1, t_2), liquid with solute from channel 2 enters channel 1 and mixes with liquid without solute that enters channel 1 from channel 3.

In this paper we consider porous media with periodic microstructures and void spaces consisting of interconnected thin channels. The local effect described in the above paragraph (that corresponds to Figure 1.1(e)) occurs throughout the porous media and promotes solute transport. In this paper we study this phenomenon.

In section 2 we describe our mathematical model. This model is the asymptotic limit of the Navier–Stokes equations within the void with nonslip boundary condition coupled with the convection-diffusion equation for the transport of solute. In section 3 we obtain the *macroscopic* transport equation which the solute concentration satisfies. We assume that the medium is subjected to a time periodic applied pressure gradient and obtain, by means of homogenization techniques on the model of section 2, that the solute concentration satisfies a macroscopic convection-diffusion equation. As expected, it is convected with the average fluid velocity. We obtain a relatively simple mean to compute the diffusion tensor, known in the literature as the dispersion tensor. In section 4 we provide some examples and in section 5 conclude with some discussions.

As previously mentioned, there are several methods for studying solute transport in porous media. Each method has its strengths and weaknesses. The most computational economical methods are those that compute the macroscopic properties with the use of periodic networks. This class of methods is essentially limited to [1] and its generalizations [2, 29]. The authors of [1, 2, 29] use the method of moments instead of homogenization or asymptotic techniques, as we do here. However, this is not the essential difference between those methods and the technique developed in this paper. The models in [1, 2, 29] use ad hoc rules that correspond to assuming that the volume of the channels is much smaller than the volume of the intersections, and some ad hoc mixing rules are given at the intersections. As a consequence, the physical effect that motivated our work (that of Figure 1.1(e)) is not captured well by the existing models [1, 2, 29] (see also our section 4). We believe our method is an ideal tool for studying the dependence of the dispersion tensor on the microgeometry and will prove to be very useful.

2. Mathematical model.

2.1. Preliminaries. Fluid flow and solute transport in channels.

Figure 2.1 shows a two-dimensional channel with length ℓ and width δ filled with a Newtonian incompressible fluid that is subjected to pressures $p = p_a$ and $p = p_b$ at the ends of the channel. Let $\hat{\mathbf{e}}$ be the vector of unit length parallel to the channel displayed in Figure 2.1. Let y be the coordinate in the direction perpendicular to the channel. At low Reynolds numbers (low velocities), the fluid velocity of the steady state flow is of the form $u(y)\hat{\mathbf{e}}$ with u satisfying $(p_b - p_a)/\ell = \mu u''$, where μ is the fluid viscosity and u'' is the second derivative of u. In addition, the fluid velocity satisfies nonslip boundary conditions at the channel walls, i.e., $u = 0$ at the walls. Simple calculations show that the velocity has a parabolic profile (see Figure 2.1) and its spatial average across the channel is

$$\mathbf{v} = \frac{\delta^2}{12\mu\ell}(p_a - p_b)\hat{\mathbf{e}} \tag{2.1}$$

(see [4]). This type of flow is known as Poiseuille flow.

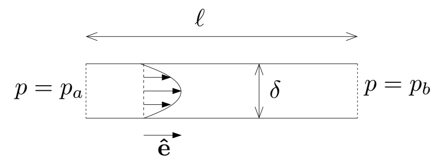


FIG. 2.1. *Velocity profile of a Poseuille flow within a straight channel (indicated by arrows).*

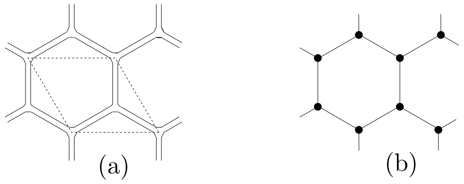


FIG. 2.2. (a) *Example of a periodic network of interconnected channels. The period cell is shown by dashed lines.* (b) *Associated graph. The lines are the edges and the solid circles the nodes.*

Taylor studied solute transport in channels at low Reynolds numbers [63]; see also Aris [3]. The result relevant to us is the following. Let D be the coefficient of diffusion of the solute in the host liquid, ℓ the length of the channel, δ its diameter, and v the spatial average of the norm of the fluid velocity. If

$$\frac{\delta^2}{D} \ll \frac{\ell}{v} \ll \frac{\ell^2}{D}, \tag{2.2}$$

the evolution of solute concentration is described by these two rules:

(2.3)
Rule 1: The concentration of solute is homogeneous in slices (of infinitesimal thickness) perpendicular to the channel.
Rule 2: The solute concentration is convected (or advected) with the average fluid velocity within the channel.

The validity of these two rules can be easily understood as follows. The time required by diffusion to homogenize the solute concentration in slices perpendicular to the channel is of order $O(\delta^2/D)$. Since the time required for solute to be convected from one end to the opposite end of the channel is $O(\ell/v)$, the validity of Rule 1 results from $\delta^2/D \ll \ell/v$. On the other hand, Taylor showed that, in the direction of the tube, the time for solute to disperse distances of $O(\ell)$ is $O(\ell^2/D^\star)$, where $D^\star = O(D + v^2\delta^2/D)$. Thus, the validity of Rule 2 results if $\ell/v \ll \ell^2/D^\star$. Simple algebra shows that, in fact, the two conditions $\delta^2/D \ll \ell/v$ and $\ell/v \ll \ell^2/D^\star$ are equivalent to (2.2).

2.2. Microgeometry. We consider two-dimensional porous media with periodic microstructures. We denote the void or pore space (i.e., the space filled by fluid) by Ω_p. Note that $\Omega_p \subseteq \mathbb{R}^2$. Since the microstructures are periodic, there exist two linearly independent vectors $\mathbf{w}$ and $\mathbf{q}$ such that

$$\Omega_p = \Omega_p + n\mathbf{w} + m\mathbf{q} \tag{2.4}$$

for all pairs of integers n and m. We assume that Ω_p is a collection interconnected thin channels (see Figure 2.2(a)). We assume that exactly three channels merge at each intersection. We associate a periodic graph with the microstructure of the medium in a natural way, as illustrated in Figure 2.2(b); the edges are the channels and the nodes the intersection of channels. We denote by $\mathcal{N}$ the set of nodes. We identify the nodes with their location, and thus $\mathcal{N} \subset \mathbb{R}^2$. We denote by $\mathcal{E}$ the set of edges. Given an edge e, its width (i.e., the width of the channel that corresponds to e) is denoted by δ_e and its length by ℓ_e. We assume that the widths of the channels are much smaller than their lengths. We also assume that the void space Ω_p is a connected set.

2.3. Fluid flow. Microscopic description. The fluid that fills Ω_p is an incompressible Newtonian fluid with constant density ρ and constant viscosity μ and satisfies nonslip boundary conditions, i.e., the fluid velocity vanishes at the channels walls (i.e., at the boundary of Ω_p).

For each node $\mathbf{a} \in \mathcal{N}$, we denote by $p_{\mathbf{a}}$ the pressure at $\mathbf{a}$. Note that $p_{\mathbf{a}} = p_{\mathbf{a}}(t)$ is a function of time t. We assume that the medium is subjected to an applied pressure gradient $\mathbf{G} = \mathbf{G}(t)$ that is periodic in t with period t_0. Thus, the pressures at the nodes satisfy the condition

$$p_{\mathbf{a}+n\mathbf{w}+m\mathbf{q}} = p_{\mathbf{a}} + \mathbf{G} \cdot (n\mathbf{w} + m\mathbf{q}) \tag{2.5}$$

for all integers n and m and all nodes $\mathbf{a}$, where, as described above, $\mathbf{w}$ and $\mathbf{q}$ are the vectors that determine the periodicity of the microgeometry, and we use the notation $\mathbf{r} \cdot \mathbf{s} = r_1 s_1 + r_2 s_2$ for all vectors $\mathbf{r}, \mathbf{s}$, and r_i is the ith component of the vector $\mathbf{r}$.

If e is an edge, we denote by $\mathbf{v}_e$ the average of the velocity field within the channel e. We assume that the variation of $\mathbf{G}(t)$ in time is slow enough that the pressure difference between the two ends of a channel creates a Poiseuille flow within that channel, and thus, for each edge e, according to our review (equation (2.1)), we have

$$\mathbf{v}_e = -\frac{\delta_e^2}{12\mu\ell_e}\,(p_{\mathbf{b}} - p_{\mathbf{a}})\,\frac{\mathbf{b}-\mathbf{a}}{\|\mathbf{b}-\mathbf{a}\|}, \text{ where } \mathbf{a} \text{ and } \mathbf{b} \text{ are the endpoints of } e, \tag{2.6}$$

and we use the standard notation for the Euclidean norm $\|\mathbf{r}\| = \sqrt{r_1^2 + r_2^2}$.

The rate at which the volume of fluid enters an intersection is equal to the rate at which it leaves the intersection, i.e., conservation of mass. This implies that, for each node $\mathbf{a}$, we have

$$\sum_{\{e \in \mathcal{E} : \mathbf{a} \text{ is an endpoint of } e\}} \delta_e \mathbf{v}_e \cdot \frac{\mathbf{b}-\mathbf{a}}{\|\mathbf{b}-\mathbf{a}\|} = 0, \text{ where } \mathbf{b} \text{ is the endpoint of } e \text{ not equal to } \mathbf{a}. \tag{2.7}$$

The velocities within all the channels are uniquely determined by the system (2.5)–(2.7). This well-known system (similar models were used as early as [33, 34, 35]; see also [61, 30]) is the asymptotics of the Navier–Stokes equations within the void with nonslip boundary conditions in the limit when the widths of the channels are much smaller than their lengths, and the time variations of the applied pressure gradient $\mathbf{G}(t)$ are slow enough. Note that the resultant velocity field is periodic in space with the same period as the microstructure.

In practice, we first solve for the pressure at the nodes and then for the velocities within the channels. More precisely, using the expression for the velocities in (2.6), we reduce (2.7) into

$$\sum_{\{\mathbf{b} \in \mathcal{N} : \mathbf{b} \text{ is connected to } \mathbf{a} \text{ by an edge}\}} \frac{\delta_e^3}{12\mu\ell_e}\,(p_{\mathbf{b}} - p_{\mathbf{a}}) = 0 \text{ for all } \mathbf{a} \in \mathcal{N}, \tag{2.8}$$

which together with condition (2.5) reduce, for each fixed t, to a system of linear equations, where the number of unknowns is equal to the number of nodes in a single period cell minus one. Once the pressure at the nodes is obtained, the velocities in the edges are easily computed with (2.6).

2.4. Solute transport. Microscopic description. For each $e \in \mathcal{E}$ we use the notation $v_e = \|\mathbf{v}_e\|$. We assume that

$$v_e = O(v),\ \ell_e = O(\ell),\ \delta_e = O(\delta) \text{ for all } e \in \mathcal{E}, \tag{2.9}$$

where v, ℓ, and δ are parameters that satisfy (2.2), and thus the transport of solute concentration within each channel is given by Rules 1 and 2 (see (2.3)).

Given an edge e, its endpoint with smallest pressure will be called its head and will be denoted by $\mathbf{h}(e)$. Analogously, $\mathbf{k}(e)$, the tail of the edge e, is the endpoint of e with largest pressure. Thus, fluid within an edge e (or channel) flows from its tail $\mathbf{k}(e)$ to its head $\mathbf{h}(e)$. Note that, since the fluid flow is time dependent, an endpoint of an edge may be its head for some period of time and its tail for other times.

We parametrize each edge e (more precisely, the segment joining the tail and head of e) by

$$\mathbf{x}_e(s) = \mathbf{k}(e) + s\frac{\mathbf{h}(e) - \mathbf{k}(e)}{\|\mathbf{h}(e) - \mathbf{k}(e)\|} \tag{2.10}$$

and we denote by $u_e(s,t)$ the solute concentration in the channel e at the point $\mathbf{x}_e(s)$ and time t. Note that $\mathbf{x}_e(0) = \mathbf{k}(e)$ and $\mathbf{x}_e(\ell_e) = \mathbf{h}(e)$ because $\ell_e = \|\mathbf{h}(e) - \mathbf{k}(e)\|$. Thus, the channel is parametrized by $\mathbf{x}_e(s)$ with $0 \leq s \leq \ell_e$. The fact that solute concentration in a channel is convected with the average fluid velocity within the channel translates into

$$\frac{\partial u_e}{\partial t} + v_e\frac{\partial u_e}{\partial s} = 0 \text{ for } 0 \leq s \leq \ell_e,\ t \geq 0, \text{ and all } e \in \mathcal{E}. \tag{2.11}$$

Let e be an edge and $\mathbf{k}(e)$ its tail (at a fixed time t). One of two cases is possible: $\mathbf{k}(e)$ is the head of two other edges, or $\mathbf{k}(e)$ is the head of only one other edge. Assume first that $\mathbf{k}(e)$ is the head of two other edges, say, β_1 and β_2, i.e., $\mathbf{h}(\beta_1) = \mathbf{h}(\beta_2) = \mathbf{k}(e)$. Conservation of solute implies that solute enters $\mathbf{k}(e)$ at the same rate that it leaves $\mathbf{k}(e)$, and thus $\delta_e v_e u_e(0,t) = \delta_{\beta_1} v_{\beta_1} u_{\beta_1}(\ell_{\beta_1}, t) + \delta_{\beta_2} v_{\beta_2} u_{\beta_2}(\ell_{\beta_2}, t)$. This condition can be written as

$$u_e(0,t) = \frac{\sum_{\{\beta:\mathbf{h}(\beta)=\mathbf{k}(e)\}} \delta_\beta v_\beta u_\beta(\ell_\beta, t)}{\sum_{\{\beta:\mathbf{h}(\beta)=\mathbf{k}(e)\}} \delta_\beta v_\beta} \tag{2.12}$$

once we note that (2.7) at $\mathbf{k}(e)$ is $\delta_e v_e = \delta_{\beta_1} v_{\beta_1} + \delta_{\beta_2} v_{\beta_2}$. We have just shown that (2.12) is valid for edges e for which its tail $\mathbf{k}(e)$ is the head of two other edges. We next show that, in fact, (2.12) is valid for all edges e. To that end, assume now that $\mathbf{k}(e)$ is the head of only one edge, say β. In other words, fluid flows into $\mathbf{k}(e)$ from only channel β. Thus, the concentration of solute going into e should be equal to the concentration of solute entering $\mathbf{k}(e)$ from β, i.e., $u_e(0,t) = u_\beta(\ell_\beta, t)$. This condition is, in fact, (2.12) in this case, i.e., when $\mathbf{k}(e)$ is the head of only one edge.

The system (2.11)–(2.12) uniquely determines the time evolution of the solute concentration within the channels once initial conditions and appropriate boundary conditions are given. We mention that the system (2.11)–(2.12) is not ad hoc; it is the asymptotic limit of the convection-diffusion equation for the transport of solute within the network in the parameter regime in which we work (i.e., (2.9) and (2.2)).

3. Macroscopic transport equation. We say that two edges are equivalent if one is the translation of the other by a vector of the form $n\mathbf{w} + m\mathbf{q}$, where $\mathbf{w}$ and $\mathbf{q}$ are the vectors that determine the periodicity of the microstructure (see (2.4)) and n and m are integers. Thus, two edges e_1 and e_2 are equivalent if there exist n and m integers such that $\mathbf{h}(e_2) = \mathbf{h}(e_1) + n\mathbf{w} + m\mathbf{q}$ and $\mathbf{k}(e_2) = \mathbf{k}(e_1) + n\mathbf{w} + m\mathbf{q}$ (we recall that $\mathbf{h}(e)$ denotes the head of the edge e and $\mathbf{k}(e)$ denotes its tail). This defines an equivalence relation in the set of edges. Note that the widths, lengths, and velocities of equivalent edges are equal, i.e., $\delta_{e_1} = \delta_{e_2}$, $\ell_{e_1} = \ell_{e_2}$, and $\mathbf{v}_{e_1} = \mathbf{v}_{e_2}$ if e_1 and e_2 are equivalent. In what follows we will take spatial average of quantities. Thus, we need to be able to select exactly one edge per equivalence class. We denote by $\mathcal{F}$ a set of edges that contains exactly one edge per equivalent class. For example, $\mathcal{F}$ could be all the edges whose heads are in the period cell

$$Q = \{s\mathbf{w} + r\mathbf{q} : 0 \leq s, r < 1\} \tag{3.1}$$

at a certain time.

We first observe that the area occupied by fluid within the period cell Q (i.e., the area of $\Omega_p \cap Q$) is

$$|\Omega_p \cap Q| = \sum_{e\in\mathcal{F}} \delta_e \ell_e. \tag{3.2}$$

We denote by $\mathbf{V}$ the spatial average fluid velocity, i.e.,

$$\mathbf{V} = \frac{\sum_{e\in\mathcal{F}} \delta_e \ell_e \mathbf{v}_e}{\sum_{e\in\mathcal{F}} \delta_e \ell_e} = \frac{\sum_{e\in\mathcal{F}} \delta_e \ell_e \mathbf{v}_e}{|\Omega_p \cap Q|}. \tag{3.3}$$

Note that assumption (2.9) implies that $\|\mathbf{V}\| = O(v)$. Assume that t_0, the period of the applied pressure gradient $\mathbf{G}$, satisfies

$$t_0 \gg \frac{\ell}{v} \qquad \text{(more precisely } t_0 \gg \max \ell_e/v_e \text{ most of the time);} \tag{3.4}$$

i.e., the time required for solute concentration to be convected across a channel is much smaller than the period of the applied pressure gradient. In Appendix A we show that, *macroscopically*, the solute concentration is convected with the average fluid velocity $\mathbf{V}$ and dispersed with dispersion tensor

$$D_{ij}^{\text{eff}} = \frac{1}{t_0} \int_0^{t_0} D_{ij}^\star(t) \mathrm{d}t, \tag{3.5}$$

where

$$D_{ij}^\star = \frac{1}{2|\Omega_p \cap Q|} \left\{ \sum_{e\in\mathcal{F}} \delta_e \ell_e \left(\frac{\ell_e}{v_e} \left[\mathbf{v}_e - \mathbf{V}\right]_i \left[\mathbf{v}_e - \mathbf{V}\right]_j \right.\right.$$
$$\left.\left. + \left[\mathbf{V} - \mathbf{v}_e\right]_i \left[\mathbf{f}_{\mathbf{k}(e)}\right]_j + \left[\mathbf{V} - \mathbf{v}_e\right]_j \left[\mathbf{f}_{\mathbf{k}(e)}\right]_i \right) \right\}, \tag{3.6}$$

$[\mathbf{y}]_i$ denotes the ith component of the vector $\mathbf{y}$, and the family of vectors $(\mathbf{f_a})_{\mathbf{a}\in\mathcal{N}}$ is a solution periodic in space and time (i.e., $\mathbf{f_a}(t) = \mathbf{f}_{\mathbf{a}+n\mathbf{w}+m\mathbf{q}}(t + pt_0)$ for all integers n, m, and p) of the following system:

$$\sum_{\{e:\mathbf{h}(e)=\mathbf{a}\}} \delta_e v_e (\mathbf{f}_{\mathbf{k}(e)} - \mathbf{f_a}) = \sum_{\{e:\mathbf{h}(e)=\mathbf{a}\}} \delta_e \ell_e \left(\mathbf{v}_e - \mathbf{V}\right) \text{ for all } \mathbf{a} \in \mathcal{N}. \tag{3.7}$$

More precisely, for each $\mathbf{a} \in \mathcal{N}$, let $u_{\mathbf{a}}(t)$ be the solute concentration that leaves the intersection $\mathbf{a}$ at time t, i.e.,

$$u_{\mathbf{a}}(t) = u_e(0,t) \text{ if } \mathbf{a} = \mathbf{k}(e) \text{ at time } t. \tag{3.8}$$

Note that $u_{\mathbf{a}}(t)$ is well defined because $u_{e_1}(0,t) = u_{e_2}(0,t)$ if e_1 and e_2 are two edges that have the same tail at time t, i.e., $\mathbf{k}(e_1) = \mathbf{k}(e_2)$. In Appendix A we show that

$$u_{\mathbf{a}}(t) \simeq u\left(\mathbf{a},t\right) \text{ for } t = O\left(t_0^2 \frac{v}{\ell}\right), \tag{3.9}$$

where $u(\mathbf{x},t)$ satisfies

$$\frac{\partial u}{\partial t} + \mathbf{V} \cdot \nabla u = \sum_{i,j} D_{ij}^{\text{eff}} \frac{\partial^2 u}{\partial x_i \partial x_j}, \tag{3.10}$$

where ∇u is the gradient of u with respect to $\mathbf{x}$ and u is subjected to appropriate boundary and initial conditions that depend on the particular problem under consideration. We note that $\mathbf{D}^{\text{eff}}$ is usually referred to as the dispersion tensor.

4. Examples and observations.

4.1. Constant applied pressure gradient. As a first general example, we consider the case when the applied pressure gradient $\mathbf{G}$ is time independent. In this case, the system for the pressure at the nodes (2.5) and (2.8) is time independent and so are the velocities within the channels (see (2.6)). The spatially periodic family of vectors $(\mathbf{f_a})_{\mathbf{a}\in\mathcal{N}}$, solution of system (3.7), is also time independent, and the expression for the dispersion tensor simplifies to

$$\begin{aligned} D_{ij}^{\text{eff}} = D_{ij}^{\star} = \frac{1}{2\sum_{e\in\mathcal{F}} \delta_e \ell_e} \sum_{e\in\mathcal{F}} \delta_e \ell_e \Bigg(& \frac{\ell_e}{v_e} \left[\mathbf{v}_e - \mathbf{V}\right]_i \left[\mathbf{v}_e - \mathbf{V}\right]_j \\ & + \left[\mathbf{V} - \mathbf{V}_e\right]_i \left[\mathbf{f}_{\mathbf{k}(e)}\right]_j + \left[\mathbf{V} - \mathbf{v}_e\right]_j \left[\mathbf{f}_{\mathbf{k}(e)}\right]_i \Bigg). \end{aligned} \tag{4.1}$$

4.2. Applied pressure gradient of the form $\mathbf{G}(t) = g(t)\mathbf{E}$ with $\mathbf{E}$ constant. As a second general example, we consider the case when the applied pressure gradient $\mathbf{G}$ is of the form $\mathbf{G}(t) = g(t)\mathbf{E}$ with $\mathbf{E}$ constant and $g(t)$ a real valued periodic function with period t_0. The evaluation of the dispersion tensor is also simple in this case. Let $\mathbf{D}_{\mathbf{E}}^{\text{eff}}$ be the dispersion tensor that corresponds to the applied pressure gradient $\mathbf{E}$. Then, the dispersion tensor that corresponds to the applied pressure gradient $\mathbf{G}(t) = g(t)\mathbf{E}$ is

$$\mathbf{D}^{\text{eff}} = \mathbf{D}_{\mathbf{E}}^{\text{eff}} \frac{1}{t_0} \int_0^{t_0} |g(t)| \mathrm{d}t. \tag{4.2}$$

The validity of the above equation results from simple calculation. Briefly, we first note that, if $\mathbf{v}_e^{\mathbf{E}}$ are the velocities within the channels when the applied pressure gradient is $\mathbf{E}$, then $\mathbf{v}_e = g(t)\mathbf{v}_e^{\mathbf{E}}$ are the velocities within the channels when the applied pressure gradient is $g(t)\mathbf{E}$. As a consequence, if the vectors $\mathbf{f}_a^{\mathbf{E}}$ solve system (3.7) when the applied pressure gradient is $\mathbf{E}$, then $\mathbf{f}_a = g(t)\mathbf{f}_a^{\mathbf{E}}$ is a solution of system (3.7) when the applied pressure gradient is $g(t)\mathbf{E}$. Thus, if $\mathbf{D}_{\mathbf{E}}^{\star}$ is the tensor of (3.6) when the applied pressure gradient is $\mathbf{E}$, then $\mathbf{D}^{\star} = |g(t)|\mathbf{D}_{\mathbf{E}}^{\star}$ is the tensor of (3.6) when the applied pressure gradient is $g(t)\mathbf{E}$. Note that $\mathbf{D}_{\mathbf{E}}^{\star}$ is time independent, and thus $\mathbf{D}_{\mathbf{E}}^{\text{eff}} = \mathbf{D}_{\mathbf{E}}^{\star}$ (see (3.6)). Finally, (4.2) results from (3.6).

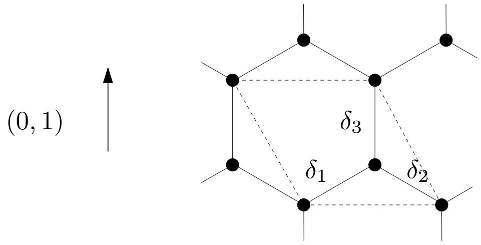


FIG. 4.1. *Graph corresponding to the microgeometry of our example. The period cell is enclosed by dashed lines. The widths of the channels are δ_1, δ_2, and δ_3.*

4.3. A concrete example. The graph that corresponds to the microgeometry of our example is shown in Figure 4.1. All the channels have the same length ℓ and form regular hexagons. The period cell is enclosed by dashed lines. The width of the channels in the period cell are, as displayed in the figure, δ_1, δ_2, and δ_3. We assume the applied pressure gradient to be of the form

$$\mathbf{G}(\mathbf{t}) = \left(0, \frac{g(t)}{\ell}\right), \tag{4.3}$$

where g is a periodic function with period t_0 and $(0,1)$ is the unit vector that points in the vertical direction (see Figure 4.1). Some algebra shows that, in this example, the use of our method leads to

$$\mathbf{V}_2(t) = -\frac{3}{16\mu\ell}\frac{\delta_3^3(\delta_1^3+\delta_2^3)}{(\delta_1^3+\delta_2^3+\delta_3^3)(\delta_1+\delta_2+\delta_3)}g(t) \tag{4.4}$$

and

$$\mathbf{D}_{22}^{\text{eff}} = \frac{9}{64\mu}\frac{\delta_3^3(\delta_1^3+\delta_2^3)(\delta_1+\delta_2)^2(\delta_1-\delta_2)^2}{\delta_1\delta_2(\delta_1^3+\delta_2^3+\delta_3^3)(\delta_1+\delta_2+\delta_3)^3}\frac{1}{t_0}\int_0^{t_0}|g(t)|\mathrm{d}t. \tag{4.5}$$

To discuss the above formulas in a more concrete context, assume that the material occupies the region $x_2 > 0$. Also assume that the material is attached to a reservoir of solute located at $x_2 < 0$ and that initially there is no solute within the material (for $x_2 > 0$). Due to symmetry, the solute concentration u, solution of (3.10), in this example depends only on x_2. Thus, we need only $\mathbf{V}_2$ and $\mathbf{D}_{22}^{\text{eff}}$, which are given by (4.4) and (4.5), respectively.

As a first observation, note that $\mathbf{D}_{22}^{\text{eff}} = 0$ if $\delta_1 = \delta_2$. Thus, after each period, solute is convected a distance $\int_0^{t_0}\mathbf{V}(t)\,\mathrm{d}t$ but is not dispersed in our asymptotic limit; there is a smaller order effective dispersion that results from an effect known as Taylor dispersion inside the channels [63, 3]. Note that this is in accordance with the physical effect described in the introduction as shown in Figure 1.1(e). The mixing of solute with the host liquid occurs when solute from two different channels and at different concentrations flows into the same intersection (in Figure 1.1(e) one of the channels had zero solute concentration). Due to symmetry in our example when $\delta_1 = \delta_2$, whenever solute from two channels flows into the same intersection, the concentration in both channels is the same. This is illustrated in Figure 4.2(a), where we show that solute reaches the upper ends of all the channels attached to the reservoir at the same time if $\delta_1 = \delta_2$.

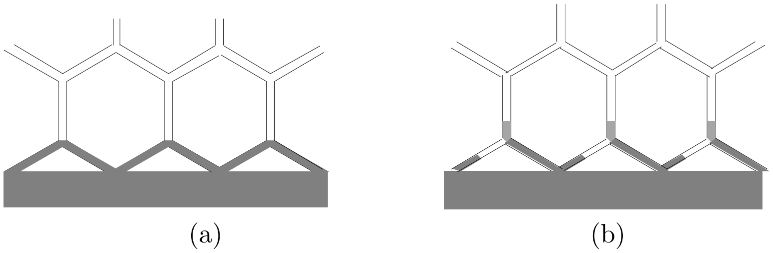


Fig. 4.2. *Microgeometry that corresponds to the graph of Figure* 4.1. *In* (a), $\delta_1 = \delta_2$. *In* (b), $\delta_1 < \delta_2$. *The shaded areas represent the solute concentration. The darker the shade, the larger the solute concentration.*

On the other hand, in Figure 4.2(b) we display an example where $\delta_1 < \delta_2$. As illustrated in that figure, the time required for solute from the reservoir to travel through the thinner channels is longer than the travel time through the thicker channels. Thus, the effect illustrated in Figure 1.1(e) does occur and, as (4.5) implies, we have that $D^{\text{eff}}_{22} \neq 0$. Note in particular that, in the case $\int_0^{t_0} g(t)\,\mathrm{d}t = 0$, there is no convection after a complete period. Thus, there will be much more transport of solute in our second example, where $D^{\text{eff}}_{22} \neq 0$, than in the example of the previous paragraph, where $D^{\text{eff}}_{22} = 0$.

At first glance, (4.5) seems to lead to the following contradiction. On one hand, (4.5) shows that $D^{\text{eff}}_{22} \to \infty$ as $\delta_1 \to 0$ while keeping δ_2 and δ_3 fixed. However, $\delta_1 = 0$ means not having the channels with width δ_1, and thus not having intersection of three channels. According to our discussions, we would expect $D^{\text{eff}}_{22} = 0$ in this case. This apparent contradiction is resolved by (3.4) and (3.9) which state that (4.5) is valid for $t = O(t_0^2 v/\ell)$ and $t_0 \gg \ell/v_1$, and thus we need $t \gg \ell/v_1$. Finally, we note that, since $v_1 \to 0$ as $\delta_1 \to 0$, this is a singular limit, which resolves this apparent contradiction. In other words, the smaller δ_1, the longer we have to wait for dispersion to occur and for our asymptotics to be valid. As $\delta_1 \to 0$, we would have to wait an infinitely long time.

5. Discussion. As mentioned in the introduction, there are several methods for studying solute transport in porous media. For their computational efficiency and their flexibility in modeling microstructures, methods that compute the macroscopic properties with the use of periodic networks are very useful. So far, this class of methods is essentially limited to [1] and its generalizations [2, 29]. Moreover, as mentioned in the introduction, the models in [1, 2, 29] use ad hoc rules that prevent them from accurately modeling the physical effect that motivated the present work, i.e., that of Figure 1.1(e).

Thus, while a large body of work exists in solute transport in porous media, the work introduced here is new and, we believe, will prove powerful in providing new understanding of the dependence of solute transport on the microgeometry. The strengths of our method include the following: (1) This method is exact to first order, i.e., has a small error. More precisely, it is the asymptotic limit of the well-established Navier–Stokes system for fluid flow and the convection-diffusion equation for solute transport (there are no ad hoc rules imposed). (2) The asymptotic limit used results from considering the simple but, we believe, fundamental effect of Figure 1.1(e). While this local effect has been identified and appears in standard texts on porous media [30], its global consequence (i.e., the combined effect of this phenomenon at all intersections) has not been well studied. This method is an ideal tool for those studies. (3) This method is relatively computationally inexpensive, essentially solving a linear system whose number of variables is equal to two times the number of nodes in a period cell. We mention that the extension of our method to three dimension is immediate.

Appendix A. Asymptotic approximation.

A.1. Dimensionless variables, parameters, and equations. We first define the small dimensionless parameter ε as

$$\varepsilon = \frac{\ell}{vt_0} \ll 1 \tag{A.1}$$

for each edge e, the dimensionless parameters as

$$\bar{\ell}_e = \frac{\ell_e}{\ell} \text{ and } \bar{\delta}_e = \frac{\delta_e}{\delta}, \tag{A.2}$$

the dimensionless velocities and their norms as

$$\bar{\mathbf{v}}_e = \frac{\mathbf{v}_e}{v} \text{ and } \bar{v}_e = \|\bar{\mathbf{v}}_e\|, \tag{A.3}$$

and the dimensionless average velocity as

$$\bar{\mathbf{V}} = \frac{\mathbf{V}}{v}. \tag{A.4}$$

The velocities will be periodic with the same period t_0 as the applied pressure gradient. This motivates the choice of the dimensionless time

$$\bar{t} = \frac{t}{t_0}. \tag{A.5}$$

We regard the dimensionless velocities as 1-periodic functions of the dimensionless time; i.e., $\bar{\mathbf{v}}_e = \bar{\mathbf{v}}_e(\bar{t})$ and $\bar{\mathbf{V}} = \bar{\mathbf{V}}(\bar{t})$ are periodic with period 1.

Since the velocities are of order v, distances traveled by convection in periods of times of order t_0 are of order vt_0. This motivates the use of the following space dimensionless variable:

$$\bar{\mathbf{x}} = \frac{\mathbf{x}}{vt_0}. \tag{A.6}$$

Accordingly, the dimensionless nodes are the set

$$\bar{\mathcal{N}} = \frac{\mathcal{N}}{vt_0}, \tag{A.7}$$

the dimensionless head and tail of each edge e are

$$\bar{\mathbf{h}}(e) = \frac{\mathbf{h}(e)}{vt_0} \text{ and } \bar{\mathbf{k}}(e) = \frac{\mathbf{k}(e)}{vt_0}, \tag{A.8}$$

respectively, and the dimensionless vectors that determine the periodicity of the microstructure are

$$\bar{\mathbf{w}} = \frac{\mathbf{w}}{vt_0} \text{ and } \bar{\mathbf{q}} = \frac{\mathbf{q}}{vt_0}. \tag{A.9}$$

Thus, parametrizing the segment joining $\bar{\mathbf{k}}(e)$ and $\bar{\mathbf{h}}(e)$ by

$$\bar{\mathbf{x}}_e(\bar{s}) = \bar{\mathbf{k}}(e) + \bar{s}\frac{\bar{\mathbf{h}}(e) - \bar{\mathbf{k}}(e)}{\|\bar{\mathbf{h}}(e) - \bar{\mathbf{k}}(e)\|}, \tag{A.10}$$

we have that (2.11) becomes

$$\frac{\partial \bar{u}_e}{\partial \bar{t}} + \bar{v}_e \frac{\partial \bar{u}_e}{\partial \bar{s}} = 0 \text{ for } 0 \leq \bar{s} \leq \varepsilon \bar{\ell}_e,\ \bar{t} \geq 0, \text{ and all } e \in \mathcal{E}, \tag{A.11}$$

where $\bar{u}_e(\bar{s}, \bar{t})$ is the solute concentration in the channel e at the point $vt_0\bar{\mathbf{x}}_e(\bar{s})$ and time $t_0\bar{t}$. Note that $\bar{\mathbf{k}}(e) = \bar{\mathbf{x}}_e(0)$ and $\bar{\mathbf{h}}(e) = \bar{\mathbf{x}}_e(\varepsilon\bar{\ell}_e)$.

On the other hand, (2.12) becomes

$$\bar{u}_e(0, \bar{t}) = \frac{\sum_{\{\beta : \bar{\mathbf{h}}(\beta) = \bar{\mathbf{k}}(e)\}} \bar{\delta}_\beta \bar{v}_\beta \bar{u}_\beta(\varepsilon\bar{\ell}_\beta, \bar{t})}{\sum_{\{\beta : \bar{\mathbf{h}}(\beta) = \bar{\mathbf{k}}(e)\}} \bar{\delta}_\beta \bar{v}_\beta}. \tag{A.12}$$

A.2. Solute transport within each channel. Let e be an edge. Solute is convected from $\bar{\mathbf{k}}(e)$, the tail of e, to $\bar{\mathbf{h}}(e)$, the head of e. Thus, the solute concentration at $\bar{\mathbf{h}}(e)$ at time $\bar{t}$ is equal to the solute concentration at $\bar{\mathbf{k}}(e)$ at an earlier time $\bar{t} - \Delta\bar{t}_e$, i.e.,

$$\bar{u}_e(0, \bar{t} - \Delta\bar{t}_e) = \bar{u}_e(\varepsilon\bar{\ell}_e, \bar{t}). \tag{A.13}$$

We next compute $\Delta\bar{t}_e$.

Let $S(\tau)$ be the solution of

$$S'(\tau) = \bar{v}_e(\tau) \text{ and } S(\bar{t}) = \varepsilon\bar{\ell}_e, \tag{A.14}$$

where S' is the derivative of S. We claim that $\Delta\bar{t}_e$ is the solution of

$$S(\bar{t} - \Delta\bar{t}_e) = 0. \tag{A.15}$$

This is due to the fact that $\bar{u}_e(S(\tau), \tau)$ is independent of τ, and thus $\bar{u}_e(0, \bar{t} - \Delta\bar{t}_e) = \bar{u}_e(S(\bar{t} - \Delta\bar{t}_e), \bar{t} - \Delta\bar{t}_e) = \bar{u}_e(S(\bar{t}), \bar{t}) = \bar{u}_e(\varepsilon\bar{\ell}_e, \bar{t})$.

From (A.14) and (A.15), we note that $\Delta\bar{t}_e = O(\varepsilon)$. Thus, we Taylor expand (A.15) to get

$$0 = S(\bar{t} - \Delta\bar{t}_e) \simeq S(\bar{t}) - S'(\bar{t})\Delta\bar{t}_e + \frac{S''(\bar{t})}{2}(\Delta\bar{t}_e)^2. \tag{A.16}$$

Next we note that $S(\bar{t}) = \varepsilon\bar{\ell}_e$ and $S'(\bar{t}) = \bar{v}_e(\bar{t})$ (see (A.14)). Thus, $S''(\bar{t}) = \bar{v}_e'(\bar{t})$, and we conclude from (A.16) that

$$0 \simeq \varepsilon\bar{\ell}_e - \bar{v}_e(\bar{t})\Delta\bar{t}_e + \frac{\bar{v}_e'(\bar{t})}{2}(\Delta\bar{t}_e)^2. \tag{A.17}$$

We now set $\Delta\bar{t}_e = \varepsilon\Delta\bar{t}_1 + \varepsilon^2\Delta\bar{t}_2$, plug this expression into (A.17), collect powers of ε to obtain equations for $\Delta\bar{t}_1$ and $\Delta\bar{t}_2$, and finally get

$$\Delta\bar{t}_e \simeq \varepsilon\frac{\bar{\ell}_e}{\bar{v}_e} + \varepsilon^2\frac{\bar{\ell}_e^2}{2}\frac{\bar{v}_e'}{\bar{v}_e^3}, \tag{A.18}$$

where $\bar{v}_e$ and $\bar{v}_e'$ are evaluated at $\bar{t}$. Note that $\Delta\bar{t}_e$ is a function of $\bar{t}$.

A.3. Continuum approximation. Using (A.13), (A.12) becomes

$$\bar{u}_e(0,\bar{t}) = \frac{\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{k}}(e)\}} \bar{\delta}_\beta \bar{v}_\beta \bar{u}_\beta(0,\bar{t}-\Delta\bar{t}_\beta)}{\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{k}}(e)\}} \bar{\delta}_\beta \bar{v}_\beta}, \tag{A.19}$$

where $\Delta\bar{t}_\beta$ is given by (A.18).

For each dimensionless node $\bar{\mathbf{a}} \in \bar{\mathcal{N}}$, let $\bar{u}_{\bar{\mathbf{a}}}(\bar{t})$ be the solute concentration that leaves the intersection $\bar{\mathbf{a}}$, i.e.,

$$\bar{u}_{\bar{\mathbf{a}}}(\bar{t}) = \bar{u}_e(0,\bar{t}) \text{ if } \bar{\mathbf{a}} = \bar{\mathbf{k}}(e) \text{ at time } \bar{t}. \tag{A.20}$$

With the notation introduced in (A.20), (A.19) becomes

$$\bar{u}_{\bar{\mathbf{a}}}(\bar{t}) = \frac{\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}} \bar{\delta}_\beta \bar{v}_\beta \bar{u}_{\bar{\mathbf{k}}(\beta)}(\bar{t}-\Delta\bar{t}_\beta)}{\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}} \bar{\delta}_\beta \bar{v}_\beta}. \tag{A.21}$$

A.3.1. Ansatz and expansions. We now use standard asymptotic techniques to obtain the macroscopic transport equation for the solute concentration. We propose the ansatz

$$\bar{u}_{\bar{\mathbf{a}}}(\bar{t}) = \rho(\bar{\mathbf{a}},\bar{t},\varepsilon\bar{t}) + \varepsilon f_{\hat{\mathbf{a}}}(\bar{\mathbf{a}},\bar{t},\varepsilon\bar{t}) + \varepsilon^2 g_{\hat{\mathbf{a}}}(\bar{\mathbf{a}},\bar{t},\varepsilon\bar{t}), \text{ where } \hat{\mathbf{a}} = \frac{\bar{\mathbf{a}}}{\varepsilon}, \tag{A.22}$$

$\rho(\bar{\mathbf{x}},\bar{t},\tau)$ is a smooth function of its variables periodic in $\bar{t}$ with period 1, and for each $\hat{\mathbf{a}} \in \bar{\mathcal{N}}/\varepsilon = \mathcal{N}/\ell$ the functions $f_{\hat{\mathbf{a}}}(\bar{\mathbf{x}},\bar{t},\tau)$ and $g_{\hat{\mathbf{a}}}(\bar{\mathbf{x}},\bar{t},\tau)$ are smooth functions of $\bar{\mathbf{x}}$, $\bar{t}$, and τ and are also periodic in $\bar{t}$ with period 1. The family of functions $f_{\hat{\mathbf{a}}}$ and $g_{\hat{\mathbf{a}}}$ are periodic in $\hat{\mathbf{a}}$ in the sense that $f_{\hat{\mathbf{a}}+(n\bar{\mathbf{w}}+m\bar{\mathbf{q}})/\varepsilon}(\bar{\mathbf{x}},\bar{t},\tau) = f_{\hat{\mathbf{a}}}(\bar{\mathbf{x}},\bar{t},\tau)$ and $g_{\hat{\mathbf{a}}+(n\bar{\mathbf{w}}+m\bar{\mathbf{q}})/\varepsilon}(\bar{\mathbf{x}},\bar{t},\tau) = g_{\hat{\mathbf{a}}}(\bar{\mathbf{x}},\bar{t},\tau)$ for all integers n and m and $\hat{\mathbf{a}} \in \bar{\mathcal{N}}/\varepsilon$.

Let $\bar{\mathbf{x}}$ be a point that we hold fixed for the moment. Let $\bar{\mathbf{a}}$ be a dimensionless node $\bar{\mathbf{a}} \in \bar{\mathcal{N}}$ such that $\|\bar{\mathbf{x}} - \bar{\mathbf{a}}\| = O(\varepsilon)$. We write

$$\bar{\mathbf{a}} = \bar{\mathbf{x}} + \varepsilon(\hat{\mathbf{a}} - \hat{\mathbf{x}}), \text{ where } \hat{\mathbf{a}} = \frac{\bar{\mathbf{a}}}{\varepsilon} \text{ and } \hat{\mathbf{x}} = \frac{\bar{\mathbf{x}}}{\varepsilon}. \tag{A.23}$$

We now plug this expression for $\bar{\mathbf{a}}$ into the right-hand side of (A.22) to get

$$\begin{aligned}\bar{u}_{\bar{\mathbf{a}}}(\bar{t}) = \rho(\bar{\mathbf{x}} + \varepsilon(\hat{\mathbf{a}} - \hat{\mathbf{x}}),\bar{t},\varepsilon\bar{t}) + \varepsilon f_{\hat{\mathbf{a}}}(\bar{\mathbf{x}} + \varepsilon(\hat{\mathbf{a}} - \hat{\mathbf{x}}),\bar{t},\varepsilon\bar{t})\\ +\varepsilon^2 g_{\hat{\mathbf{a}}}(\bar{\mathbf{x}} + \varepsilon(\hat{\mathbf{a}} - \hat{\mathbf{x}}),\bar{t},\varepsilon\bar{t}).\end{aligned} \tag{A.24}$$

We now Taylor expand the right-hand side of the above equality around the point $(\bar{\mathbf{x}},\bar{t},\varepsilon\bar{t})$ to get

$$\begin{aligned}\bar{u}_{\bar{\mathbf{a}}}(\bar{t}) \simeq \rho + \varepsilon\left\{\sum_i \frac{\partial\rho}{\partial\bar{x}_i}[\hat{\mathbf{a}} - \hat{\mathbf{x}}]_i + f_{\hat{\mathbf{a}}}\right\}\\ +\varepsilon^2\left\{\frac{1}{2}\sum_{i,j}\frac{\partial^2\rho}{\partial\bar{x}_i\partial\bar{x}_j}[\hat{\mathbf{a}} - \hat{\mathbf{x}}]_i[\hat{\mathbf{a}} - \hat{\mathbf{x}}]_j + \sum_i \frac{\partial f_{\hat{\mathbf{a}}}}{\partial\bar{x}_i}[\hat{\mathbf{a}} - \hat{\mathbf{x}}]_i + g_{\hat{\mathbf{a}}}\right\},\end{aligned} \tag{A.25}$$

where ρ, $f_{\hat{\mathbf{a}}}$, $g_{\hat{\mathbf{a}}}$ and their derivatives are evaluated at $(\bar{\mathbf{x}},\bar{t},\varepsilon\bar{t})$. In the above equation we neglected terms of order ε^3.

Now let β be an edge whose dimensionless head is $\bar{\mathbf{a}}$, i.e., $\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}$. Note that $\|\bar{\mathbf{k}}(\beta)-\bar{\mathbf{a}}\|=\varepsilon\bar{\ell}_\beta$. Thus, since $\|\bar{\mathbf{x}}-\bar{\mathbf{a}}\|=O(\varepsilon)$, we have $\|\bar{\mathbf{x}}-\bar{\mathbf{k}}(\beta)\|=O(\varepsilon)$. We write

$$\bar{\mathbf{k}}(\beta)=\bar{\mathbf{x}}+\varepsilon(\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}),\ \text{where } \hat{\mathbf{k}}(\beta)=\frac{\bar{\mathbf{k}}(\beta)}{\varepsilon} \text{ and } \hat{\mathbf{x}}=\frac{\bar{\mathbf{x}}}{\varepsilon}. \tag{A.26}$$

From (A.24), but replacing $\bar{\mathbf{a}}$ by $\bar{\mathbf{k}}(\beta)$ and $\bar{t}$ by $\bar{t}-\Delta\bar{t}_\beta$, we get

$$\begin{aligned}\bar{u}_{\bar{\mathbf{k}}(\beta)}(\bar{t}-\Delta\bar{t}_\beta)=&\rho(\bar{\mathbf{x}}+\varepsilon(\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}),\bar{t}-\Delta\bar{t}_\beta,\varepsilon\bar{t}-\varepsilon\Delta\bar{t}_\beta)\\ &+\varepsilon f_{\hat{\mathbf{k}}(\beta)}(\bar{\mathbf{x}}+\varepsilon(\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}),\bar{t}-\Delta\bar{t}_\beta,\varepsilon\bar{t}-\varepsilon\Delta\bar{t}_\beta)\\ &+\varepsilon^2 g_{\hat{\mathbf{k}}(\beta)}(\bar{\mathbf{x}}+\varepsilon(\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}),\bar{t}-\Delta\bar{t}_\beta,\varepsilon\bar{t}-\varepsilon\Delta\bar{t}_\beta).\end{aligned} \tag{A.27}$$

We now Taylor expand the right-hand side of the above equation around the point $(\bar{\mathbf{x}},\bar{t},\varepsilon\bar{t})$ and make use of the expression (A.18) (with e replaced by β) to get

$$\begin{aligned}\bar{u}_{\bar{\mathbf{k}}(\beta)}(\bar{t}-\Delta\bar{t}_\beta)\simeq\rho+\varepsilon\left\{\sum_i\frac{\partial\rho}{\partial\bar{x}_i}[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}]_i-\frac{\partial\rho}{\partial\bar{t}}\frac{\bar{\ell}_\beta}{\bar{v}_\beta}+f_{\hat{\mathbf{k}}(\beta)}\right\}\\ +\varepsilon^2\left\{\frac{1}{2}\sum_{i,j}\frac{\partial^2\rho}{\partial\bar{x}_i\partial\bar{x}_j}[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}]_i[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}]_j-\sum_i\frac{\partial^2\rho}{\partial\bar{t}\partial\bar{x}_i}[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}]_i\frac{\bar{\ell}_\beta}{\bar{v}_\beta}\right.\\ \left.+\frac{1}{2}\frac{\partial^2\rho}{\partial\bar{t}^2}\frac{\bar{\ell}_\beta^2}{\bar{v}_\beta^2}-\frac{\partial\rho}{\partial\bar{t}}\frac{\bar{\ell}_\beta^2}{2}\frac{\bar{v}'_\beta}{\bar{v}_\beta^3}-\frac{\partial\rho}{\partial\tau}\frac{\bar{\ell}_\beta}{\bar{v}_\beta}+\sum_i\frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial\bar{x}_i}[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}]_i-\frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial\bar{t}}\frac{\bar{\ell}_\beta}{\bar{v}_\beta}+g_{\hat{\mathbf{k}}(\beta)}\right\},\end{aligned} \tag{A.28}$$

where ρ, $f_{\hat{\mathbf{k}}(\beta)}$, $g_{\hat{\mathbf{k}}(\beta)}$ and their derivatives are evaluated in $(\bar{x},\bar{t},\varepsilon\bar{t})$, and $\bar{v}_\beta$ and its derivative are evaluated at $\bar{t}$. In the above equation we neglected terms of order ε^3.

Now we plug the expressions for $\bar{u}_{\hat{\mathbf{a}}}(\bar{t})$ and $\bar{u}_{\bar{\mathbf{k}}(\beta)}(\bar{t}-\Delta\bar{t}_\beta)$ given in (A.25) and (A.28) into (A.21), neglect terms of ε^2, and make simple algebraic manipulations (which include dividing by ε) to obtain

$$\sum_{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}}\bar{\delta}_\beta\bar{v}_\beta(f_{\hat{\mathbf{k}}(\beta)}-f_{\hat{\mathbf{a}}})=\left(\sum_{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}}\bar{\delta}_\beta\bar{\ell}_\beta\right)\frac{\partial\rho}{\partial\bar{t}}+\left(\sum_{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}}\bar{\delta}_\beta\bar{\ell}_\beta\bar{\mathbf{v}}_\beta\right)\cdot\bar{\nabla}\rho. \tag{A.29}$$

We require the above equation to be valid for all $(\bar{x},\bar{t},\tau)$, not just at $\tau=\varepsilon\bar{t}$.

A.3.2. Fredholm alternative. First equation for ρ. Convection with average velocity. We recall that two edges are equivalent if one is the translation of the other by a vector of the form $n\mathbf{w}+m\mathbf{q}$, where n and m are integers, and we denote by $\mathcal{F}$ a set of edges that contains exactly one edge per equivalent class (see section 3). Analogously, we also say that two dimensionless nodes $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are equivalent if $\bar{\mathbf{b}}=\bar{\mathbf{a}}+n\bar{\mathbf{w}}+m\bar{\mathbf{q}}$ for some n and m integers, and we denote by $\bar{\mathcal{M}}$ a set of dimensionless nodes that contains exactly one dimensionless node per equivalent class. For example, $\bar{\mathcal{M}}$ could be all the dimensionless nodes included in the dimensionless period cell $\bar{Q}=\{s\bar{\mathbf{w}}+r\bar{\mathbf{q}}:0\le s,r<1\}$, i.e., $\bar{\mathcal{M}}=\bar{\mathcal{N}}\cap\bar{Q}$.

Let $y_{\hat{\mathbf{a}}}$ be defined for all $\hat{\mathbf{a}}\in\bar{\mathcal{N}}/\varepsilon$ and have the same periodicity as $f_{\hat{\mathbf{a}}}$, i.e., $y_{\hat{\mathbf{a}}+(n\bar{\mathbf{w}}+m\bar{\mathbf{q}})/\varepsilon}=y_{\hat{\mathbf{a}}}$ for all integers n and m and all $\hat{\mathbf{a}}\in\bar{\mathcal{N}}/\varepsilon$. Note that

$$\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}}y_{\hat{\mathbf{a}}}\sum_{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}}\bar{\delta}_\beta\bar{v}_\beta f_{\hat{\mathbf{k}}(\beta)}=\sum_{\beta\in\mathcal{F}}\bar{\delta}_\beta\bar{v}_\beta y_{\hat{\mathbf{h}}(\beta)}f_{\hat{\mathbf{k}}(\beta)}=\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}}f_{\hat{\mathbf{a}}}\sum_{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}}\bar{\delta}_\beta\bar{v}_\beta y_{\hat{\mathbf{h}}(\beta)}. \tag{A.30}$$

Thus,

$$\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}} y_{\hat{\mathbf{a}}} \sum_{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}} \bar{\delta}_\beta \bar{v}_\beta (f_{\hat{\mathbf{k}}(\beta)} - f_{\hat{\mathbf{a}}}) = \sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}} f_{\hat{\mathbf{a}}} \sum_{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}} \bar{\delta}_\beta \bar{v}_\beta (y_{\hat{\mathbf{h}}(\beta)} - y_{\hat{\mathbf{a}}}). \tag{A.31}$$

The above expression is equal to 0 for all periodic $f_{\hat{a}}$ if and only if

$$\sum_{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}} \bar{\delta}_\beta \bar{v}_\beta (y_{\hat{\mathbf{h}}(\beta)} - y_{\hat{\mathbf{a}}}) = 0 \text{ for all } \bar{\mathbf{a}} \in \bar{\mathcal{N}}. \tag{A.32}$$

A simple calculation shows that $y_{\hat{\mathbf{b}}} = y_{\hat{\mathbf{a}}}$ for all $\hat{\mathbf{a}}, \hat{\mathbf{b}} \in \bar{\mathcal{N}}/\varepsilon$ (if $\bar{v}_e \neq 0$ for all edges e. This is a generic condition that we assume is satisfied).

If we multiply the left-hand side of (A.29) by $y_{\hat{\mathbf{a}}}$ and add over all $\mathbf{a} \in \bar{\mathcal{M}}$, we obtain the expression in (A.31). Thus, the above discussion, the Fredholm alternative theory, and simple manipulations imply that there exists a solution $f_{\hat{\mathbf{a}}}$ (of (A.29)) that satisfies $f_{\hat{\mathbf{a}}+(n\bar{\mathbf{w}}+m\bar{\mathbf{z}})/\varepsilon} = f_{\hat{\mathbf{a}}}$ for all integers n and m if and only if

$$\frac{\partial \rho}{\partial \bar{t}} + \bar{\mathbf{V}} \cdot \bar{\nabla} \rho = 0, \tag{A.33}$$

where $\bar{\nabla}\rho$ is the gradient of ρ with respect to $\bar{\mathbf{x}}$.

A.3.3. Further expansions. Next we take derivatives of (A.33) with respect to $\bar{x}_i$ to obtain

$$\frac{\partial^2 \rho}{\partial \bar{t} \partial \bar{x}_i} = -\sum_j [\bar{\mathbf{V}}]_j \frac{\partial^2 \rho}{\partial \bar{x}_i \partial \bar{x}_j}. \tag{A.34}$$

Analogously, taking derivatives of (A.33) with respect to $\bar{t}$ and using (A.34), we obtain

$$\frac{\partial^2 \rho}{\partial \bar{t}^2} = -\sum_i [\bar{\mathbf{V}}']_i \frac{\partial \rho}{\partial \bar{x}_i} - \sum_i [\bar{\mathbf{V}}]_i \frac{\partial^2 \rho}{\partial \bar{x}_i \partial \bar{t}} = -\sum_i [\bar{\mathbf{V}}']_i \frac{\partial \rho}{\partial \bar{x}_i} + \sum_{i,j} [\bar{\mathbf{V}}]_i [\bar{\mathbf{V}}]_j \frac{\partial^2 \rho}{\partial \bar{x}_i \partial \bar{x}_j}. \tag{A.35}$$

Using the last three identities, (A.28) becomes

$$\begin{aligned} \bar{u}_{\bar{\mathbf{k}}(\beta)}(\bar{t} - \Delta \bar{t}_\beta) = \rho + \varepsilon \left\{ \sum_i \frac{\partial \rho}{\partial \bar{x}_i} \left[\hat{\mathbf{k}}(\beta) - \hat{\mathbf{x}} + \frac{\bar{\ell}_\beta}{\bar{v}_\beta} \bar{\mathbf{V}} \right]_i + f_{\hat{\mathbf{k}}(\beta)} \right\} \\ + \varepsilon^2 \left\{ \frac{1}{2} \sum_{i,j} \frac{\partial^2 \rho}{\partial \bar{x}_i \partial \bar{x}_j} \left[\hat{\mathbf{k}}(\beta) - \hat{\mathbf{x}} + \frac{\bar{\ell}_\beta}{\bar{v}_\beta} \bar{\mathbf{V}} \right]_i \left[\hat{\mathbf{k}}(\beta) - \hat{\mathbf{x}} + \frac{\bar{\ell}_\beta}{\bar{v}_\beta} \bar{\mathbf{V}} \right]_j \right. \\ \left. - \frac{1}{2} \frac{\bar{\ell}_\beta}{\bar{v}_\beta} \sum_i \left(\frac{\bar{\ell}_\beta}{\bar{v}_\beta} [\bar{\mathbf{V}}]_i \right)' \frac{\partial \rho}{\partial \bar{x}_i} - \frac{\partial \rho}{\partial \tau} \frac{\bar{\ell}_\beta}{\bar{v}_\beta} + \sum_i \frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial \bar{x}_i} [\hat{\mathbf{k}}(\beta) - \hat{\mathbf{x}}]_i - \frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial \bar{t}} \frac{\bar{\ell}_\beta}{\bar{v}_\beta} + g_{\hat{\mathbf{k}}(\beta)} \right\}. \end{aligned} \tag{A.36}$$

We now plug into (A.21) the expressions for $\bar{u}_{\hat{\mathbf{k}}(\beta)}(\bar{t} - \Delta \bar{t}_\beta)$ and $\bar{u}_a(\bar{t})$ given by (A.36) and (A.25), respectively, to obtain, after some algebraic manipulations and making use of (A.29), that

(A.37)
$$\left(\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\right)\frac{\partial\rho}{\partial\tau}+\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\left[\bar{\delta}_\beta\bar{v}_\beta\left(\bar{\nabla}f_{\hat{\mathbf{a}}}\cdot(\hat{\mathbf{a}}-\hat{\mathbf{x}})-\bar{\nabla}f_{\hat{\mathbf{k}}(\beta)}\cdot\left(\hat{\mathbf{k}}(\beta)-\hat{\mathbf{x}}\right)\right)\right.$$
$$\left.+\frac{\bar{\delta}_\beta}{2}\bar{\ell}_\beta^2\left(\frac{\bar{\mathbf{V}}}{\bar{v}_\beta}\right)'\cdot\bar{\nabla}\rho+\bar{\delta}_\beta\bar{\ell}_\beta\frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial\bar{t}}\right]$$
$$=\frac{1}{2}\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{v}_\beta\sum_{i,j}\frac{\partial^2\rho}{\partial\bar{x}_i\partial\bar{x}_j}\left(\left[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{a}}+\frac{\bar{\ell}_\beta}{\bar{v}_\beta}\bar{\mathbf{V}}\right]_i\left[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{a}}+\frac{\bar{\ell}_\beta}{\bar{v}_\beta}\bar{\mathbf{V}}\right]_j\right.$$
$$\left.2+[\hat{\mathbf{a}}-\hat{\mathbf{x}}]_i\left[\hat{\mathbf{k}}(\beta)-\hat{\mathbf{a}}+\frac{\bar{\ell}_\beta}{\bar{v}_\beta}\bar{\mathbf{V}}\right]_j\right)+\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{v}_\beta\left(g_{\hat{\mathbf{k}}(\beta)}-g_{\hat{\mathbf{a}}}\right).$$

Taking the gradient of (A.29), taking the dot product of the result with $(\hat{\mathbf{a}}-\hat{\mathbf{x}})$, making use of previous equations, performing some algebraic manipulations, and noting that $\bar{v}_\beta(\hat{\mathbf{a}}-\hat{\mathbf{k}}(\beta))=\bar{\ell}_\beta\bar{\mathbf{v}}_\beta$, (A.37) can be reduced to

(A.38)
$$\left(\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\right)\frac{\partial\rho}{\partial\tau}+\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\left[\bar{\nabla}f_{\hat{\mathbf{k}}(\beta)}\cdot\bar{\mathbf{v}}_\beta+\frac{\bar{\ell}_\beta}{2}\left(\frac{\bar{\mathbf{V}}}{\bar{v}_\beta}\right)'\cdot\bar{\nabla}\rho+\frac{\partial f_{\hat{\mathbf{k}}(\beta)}}{\partial\bar{t}}\right]$$
$$=\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{v}_\beta\left(g_{\hat{\mathbf{k}}(\beta)}-g_{\hat{\mathbf{a}}}\right)+\frac{1}{2}\sum_{i,j}\frac{\partial^2\rho}{\partial\bar{x}_i\partial\bar{x}_j}\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\frac{\bar{\delta}_\beta\bar{\ell}_\beta^2}{\bar{v}_\beta}\left[\bar{\mathbf{v}}_\beta-\bar{\mathbf{V}}\right]_i\left[\bar{\mathbf{v}}_\beta-\bar{\mathbf{V}}\right]_j.$$

A.4. Fredholm alternative. Long time equation for ρ. Dispersion tensor. Following the same arguments to obtain (A.33), we have that there exists a solution $g_{\hat{\mathbf{a}}}$ of (A.38) that is periodic in $\hat{\mathbf{a}}$ if and only if

$$\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}}\left\{\left(\sum_{\{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\right)\frac{\partial f_{\hat{\mathbf{a}}}}{\partial\bar{t}}+\left(\sum_{\{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\bar{\mathbf{v}}_\beta\right)\cdot\bar{\nabla}f_{\hat{\mathbf{a}}}\right\}$$
$$\text{(A.39)}\quad+\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}}\left\{\left(\sum_{\{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\right)\frac{\partial\rho}{\partial\tau}+\left(\sum_{\{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}\}}\frac{\bar{\delta}_\beta}{2}\bar{\ell}_\beta^2\left(\frac{\bar{\mathbf{V}}}{\bar{v}_\beta}\right)'\right)\cdot\bar{\nabla}\rho\right\}$$
$$=\frac{1}{2}\sum_{i,j}\frac{\partial^2\rho}{\partial\bar{x}_i\partial\bar{x}_j}\sum_{\bar{\mathbf{a}}\in\bar{\mathcal{M}}}\sum_{\{\beta:\bar{\mathbf{k}}(\beta)=\bar{\mathbf{a}}\}}\frac{\bar{\delta}_\beta\bar{\ell}_\beta^2}{\bar{v}_\beta}\left[\bar{\mathbf{v}}_\beta-\bar{\mathbf{V}}\right]_i\left[\bar{\mathbf{v}}_\beta-\bar{\mathbf{V}}\right]_j.$$

Let $\mathbf{F}_{\hat{\mathbf{a}}}=\mathbf{F}_{\hat{\mathbf{a}}}(\bar{t})$ be a solution of

$$\text{(A.40)}\qquad\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{v}_\beta\left(\mathbf{F}_{\hat{\mathbf{k}}(\beta)}-\mathbf{F}_{\hat{\mathbf{a}}}\right)=\sum_{\{\beta:\bar{\mathbf{h}}(\beta)=\bar{\mathbf{a}}\}}\bar{\delta}_\beta\bar{\ell}_\beta\left(\bar{\mathbf{v}}_\beta-\bar{\mathbf{V}}\right)$$

that is periodic in $\bar{t}$ and $\hat{\mathbf{a}}$, i.e., $\mathbf{F}_{\hat{\mathbf{a}}+(n\bar{\mathbf{w}}+m\bar{\mathbf{q}})/\varepsilon}(\bar{t}+p)=\mathbf{F}_{\hat{\mathbf{a}}}(\bar{t})$ for all integers n,m,p. A simple calculation shows that $f_{\hat{\mathbf{a}}}$ is a periodic (in $\hat{\mathbf{a}}$ and $\bar{t}$) solution of (A.29) if and only if

$$\text{(A.41)}\qquad f_{\hat{\mathbf{a}}}(\bar{\mathbf{x}},\bar{t},\tau)=\mathbf{F}_{\hat{\mathbf{a}}}(\bar{t})\cdot\bar{\nabla}\rho(\bar{x},\bar{t},\tau)+\psi(\bar{x},\bar{t},\tau),$$

where ψ is an arbitrary function that is periodic on $\bar{t}$.

Note that $\{\beta : \bar{\mathbf{k}}(\beta) = \bar{\mathbf{a}} \text{ and } \bar{\mathbf{a}} \in \bar{\mathcal{M}}\}$ is a set that contains exactly one edge per equivalent class. Thus, the sums in equation (A.39) are spatial averages, i.e., $\sum_{\bar{\mathbf{a}} \in \bar{\mathcal{M}}} \sum_{\{\beta : \bar{\mathbf{k}}(\beta) = \bar{\mathbf{a}}\}} = \sum_{\beta \in \mathcal{F}}$, where we recall that $\mathcal{F}$ is any set of edges that contains exactly one edge per equivalent class. Using this observation, (A.41), and some simple algebraic manipulations, we transform (A.39) in

$$\begin{aligned}
&\left(\sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta\right) \left(\frac{\partial \rho}{\partial \tau} + \frac{\partial \psi}{\partial \bar{t}} + \bar{\mathbf{V}} \cdot \bar{\nabla} \psi\right) + \left(\sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta \mathbf{F}'_{\hat{\mathbf{k}}(\beta)} + \frac{\bar{\delta}_\beta}{2} \bar{\ell}_\beta^2 \left(\frac{\bar{\mathbf{V}}}{\bar{v}_\beta}\right)'\right) \cdot \bar{\nabla} \rho \\
&= \frac{1}{2} \sum_{i,j} \frac{\partial^2 \rho}{\partial \bar{x}_i \partial \bar{x}_j} \sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta \left(\frac{\bar{\ell}_\beta}{\bar{v}_\beta} \left[\bar{\mathbf{v}}_\beta - \bar{\mathbf{V}}\right]_i \left[\bar{\mathbf{v}}_\beta - \bar{\mathbf{V}}\right]_j + 2 \left[\mathbf{F}_{\hat{\mathbf{k}}(\beta)}\right]_i \left[\bar{\mathbf{V}} - \bar{\mathbf{v}}_\beta\right]_j\right).
\end{aligned} \tag{A.42}$$

We now make the change of variables

$$\bar{\mathbf{y}} = \bar{\mathbf{x}} - \int_0^{\bar{t}} \bar{\mathbf{V}}(s)\, \mathrm{d}s. \tag{A.43}$$

To avoid confusion, we denote ρ by $\bar{\rho}$ when the new independent variables $(\bar{\mathbf{y}}, \bar{t}, \tau)$ are used. In these new variables, (A.33) becomes $\partial \bar{\rho}/(\partial \bar{t}) = 0$, from which we get that $\bar{\rho}$ is a function that depends only on $\bar{\mathbf{y}}$ and τ, i.e., $\bar{\rho} = \bar{\rho}(\bar{\mathbf{y}}, \tau)$.

Analogously, $\bar{\psi}$ is simply ψ, but only when the new independent variables $(\bar{\mathbf{y}}, \bar{t}, \tau)$ are used. The changes that occur in (A.42) when the new variables are used are the following: (1) The term $\bar{\mathbf{V}} \cdot \bar{\nabla} \psi$ is removed; (2) derivatives with respect to $\bar{\mathbf{x}}$ are replaced by derivatives with respect to $\bar{\mathbf{y}}$; and (3) ρ is replaced by $\bar{\rho}$ and ψ by $\bar{\psi}$. After making this change of variables, we integrate the resulting equation (A.42) with respect to $\bar{t}$ over a period, from $\bar{t} = \bar{t}_0$ to $\bar{t} = \bar{t}_0 + 1$, and divide by $\sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta$ to obtain

$$\frac{\partial \bar{\rho}}{\partial \tau}(\bar{\mathbf{y}}, \tau) + \bar{\psi}(\bar{\mathbf{y}}, \bar{t}_0 + 1, \tau) - \bar{\psi}(\bar{\mathbf{y}}, \bar{t}_0, \tau) = \sum_{i,j} \bar{D}_{ij} \frac{\partial^2 \bar{\rho}}{\partial y_i \partial y_j}(\bar{\mathbf{y}}, \tau), \tag{A.44}$$

where

$$\begin{aligned}
\bar{D}_{ij} = \frac{1}{2 \sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta} \sum_{\beta \in \mathcal{F}} \bar{\delta}_\beta \bar{\ell}_\beta \int_0^1 &\left(\frac{\bar{\ell}_\beta}{\bar{v}_\beta} \left[\bar{\mathbf{v}}_\beta - \bar{\mathbf{V}}\right]_i \left[\bar{\mathbf{v}}_\beta - \bar{\mathbf{V}}\right]_j \right. \\
&\left. + \left[\mathbf{F}_{\hat{\mathbf{k}}(\beta)}\right]_i \left[\bar{\mathbf{V}} - \bar{\mathbf{v}}_\beta\right]_j + \left[\mathbf{F}_{\hat{\mathbf{k}}(\beta)}\right]_j \left[\bar{\mathbf{V}} - \bar{\mathbf{v}}_\beta\right]_i\right) \mathrm{d}\bar{t}.
\end{aligned} \tag{A.45}$$

Finally, the dependence of $\bar{\rho}$ in τ is obtained by requiring $\bar{\psi}$ to be bounded. Given (A.44), we note that this occurs only if

$$\frac{\partial \bar{\rho}}{\partial \tau}(y, \tau) = \sum_{i,j} \bar{D}_{ij} \frac{\partial^2 \bar{\rho}}{\partial y_i \partial y_j}(y, \tau). \tag{A.46}$$

The result stated in section 3 is obtain by going back to the original dimensional variables.

REFERENCES

[1] P. M. Adler and H. Brenner, *Transport processes in spatially periodic capillary networks*–II. *Taylor dispersion with mixing vertices*, PhysicoChem. Hydrodyn., 5 (1984), pp. 269–285.

[2] P. M. Adler, *Porous Media. Geometry and Transports*, Butterworth-Heinemann, Stoneham, MA, 1992.

[3] R. Aris, *On the dispersion of solute matter in a fluid flowing through a tube*, Proc. Roy. Soc. London Ser. A, 235 (1956), pp. 67–77.

[4] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, New York, 1973.

[5] J. Bear, *On the tensor form of dispersion in porous media*, J. Geophys. Res., 66 (1961), pp. 1185–1197.

[6] J. Bear, *Dynamics of Fluids in Porous Media*, Dover Publications, New York, 1972.

[7] Berkowitz and Balberg, *Percolation theory and its applications to ground water hydrology*, Water Resour. Res., 29 (1993), pp. 775–794.

[8] B. Berkowitz and R. P. Ewing, *Percolation theory and network modeling applications in soil physics*, Surveys in Geophysics, 19 (1998), pp. 23–72.

[9] B. Berkowitz, *Dispersion in Heterogeneous Geological Formations*, Kluwer Academic Publishers, London, 2001.

[10] B. Berkowitz, *Characterizing flow and transport in fractured geological media: A review*, in Adv. Water Res., 25 (2002), pp. 861–884.

[11] H. Brenner, *Dispersion resulting from flow through spatially periodic porous media*, Philos. Trans. Roy. Soc. London Ser. A, 297 (1980), pp. 81–133.

[12] H. Brenner, *A general theory of Taylor dispersion phenomena*, PhysicoChem. Hydrodyn., 1 (1980), pp. 91–123.

[13] H. Brenner and P. M. Adler, *Transport Processes in Porous Media*, Hemisphere/McGraw-Hill, New York, 1985.

[14] H. Brenner and D. A. Edwards, *Macrotransport Processes*, Butterworth-Heinemann Seri. Chem. Engrg., Butterworth-Heinemann, Boston, MA, 1993.

[15] C. Bruderer and Y. Bernabe, *Network modeling of dispersion: Transition from Taylor dispersion in homogeneous networks to mechanical dispersion in very heterogeneous ones*, Water Resour. Res., 37 (2001), pp. 897–908.

[16] C. Bruderer-Wang, P. Cowie, Y. Bernabe, and I. Main, *Relating flow channeling to tracer dispersion in heterogeneous networks*, Adv. in Water Res., 27 (2004), pp. 843–855.

[17] M. L. Brusseau, *Transport of reactive contaminants in heterogeneous porous media*, Rev. Geophys., 32 (1994), pp. 285–313.

[18] R. P. Carbonell, *Effect of pore distribution and flow segregation on dispersion in porous media*, Chem. Eng. Sci., 34 (1979), pp. 1031–1039.

[19] H. R. Cedergren, *Seepage, Drainage, and Flow Nets*, John Wiley, New York, 1989.

[20] E. Charlaix and H. Gayvallet, *Hydrodynamic dispersion in networks of capillaries of random permeability*, Europhys. Lett., 16 (1991), pp. 259–264.

[21] K. H. Coats, *Dead-end pore volume and dispersion in porous media*, Soc. Pet. Eng. J., 4 (1964), pp. 73–84.

[22] J. H. Cushman, *Dynamics of Fluids in Hierarchical Porous Media*, Academic Press, New York, 1990.

[23] J. H. Cushman, L. S. Bennethum, and B. X. Hu, *A primer on upscaling tools for porous media*, Adv. in Water Res., 25 (2002), pp. 1043–1067.

[24] G. Dagan, *Theory of solute transport by groundwater*, in Ann. Rev. Fluid Mech., 19 (1987), pp. 183–215.

[25] G. Dagan, *Flow and Transport in Porous Formations*, Springer-Verlag, New York, 1989.

[26] G. Dagan, *Transport in heterogeneous porous formations: Spatial moments, ergodicity, and effective dispersion*, Water Resour. Res., 26 (1990), pp. 1281–1290.

[27] L. de Arcangelis, J. Koplik, S. Redner, and D. Wilkinson, *Hydrodynamic dispersion in network models of porous media*, Phys. Rev. Lett., 57 (1986), pp. 996–999.

[28] S. Didierjean, H. P. Amaral Souto, R. Delannay, and C. Moyne, *Dispersion in periodic porous media. Experience versus theory for two-dimensional systems*, Chem. Eng. Sci., 52 (1997), pp. 1861–1874.

[29] K. D. Dorfman and H. Brenner, *Generalized Taylor-Aris dispersion in discrete spatially periodic networks: Microfluidic applications*, Phys. Rev. E, 65 (2002), 021103.

[30] F. A. L. Dullien, *Porous Media. Fluid Transport and Pore Structure*, 2nd ed., Academic Press, New York, 1992.

[31] D. A. Edwards, M. Shapiro, H. Brenner, and M. Shapira, *Dispersion of inert solutes in spatially periodic two-dimensional porous media*, Transp. Porous Media, 6 (1991), pp. 337–358.

[32] A. M. M. Elfeki, G. J. M. Uffink, and F. B. J. Barends, *Groundwater Contaminant Transport. Impact of Heterogeneous Characterization: A New View on Dispersion*, A. A. Balkema, Rotterdam, The Netherlands, 1997.

[33] I. Fatt, *The network model of porous media–I. Capillary pressure characteristics*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 144–159.

[34] I. Fatt, *The network model of porous media–II. Dynamic properties of a single size tube network*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 160–163.

[35] I. Fatt, *The network model of porous media–III. Dynamic properties of networks with tube radius distributions*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 164–181.

[36] J. Feyen, D. Jacques, A. Timmerman, and J. Vanderborght, *Modeling water flow and solute transport in heterogeneous soils: A review of recent approaches*, J. Agric. Engng. Res., 70 (1998), pp. 231–256.

[37] J. J. Fried and M. A. Combarnous, *Dispersion in porous media*, Adv. Hydrosci., 7 (1971), pp. 169–282.

[38] G. R. Gavalas and S. Kim, *Periodic capillaries models of diffusion in porous solids*, Chem. Eng. Sci., 36 (1981), pp. 1111–1122.

[39] G. H. Goldsztein, *Transport of nutrients in bones*, SIAM J. Appl. Math., 65 (2005), pp. 2128–2140.

[40] D. J. Gunn and C. Pryce, *Dispersion in packed beds*, Trans. Inst. Chem. Eng., 47 (1969), pp. T342–T350.

[41] F. J. M. Horn, *Calculation of dispersion coefficients by means of moments*, AIChE J., 17 (1971), pp. 613–620.

[42] U. Hornung, *Miscible displacement*, in Homogenization and Porous Media, U. Hornung, ed., Springer, New York, 1997.

[43] M. L. Knothe Tate and U. Knothe, *An ex vivo model to study the transport processes and fluid flow in loaded bone*, J. Biomech., 33 (2000), pp. 247–254.

[44] M. L. Knothe Tate, P. Niederer, and U. Knothe, *In vivo tracer transport through the lacunocanalicular system of rat bone in an environment devoid of mechanical loading*, Bone, 22 (1998), pp. 107–117.

[45] M. L. Knothe Tate, U. Knothe, and P. Niederer, *Experimental elucidation of mechanical load-induced fluid flow and its potential role in bone metabolism and functional adaptation*, Am. J. Med. Sci., 316 (1998), pp. 189–195.

[46] D. L. Koch and J. F. Brady, *Dispersion in fixed beds*, J. Fluid Mech., 154 (1985), pp. 399–427.

[47] S. Koplik, S. Redner, and D. Wilkinson, *Transport and dispersion in random networks with percolation disorder*, Phys. Rev. A (3), 37 (1988), pp. 2619–2636.

[48] R. Mauri, *Dispersion, convection and reaction in porous media*, Phys. Fluids A, 3 (1991), pp. 743–756.

[49] C. C. Mei, *Method of homogenization applied to dispersion in porous media*, Transp. in Porous Media, 9 (1992), pp. 261–274.

[50] K. Piekarski and M. Munro, *Transport mechanism operating between blood supply and osteocytes in long bones*, Nature, 269 (1977), pp. 80–82.

[51] J. Rubinstein and R. Mauri, *Dispersion and convection in periodic porous media*, SIAM J. Appl. Math., 46 (1986), pp. 1018–1023.

[52] P. G. Saffman, *Dispersion flow through a network of capillaries*, Chem. Eng. Sci., 11 (1959), pp. 125–129.

[53] P. G. Saffman, *A theory of dispersion in a porous medium*, J. Fluid Mech., 6 (1959), pp. 321–349.

[54] P. G. Saffman, *Dispersion due to molecular diffusion and macroscopic mixing in flow through a network of capillaries*, J. Fluid Mech., 7 (1960), pp. 194–208.

[55] M. Sahimi, A. A. Heiba, B. D. Hughes, H. T. Davis, and L. E. Scriven, *Dispersion in Flow through Porous Media*, SPE 10969, Soc. Petrol. Engrs., New Orleans, LA, 1982.

[56] M. Sahimi, B. D. Hughes, L. E. Scriven, and H. T. Davis, *Stochastic transport in disordered systems*, J. Chem. Phys., 78 (1983), pp. 6849–6864.

[57] M. Sahimi, H. T. Davis, and L. E. Scriven, *Dispersion in disordered porous media*, Chem. Eng. Commun., 23 (1983), pp. 329–341.

[58] Sahimi, *Flow and Transport in Porous Media and Fractured Rock*, John Wiley, New York, 1995.

[59] J. Salles, J.-F. Thovert, R. Delannay, L. Prevors, J.-L. Auriault, and P. M. Adler, *Taylor dispersion in porous media. Determination of the dispersion tensor*, Phys. Fluids A, 5 (1993), pp. 2348–2376.

[60] M. Sardin and D. Schweich, *Modeling the nonequilibrium transport of linearly interacting solutes in porous media: A review*, Water Resourc. Res. 27 (1991), pp. 2287–2307.

[61] A. E. Scheidegger, *The Physics of Flow through Porous Media*, 2nd ed., Toronto University Press, Toronto, 1963.

[62] V. I. Selyakov and V. V. Kadet, *Percolation Models for Transport in Porous Media*, Kluwer Academic Publishers, Boston, 1996.

[63] G. I. Taylor, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London Ser. A, 223 (1953), pp. 446–468.

[64] S. W. Wheatcraft, G. A. Sharp, and S. W. Tyler, *Fluid flow and solute transport in fractal heterogeneous porous media*, in Dynamics of Fluid in Hierarchical Porous Media, J. H. Cushman, ed., Academic Press, New York, 1990, pp. 305–326.

[65] L. Wang, S. C. Cowin, S. Weinbaum, and S. P. Fritton, *Modeling tracer transport in an osteon under cyclic loading*, Ann. Biomed. Eng., 28 (2000), pp. 1200–1209.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1223–1246

DIFFUSION OF PROTEIN RECEPTORS ON A CYLINDRICAL DENDRITIC MEMBRANE WITH PARTIALLY ABSORBING TRAPS[*]

PAUL C. BRESSLOFF[†], BERTON A. EARNSHAW[†], AND MICHAEL J. WARD[‡]

Abstract. We present a model of protein receptor trafficking within the membrane of a cylindrical dendrite containing small protrusions called spines. Spines are the locus of most excitatory synapses in the central nervous system and act as localized traps for receptors diffusing within the dendritic membrane. We treat the transverse intersection of a spine and dendrite as a spatially extended, partially absorbing boundary and use singular perturbation theory to analyze the steady-state distribution of receptors. We compare the singular perturbation solutions with numerical solutions of the full model and with solutions of a reduced one-dimensional model and find good agreement between them all. We also derive a system of Fokker–Planck equations from our model and use it to exactly solve a mean first passage time (MFPT) problem for a single receptor traveling a fixed axial distance along the dendrite. This is then used to calculate an effective diffusion coefficient for receptors when spines are uniformly distributed along the length of the cable and to show how a nonuniform distribution of spines gives rise to anomalous subdiffusion.

Key words. protein receptor trafficking, diffusion-trapping, matched asymptotics, singular perturbation theory

AMS subject classification. 92C20

DOI. 10.1137/070698373

1. Introduction.

Neurons are amongst the largest and most complex cells in biology. Their intricate geometry presents many challenges for cell function, in particular with regard to the efficient delivery of newly synthesized proteins from the cell body or soma to distant locations on the axon or dendrites. The axon contains ion channels for action potential propagation and presynaptic active zones for neurotransmitter release, whereas each dendrite contains postsynaptic domains (or densities) where receptors that bind neurotransmitter tend to cluster. At most excitatory synapses in the brain, the postsynaptic density is located within a dendritic spine, which is a small, submicrometer membranous extrusion that protrudes from a dendrite. Typically spines have a bulbous head which is connected to the parent dendrite through a thin spine neck. Given that hundreds or thousands of synapses and their associated spines are distributed along the entire length of a dendrite, it follows that neurons must traffic receptors and other postsynaptic proteins over long distances (several 100μm) from the soma. This can occur by two distinct mechanisms: either by lateral diffusion in the plasma membrane [8, 26, 1, 7] or by motor-driven intracellular transport along microtubules followed by local insertion into the surface membrane (exocytosis) [17, 20, 29]. It is likely that both forms of transport occur in dendrites, depending on the type of receptor and the developmental stage of the organism.

[*] Received by the editors July 25, 2007; accepted for publication (in revised form) December 26, 2007; published electronically April 2, 2008. This work was supported by the NSF (DMS 0515725 and RTG 0354259).
http://www.siam.org/journals/siap/68-5/69837.html

[†] Department of Mathematics, University of Utah, 155 S. 1400 E., Salt Lake City, UT 84112 (bressloff@math.utah.edu, earnshaw@math.utah.edu).

[‡] Department of Mathematics, University of British Columbia, Vancouver, BC, V6T1Z2, Canada (ward@math.ubc.ca).

Recently, we constructed a one-dimensional diffusion-trapping model for the surface transport of AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole-propionic acid) receptors along a dendrite [6]. AMPA receptors respond to the neurotransmitter glutamate and mediate the majority of fast excitatory synaptic transmission in the central nervous system. Moreover, there is a large body of experimental evidence suggesting that the fast trafficking of AMPA receptors into and out of spines contributes to activity-dependent, long-lasting changes in synaptic strength [21, 22, 4, 8, 9, 16]. Single-particle tracking experiments suggest that surface AMPA receptors diffuse freely within the dendritic membrane until they encounter a spine [13, 26]. If a receptor flows into a spine, then it is temporarily confined by the geometry of the spine and through interactions with scaffolding proteins and cytoskeletal elements [4, 8]. A surface receptor may also be internalized via endocytosis and stored within an intracellular pool, where it is either recycled to the surface via exocytosis or degraded [11]. Motivated by these experimental observations, we modeled the surface transport of receptors along a dendrite as a process of diffusion in the presence of spatially localized, partially absorbing traps [6]. One of the major simplifications of our model was to reduce the cylindrical like surface of a dendrite to a one-dimensional domain by neglecting variations in receptor concentration around the circumference of the cable relative to those along the cable. We also neglected the spatial extent of each spine by treating it as a homogeneous compartment that acts as a point-like source/sink for receptors on the dendrite. This was motivated by the observation that the spine neck, which forms the junction between a synapse and its parent dendrite, varies in radius from 0.02–0.2μm [23]. This is typically an order of magnitude smaller than the spacing between neighboring spines and the radius of the dendritic cable (around 1μm). In the one-dimensional case, the introduction of point-like spines does not lead to any singularities since the associated one-dimensional Green's function for diffusion is a pointwise bounded function. We were thus able to calculate explicitly the steady-state distribution of receptors along the dendrite and spines, as well as to determine the mean first passage time (MFPT) for a receptor to reach a certain distance from the soma. This allowed us to investigate the efficacy of diffusive transport as a function of various biophysical parameters such as surface diffusivity and the rates of exo/endocytosis within each spine.

In this paper we extend our diffusion-trapping model to the more realistic case of a two-dimensional cylindrical surface. However, since the two-dimensional Green's function has logarithmic singularities, we can no longer neglect the spatial extent of a spine. Therefore, we proceed by solving the steady-state diffusion equation on a finite cylindrical surface containing a set of small, partially absorbing holes, which represent the transverse intersections of the spines with the dendrite. The solution is constructed by matching appropriate "inner" and "outer" asymptotic expansions [27, 25, 28, 24]. This leads to a system of linear equations that determines the dendritic receptor concentration on the boundary between the dendrite and each spine. We numerically solve these equations and use this to construct the steady-state distribution of receptors along the dendrite. We compare our results with numerical solutions of the full model and with a reduced one-dimensional model.

A brief outline of this paper is as follows: In section 2 we formulate our diffusion-trapping model for receptor trafficking on the boundary of a cylindrical dendritic cable. In section 3 we construct the steady-state solution to this model using singular perturbation techniques in the limit of small spine radii. In section 4 we present some numerical experiments for realistic physiological parameter values that compare our asymptotic solution of section 3 with both full numerical solutions and with the solution of a one-dimensional approximation valid for a large aspect ratio dendritic cable. Finally, in section 5 we asymptotically calculate the MFPT for the diffusion of a single tagged receptor.

2. Diffusion-trapping model on a cylinder. Consider a population of N dendritic spines distributed along a cylindrical dendritic cable of length L and radius l as shown in Figure 2.1(A). Since protein receptors are much smaller than the length and circumference of the cylinder, we can neglect the extrinsic curvature of the membrane. Therefore, as shown in Figure 2.1(B), we represent the cylindrical surface of the dendrite as a long rectangular domain Ω_0 of width $2\pi l$ and length L so that

$$\Omega_0 \equiv \{(x,y) : 0 < x < L\,, |y| < \pi l\}.$$

The cylindrical topology is preserved by imposing periodic boundary conditions along the circumference of the cylinder, that is, at $y = \pm\pi l$. At one end of the cylinder $(x = 0)$ we impose a nonzero flux boundary condition, which represents a constant source of newly synthesized receptors from the soma, and at the other end $(x = L)$ we impose a no-flux boundary condition. Each spine neck is assumed to intersect the dendritic surface transversely such that the intersection is a circle of radius $\varepsilon\rho$ centered about the point $\mathbf{r}_j = (x_j, y_j) \in \Omega_0$, where $j = 1, \dots, N$ labels the jth spine. For simplicity, we take all spines to have the same radius. Since a dendrite is usually several hundred μm in length, we will assume the separation of length scales $\varepsilon\rho \ll 2\pi l \ll L$. We then fix the units of length by setting $\rho = 1$ such that $2\pi l = \mathcal{O}(1)$ and treat ε as a small dimensionless parameter. Finally, we denote the surface of the cylinder excluding the small discs arising from the spines by Ω_ε so that

$$\Omega_\varepsilon = \Omega_0 \setminus \bigcup_{j=1}^{N} \Omega_j, \quad \Omega_j = \{\mathbf{r} : |\mathbf{r} - \mathbf{r}_j| \le \varepsilon\}.$$

Let $U(\mathbf{r}, t)$ denote the concentration of surface receptors within the dendritic membrane at position $\mathbf{r} \in \Omega_\varepsilon$ at time $t \in \mathbf{R}_+$. As a result of the small area of each spine, we assume that the receptor concentrations within each spine are spatially homogeneous. We let $R_j(t)$ denote the concentration of surface receptors in the jth spine. The dendritic surface receptor concentration evolves according to the diffusion equation

$$\frac{\partial U}{\partial t} = D\nabla^2 U, \quad (\mathbf{r}, t) \in \Omega_\varepsilon \times \mathbf{R}_+ \tag{2.1}$$

for a homogeneous surface diffusivity D, with periodic boundary conditions at the ends $y = \pm\pi l$,

$$U(x, \pi l, t) = U(x, -\pi l, t), \quad \partial_y U(x, \pi l, t) = \partial_y U(x, -\pi l, t), \tag{2.2}$$

and nonzero or zero flux conditions at the ends $x = 0, L$,

$$\partial_x U(0, y, t) = -\sigma \equiv -\frac{\sigma_0}{2\pi l D}\,, \quad \partial_x U(L, y, t) = 0. \tag{2.3}$$

Here σ_0 denotes the number of receptors per unit time entering the surface of the cylinder from the soma. At each interior boundary $\partial\Omega_j$ we impose the mixed boundary condition

$$\varepsilon\partial_n U(\mathbf{r}, t) = -\frac{\omega_j}{2\pi D}(U(\mathbf{r}, t) - R_j), \quad \mathbf{r} \in \partial\Omega_j, \quad j = 1, \dots, N, \tag{2.4}$$

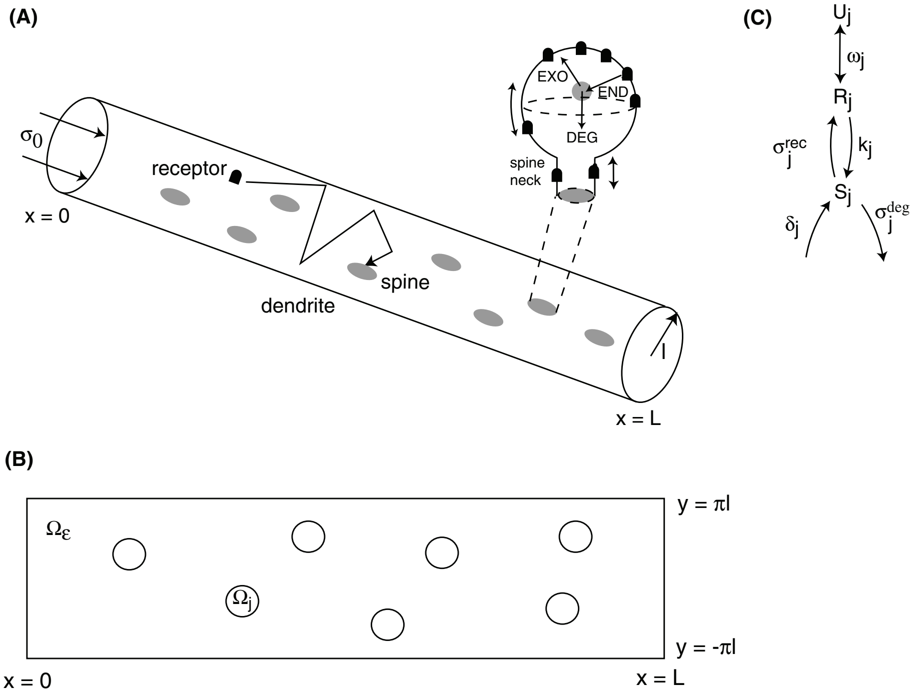


Fig. 2.1. *Diffusion-trapping model of receptor trafficking on a cylindrical dendritic cable (diagram not to scale).* (A) *A population of dendritic spines is distributed on the surface of a cylinder of length L and radius l. Each receptor diffuses freely until it encounters a spine where it may become trapped. Within a spine receptors may be internalized via endocytosis (END) and then either recycled to the surface via exocytosis (EXO) or degraded (DEG); see the inset. Synthesis of new receptors at the soma and insertion into the plasma membrane generates a surface flux σ_0 at one end of the cable.* (B) *Topologically equivalent rectangular domain with opposite sides $y = \pm\pi l$ identified.* (C) *State transition diagram for a simplified one-compartment model of a dendritic spine. Here R_j denotes the concentration of surface receptors inside the jth spine, U_j is the mean dendritic receptor concentration on the boundary between the spine neck and dendrite, and S_j is the number of receptors within the corresponding intracellular pool. Freely diffusing surface receptors can enter/exit the spine at a hopping rate ω_j, be endocytosed at a rate k_j, be exocytosed at a rate σ_j^{rec}, and be degraded at a rate σ_j^{deg}. New intracellular receptors are produced at a rate δ_j.*

where $\partial_n U$ is the outward normal derivative to Ω_ε. The flux of receptors across the boundary between the dendrite and jth spine is taken to depend on the difference in concentrations on either side of the boundary with ω_j an effective hopping rate. (This rate is determined by the detailed geometry of the spine [2].) It follows that the total number of receptors crossing the boundary per unit time is $\omega_j[U_j(t) - R_j(t)]$, where $U_j(t)$ is the mean dendritic receptor concentration on the boundary $\partial\Omega_j$ of the jth spine of length $2\pi\varepsilon$:

$$U_j = \frac{1}{2\pi\varepsilon}\int_{\partial\Omega_j} U(\mathbf{r}, t)\, d\mathbf{r}. \tag{2.5}$$

Surface receptors within the jth spine can be endocytosed at a rate k_j and stored in an intracellular pool. Intracellular receptors are either reinserted into the surface via exocytosis at a rate σ_j^{rec} or degraded at a rate σ_j^{deg}. We also allow for a local source of intracellular receptors with a production rate δ_j. Denoting the number of receptors in the jth intracellular pool by $S_j(t)$, we then have the pair of equations

$$\frac{dR_j}{dt} = \frac{\omega_j}{A_j}[U_j - R_j] - \frac{k_j}{A_j}R_j + \frac{\sigma_j^{rec}S_j}{A_j}\,, \tag{2.6}$$

$$\frac{dS_j}{dt} = -\sigma_j^{rec}S_j - \sigma_j^{deg}S_j + k_jR_j + \delta_j. \tag{2.7}$$

The first term on the right-hand side of (2.6) represents the exchange of surface receptors between the spine and parent dendrite. Since $\omega_j[U_j - R_j]$ is the number of receptors per unit time flowing across the junction between the dendritic cable and the spine, it is necessary to divide through by the surface area A_j of the spine in order to properly conserve receptor numbers. Note that in our previous one-dimensional model [6] we absorbed the factor of A_j into our definition of the rate of endocytosis k_j. (The precise variation of endocytic rate with spine area A_j will depend upon whether or not endocytosis is localized to certain hotspots within the spine [4].) The various processes described by (2.6) and (2.7) are summarized in Figure 2.1(C).

3. Steady-state analysis using asymptotic matching. In steady-state one can solve (2.6) and (2.7) for R_j in terms of the mean concentration U_j to get

$$R_j = \frac{\omega_jU_j}{\omega_j + k_j(1-\lambda_j)} + \frac{\lambda_j\delta_j}{\omega_j + k_j(1-\lambda_j)}, \tag{3.1}$$

where

$$\lambda_j \equiv \frac{\sigma_j^{rec}}{\sigma_j^{rec} + \sigma_j^{deg}}, \quad S_j = \frac{\lambda_j}{\sigma_j^{rec}}\left(k_jR_j + \delta_j\right). \tag{3.2}$$

Then U_j is determined from (2.5) and the steady-state version of (2.1):

$$\nabla^2 U = 0\,, \quad \mathbf{r} \in \Omega_\varepsilon, \tag{3.3}$$

with boundary conditions

$$U(x,\pi l) = U(x,-\pi l)\,, \quad \partial_y U(x,\pi l) = \partial_y U(x,-\pi l), \tag{3.4}$$

$$\partial_x U(0,y) = -\sigma, \quad \partial_x U(L,y) = 0, \tag{3.5}$$

where $\sigma = \sigma_0/(2\pi l D)$ from (2.3) and

$$\varepsilon\partial_n U(\mathbf{r}) = -\frac{\omega_j}{2\pi D}(U(\mathbf{r}) - R_j), \quad \mathbf{r} \in \partial\Omega_j\,, \quad j = 1,\ldots,N. \tag{3.6}$$

Since the radius of each spine is asymptotically small, we can make the simplification that $U(\mathbf{r}) = U_j$ on $\partial\Omega_j$. The substitution of (3.1) into (3.6) then yields the following reduced condition on the boundary of each spine:

$$\varepsilon\partial_n U(\mathbf{r}) = -\frac{\widehat{\omega}_j}{2\pi D}(U_j - \widehat{R}_j)\,, \quad \mathbf{r} \in \partial\Omega_j\,, \quad j = 1,\ldots,N, \tag{3.7}$$

where $\widehat{\omega}_j$ and $\widehat{R}_j$ are defined by

$$\widehat{\omega}_j \equiv \frac{\omega_j k_j(1-\lambda_j)}{\omega_j + k_j(1-\lambda_j)}\,, \qquad \widehat{R}_j \equiv \frac{\sigma_j^{rec}}{k_j}\frac{\delta_j}{\sigma_j^{deg}}. \tag{3.8}$$

Under steady-state conditions, one can view $\widehat{\omega}_j$ as an effective spine-neck hopping rate and $\widehat{R}_j$ as an effective receptor concentration within the spine.

Upon integrating the diffusion equation (3.3) over the domain Ω_ε and imposing the boundary conditions (3.4), (3.5), and (3.7), we obtain the solvability condition

$$\sigma_0 = \sum_{j=1}^{N} \widehat{\omega}_j \left[U_j - \widehat{R}_j\right]. \tag{3.9}$$

This expresses the condition that the rate at which receptors enter the dendrite from the soma is equal to the effective rate at which receptors exit the dendrite into spines and are degraded. Note that in the limit of negligible degradation of receptors in the intracellular pools so that $\sigma_j^{deg} \to 0$, it follows from (3.2) and (3.8) that $\lambda_j \to 1$, $\widehat{\omega}_j \to 0$, $\widehat{R}_j \to \infty$ with $\widehat{\omega}_j\widehat{R}_j \to -\delta_j$. Consequently, in the limit $\sigma_j^{deg} \to 0$, (3.9) reduces to $\sigma_0 \to -\sum_{j=1}\delta_j$, which has no solution for $\sigma_0 > 0$ and positive production rates $\delta_j > 0$. This shows that there is no steady-state solution when $\sigma_j^{deg} = 0$, as the number of receptors in the dendrite would grow without bound as time increases. A similar argument shows that there is also no steady-state solution in the limit of infinite spine-neck resistances such that $\omega_j \to 0$ for $j = 1, \ldots, N$. In this limit, newly synthesized receptors at the soma would not be able to diffuse from the dendrite to a spine and be degraded.

Our method of solution for the boundary value problem given by (3.3), (3.4), (3.5), and (3.7), which we denote by BVPI, proceeds in two steps. First, we solve a related problem, denoted by BVPII, in which the mixed boundary conditions (3.7) are replaced by the inhomogeneous Dirichlet conditions

$$U(\mathbf{r}) = U_j\,, \quad \mathbf{r} \in \partial\Omega_j, \quad j = 1, \ldots, N, \tag{3.10}$$

under the assumption that the U_j are known. In the singularly perturbed limit $\varepsilon \to 0$, the approximate solution for U valid away from each of the spines is shown below to be determined up to an arbitrary constant χ. Then, by substituting our asymptotic solution to BVPII into the N mixed boundary conditions (3.7) and upon satisfying the conservation equation (3.9), we obtain $N+1$ linear equations for the $N+1$ unknowns χ and $U_j, j = 1, \ldots, N$. This closed linear system of equations can be solved numerically to generate the full solution to the original boundary value problem BVPI. In order to solve BVPII asymptotically in the limit of small spine radii $\varepsilon \to 0$, we match appropriate "inner" and "outer" asymptotic expansions, following along similar lines to previous studies of boundary value problems in domains with small holes [27, 25, 28, 24].

3.1. Matching inner and outer solutions. Around each small circle $\partial\Omega_j$ we expect the solution of BVPII to develop a boundary layer where it changes rapidly from its value U_j on $\partial\Omega_j$ to another value that is required by the solution to the steady-state diffusion equation in the bulk of the domain. Therefore, Ω_ε may be decomposed into a set of $j = 1, \ldots, N$ "inner" regions, where $|\mathbf{r} - \mathbf{r}_j| = \mathcal{O}(\varepsilon)$, and an "outer" region, where $|\mathbf{r} - \mathbf{r}_j| \gg \mathcal{O}(\varepsilon)$ for all $j = 1, \ldots, N$. In the jth inner region, we introduce the stretched local variable $\mathbf{s} = \varepsilon^{-1}(\mathbf{r} - \mathbf{r}_j)$ and set $V(\mathbf{s};\varepsilon) = U(\mathbf{r}_j + \varepsilon\mathbf{s};\varepsilon)$ so that to leading order (omitting far-field boundary conditions)

$$\begin{aligned} \nabla_{\mathbf{s}}^2 V &= 0\,, \quad |\mathbf{s}| > 1, \\ V &= U_j\,, \quad |\mathbf{s}| = 1. \end{aligned} \tag{3.11}$$

This has an exact solution of the form $V = U_j + B_j \log|\mathbf{s}|$ with the unknown amplitude B_j determined by matching inner and outer solutions. This leads to an infinite logarithmic expansion of B_j in terms of the small parameter [27, 25, 28, 24]

$$\nu = -\frac{1}{\log(\varepsilon)}. \tag{3.12}$$

Since the outer solution is $\mathcal{O}(1)$ as $\nu \to 0$ and V grows logarithmically at infinity, we write $B_j = \nu A_j(\nu)$, where the function $A_j(\nu)$ is to be found. The inner solution becomes

$$V = U_j + \nu A_j(\nu) \log|\mathbf{s}|. \tag{3.13}$$

In terms of the outer variable $|\mathbf{r} - \mathbf{r}_j|$, we then obtain the following far-field behavior of the inner solution:

$$V \sim U_j + A_j(\nu) + \nu A_j(\nu) \log|\mathbf{r} - \mathbf{r}_j|. \tag{3.14}$$

This far-field behavior must then match with the near-field behavior of the asymptotic expansion of the solution in the outer region away from the N traps. The corresponding outer problem is given by

$$\nabla^2 U = 0, \quad \mathbf{r} \in \Omega_0 \backslash \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}, \tag{3.15}$$

with boundary conditions

$$\begin{aligned} &U(x, \pi l) = U(x, -\pi l)\,, \quad \partial_y U(x, \pi l) = \partial_y U(x, -\pi l), \\ &\partial_x U(0, y) = -\sigma, \quad \partial_x U(L, y) = 0, \end{aligned}$$

together with the asymptotic singularity conditions

$$U \sim U_j + A_j(\nu) + \nu A_j(\nu) \log|\mathbf{r} - \mathbf{r}_j| \quad \text{as } \mathbf{r} \to \mathbf{r}_j\,, \quad j = 1, \ldots, N. \tag{3.16}$$

Equations (3.15) and (3.16) can be reformulated in terms of an outer problem with homogeneous boundary conditions and a constant forcing term by decomposing

$$U(\mathbf{r}) = \mathcal{U}(\mathbf{r}) + u(\mathbf{r})\,, \quad u(x, y) \equiv \frac{\sigma}{2L}(x - L)^2. \tag{3.17}$$

Then (3.15) becomes

$$\nabla^2 \mathcal{U} = -\frac{\sigma}{L}\,, \quad \mathbf{r} \in \Omega_0 \backslash \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}\,, \tag{3.18}$$

with boundary conditions

$$\begin{aligned} &\mathcal{U}(x, \pi l) = \mathcal{U}(x, -\pi l)\,, \quad \partial_y \mathcal{U}(x, \pi l) = \partial_y \mathcal{U}(x, -\pi l), \\ &\partial_x \mathcal{U}(0, y) = 0, \quad \partial_x \mathcal{U}(L, y) = 0\,, \end{aligned}$$

and the asymptotic singularity conditions

$$\mathcal{U} \sim -u(\mathbf{r}_j) + U_j + A_j(\nu) + \nu A_j(\nu) \log|\mathbf{r} - \mathbf{r}_j| \quad \text{as } \mathbf{r} \to \mathbf{r}_j\,, \quad j = 1, \ldots, N. \tag{3.19}$$

In order to treat the logarithmic behavior of the outer solution at $\mathbf{r}_j$, we introduce the Neumann Green's function $G(\mathbf{r}; \mathbf{r}')$, defined as the unique solution to

$$\begin{gathered} \nabla^2 G = \frac{1}{|\Omega_0|} - \delta(\mathbf{r} - \mathbf{r}')\,, \quad \mathbf{r} \in \Omega_0\,, \\ G(x, \pi l; \mathbf{r}') = G(x, -\pi l; \mathbf{r}')\,, \quad \partial_y G(x, \pi l; \mathbf{r}') = \partial_y G(x, -\pi l; \mathbf{r}'), \\ \partial_x G(0, y; \mathbf{r}') = 0\,, \quad \partial_x G(L, y; \mathbf{r}') = 0\,, \\ \int_{\Omega_0} G(\mathbf{r}; \mathbf{r}')\, d\mathbf{r} = 0. \end{gathered} \tag{3.20}$$

Here $|\Omega_0| = 2\pi L l$ is the area of the rectangular domain Ω_0. This Green's function has a logarithmic singularity as $\mathbf{r} \to \mathbf{r}'$ so that we can decompose G as

$$G(\mathbf{r}; \mathbf{r}') = -\frac{1}{2\pi} \log|\mathbf{r} - \mathbf{r}'| + \mathcal{G}(\mathbf{r}; \mathbf{r}'), \tag{3.21}$$

where $\mathcal{G}$ is the regular part of G. We will calculate $\mathcal{G}$ explicitly in section 3.3. This property of G suggests that we replace (3.18) and (3.19) by the following single equation in Ω_0:

$$\nabla^2 \mathcal{U} = -\frac{\sigma}{L} + \sum_{j=1}^{N} 2\pi\nu A_j(\nu) \delta(\mathbf{r} - \mathbf{r}_j)\,, \quad \mathbf{r} \in \Omega_0. \tag{3.22}$$

Then, upon using the divergence theorem together with the homogeneous boundary conditions for $\mathcal{U}$, we obtain the solvability condition

$$\frac{\sigma}{L} |\Omega_0| = \sum_{j=1}^{N} 2\pi\nu A_j(\nu). \tag{3.23}$$

It readily follows that (3.22) has the solution

$$\mathcal{U}(\mathbf{r}) = -\sum_{j=1}^{N} 2\pi\nu A_j(\nu) G(\mathbf{r}; \mathbf{r}_j) + \chi, \tag{3.24}$$

where χ is a constant to be found. Since $\int_{\Omega_0} G\, d\mathbf{r} = 0$, it follows that χ can be interpreted as the spatial average of $\mathcal{U}$, defined by $\chi = |\Omega_0|^{-1} \int_{\Omega_0} \mathcal{U}\, d\mathbf{r}$. Then, as $\mathbf{r} \to \mathbf{r}_j$, the outer solution for $\mathcal{U}$ given in (3.24) has the near-field behavior

$$\mathcal{U} \sim -2\pi\nu A_j(\nu) \left[-\frac{1}{2\pi} \log|\mathbf{r} - \mathbf{r}_j| + \mathcal{G}(\mathbf{r}_j; \mathbf{r}_j) \right] - \sum_{i \neq j}^{N} 2\pi\nu A_i(\nu) G(\mathbf{r}_j; \mathbf{r}_i) + \chi \tag{3.25}$$

for each $j = 1, \ldots, N$. Upon comparing the nonsingular terms in this expression and that of (3.19), we obtain the following system of equations:

$$(1 + 2\pi\nu \mathcal{G}_{jj}) A_j + \sum_{i \neq j}^{N} 2\pi\nu G_{ji} A_i = u_j - U_j + \chi\,, \quad j = 1\,, \ldots, N\,, \tag{3.26}$$

where $u_j \equiv u(\mathbf{r}_j)$, $G_{ji} \equiv G(\mathbf{r}_j; \mathbf{r}_i)$, and $\mathcal{G}_{jj} \equiv \mathcal{G}(\mathbf{r}_j; \mathbf{r}_j)$.

3.2. Calculation of boundary concentrations U_j. Equations (3.13) and (3.24) are the inner and outer solutions of BVPII, respectively, where the $N+1$ coefficients χ and A_j for $j=1,\dots,N$ are determined from the N linear equations (3.26) together with the solvability condition (3.23). We can now generate the solution to the original problem BVPI by substituting the inner solution (3.13) into the mixed boundary conditions (3.7). This gives

$$2\pi\nu A_j(\nu) = \frac{\widehat{\omega}_j}{D}[U_j - \widehat{R}_j] \equiv V_j. \tag{3.27}$$

Substituting (3.27) into the solvability condition (3.23) shows that the latter is equivalent to the conservation equation (3.9). Furthermore, upon substituting (3.27) into (3.26) we obtain the system of linear equations

$$\left[(2\pi\nu)^{-1} + \mathcal{G}_{jj}\right]\frac{\widehat{\omega}_j}{D}[U_j - \widehat{R}_j] + \sum_{i\neq j}^{N} G_{ji}\frac{\widehat{\omega}_i}{D}[U_i - \widehat{R}_i] = u_j - U_j + \chi\,, \quad j=1,\dots,N. \tag{3.28}$$

This system, together with the conservation equation (3.9), gives $N+1$ equations for the $N+1$ unknowns χ and $U_j, j=1,\dots,N$. This system depends on the flux σ_0 of receptors from the soma, the number and the locations of the dendritic spines, and the aspect ratio of Ω_0. Upon solving this system for U_j and χ, the dendritic receptor concentration in the bulk of the dendritic membrane, obtained from (3.17), (3.24), and (3.27), is given by

$$U(\mathbf{r}) = u(\mathbf{r}) - \sum_{j=1}^{N}\frac{\widehat{\omega}_j}{D}[U_j - \widehat{R}_j]G(\mathbf{r};\mathbf{r}_j) + \chi. \tag{3.29}$$

This approximate solution for U is valid for distances larger than $\mathcal{O}(\varepsilon)$ away from the centers $\mathbf{r}_j$, for $j=1,\dots,N$, of the spines. Moreover, the distribution R_j of receptors within the spines is given in terms of U_j by (3.1).

There are two important remarks. First, we emphasize that the system (3.26) together with (3.23) contains all of the logarithmic correction terms in the asymptotic solution to BVPI. The error made in this approximation is transcendentally small of order $\mathcal{O}(\varepsilon)$, which is asymptotically smaller than any power of ν. A precise estimate of such transcendentally small terms for a related problem is given in Appendix A of [25]. Second, we remark that in [6] we have previously derived one-dimensional versions of (3.9), (3.28), and (3.29). In contrast to the two-dimensional case, the one-dimensional Neumann Green's function is nonsingular so that one can represent the spines as point sources/sinks on the dendrite, and singular perturbation theory is not needed.

It is convenient to introduce a matrix solution to (3.28). We first introduce the matrix $\mathbf{B}$ with elements

$$B_{jj} = 2\pi\left(\frac{D}{\widehat{\omega}_j} + \mathcal{G}_{jj}\right)\,, \quad B_{ji} = 2\pi G_{ji}\,,\ j\neq i. \tag{3.30}$$

Then (3.28) can be written in the compact form

$$\sum_{i=1}^{N}(\delta_{i,j} + \nu B_{ji})V_i = 2\pi\nu[u_j - \widehat{R}_j + \chi]\,,$$

where V_i is defined in (3.27). Defining $\mathbf{M} = (\mathbf{I} + \nu\mathbf{B})^{-1}$, where $\mathbf{I}$ is the $N \times N$ identity matrix, we have

$$V_j = 2\pi\nu \sum_{i=1}^{N} M_{ji}(u_i - \widehat{R}_i + \chi). \tag{3.31}$$

The constant χ is then determined by substituting (3.31) into the solvability condition (3.9). This yields

$$\frac{\sigma_0}{D} = \sum_{j=1}^{N} V_j = 2\pi\nu \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ji}(u_i - \widehat{R}_i + \chi).$$

Upon solving this equation for χ we get

$$\chi = \frac{\frac{\sigma_0}{2\pi\nu D} - \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ji}(u_i - \widehat{R}_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} M_{ji}}. \tag{3.32}$$

Since $M_{ji} = \delta_{i,j} + \mathcal{O}(\nu)$, it follows that to leading order in ν

$$\chi = \frac{\sigma_0}{2\pi N D\nu} + \mathcal{O}(1), \quad U_j = \widehat{R}_j + \frac{\sigma_0}{N\widehat{\omega}_j} + \mathcal{O}(\nu). \tag{3.33}$$

The singular nature of the constant χ as $\nu \to 0$, and hence the solution $U(\mathbf{r})$, reflects the fact that for fixed somatic flux σ_0, the flux in the neighborhood of each spine boundary $\partial\Omega_j$ diverges as $\varepsilon \to 0$. This is necessary in order to maintain the solvability condition (3.9). Note, in particular, that $\widehat{\omega}_j[U_j - \widehat{R}_j]$ gives the number of receptors flowing across the boundary per unit time, and its size essentially remains fixed as ε decreases. Thus, in this limit, the flux through the boundary increases, resulting in a steeper concentration gradient in a neighborhood of the spine boundary. If the hopping rate $\widehat{\omega}_j$ decreases as ε decreases, then the boundary concentration U_j will also diverge in order to maintain (3.9).

3.3. Evaluation of Green's function. To evaluate the Green's function G satisfying (3.20), we begin by writing its Fourier series representation,

$$\begin{aligned} G(\mathbf{r};\mathbf{r}') = {} & \frac{2}{|\Omega_0|}\sum_{n=1}^{\infty} \frac{\cos\left(\frac{\pi n x}{L}\right)\cos\left(\frac{\pi n x'}{L}\right)}{\left(\frac{\pi n}{L}\right)^2} + \frac{2}{|\Omega_0|}\sum_{m=1}^{\infty} \frac{\cos\left(\frac{m(y-y')}{l}\right)}{\left(\frac{m}{l}\right)^2} \\ & + \frac{4}{|\Omega_0|}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \frac{\cos\left(\frac{\pi n x}{L}\right)\cos\left(\frac{\pi n x'}{L}\right)\cos\left(\frac{m(y-y')}{l}\right)}{\left(\frac{\pi n}{L}\right)^2 + \left(\frac{m}{l}\right)^2}. \end{aligned} \tag{3.34}$$

Upon recalling the formula (cf. p. 46 of [12])

$$\sum_{k=1}^{\infty} \frac{\cos(k\theta)}{k^2 + b^2} = \frac{\pi}{2b}\frac{\cosh(b(\pi - |\theta|))}{\sinh(\pi b)} - \frac{1}{2b^2}, \quad |\theta| \leq 2\pi\,, \tag{3.35}$$

we can sum the third term of (3.34) over the index n to obtain

$$\frac{1}{2\pi}\sum_{m=1}^{\infty}\frac{\cos\left(\frac{m(y-y')}{l}\right)\left[\cosh\left(\frac{m(L-|x-x'|)}{l}\right)+\cosh\left(\frac{m(L-|x+x'|)}{l}\right)\right]}{m\sinh\left(\frac{Lm}{l}\right)} - \frac{2}{|\Omega_0|}\sum_{m=1}^{\infty}\frac{\cos\left(\frac{m(y-y')}{l}\right)}{\left(\frac{m}{l}\right)^2}. \tag{3.36}$$

Notice that the second sum of (3.36) cancels the second sum of (3.34). Then, using the angle addition formula for hyperbolic cosine and the relation $\cosh(x)-\sinh(x)=\mathrm{e}^{-x}$, we derive the following key identity for any constants a, b, and c:

$$\frac{\cosh(a-b)+\cosh(a-c)}{\sinh a} = \frac{1}{1-\mathrm{e}^{-2a}}\left[\mathrm{e}^{-b}+\mathrm{e}^{-c}+\mathrm{e}^{b-2a}+\mathrm{e}^{c-2a}\right]. \tag{3.37}$$

We use this identity to rewrite the first sum in (3.36) and then substitute the resulting expression into (3.34). This yields

$$G(\mathbf{r};\mathbf{r}') = \frac{H(x;x')}{2\pi l} + \sum_{m=1}^{\infty}\frac{\left(z_+^m+\overline{z_+}^m+z_-^m+\overline{z_-}^m+\zeta_+^m+\overline{\zeta_+}^m+\zeta_-^m+\overline{\zeta_-}^m\right)}{4\pi m(1-q^m)}, \tag{3.38}$$

where $q\equiv \mathrm{e}^{-2L/l}$. Here $z_\pm$ and $\zeta_\pm$ are defined by $z_\pm \equiv \mathrm{e}^{r_\pm/l}$ and $\zeta_\pm\equiv \mathrm{e}^{\rho_\pm/l}$, where

$$r_+ \equiv -|x+x'|+i(y-y'), \quad r_-\equiv -|x-x'|+i(y-y'), \tag{3.39}$$

$$\rho_+ \equiv |x+x'|-2L+i(y-y'), \quad \rho_-\equiv |x-x'|-2L+i(y-y'), \tag{3.40}$$

and $\bar{\cdot}$ denotes complex conjugate. Moreover, in (3.38), $H(x;x')$ is defined by

$$H(x;x') \equiv \frac{2}{L}\sum_{n=1}^{\infty}\frac{\cos\left(\frac{\pi n x}{L}\right)\cos\left(\frac{\pi n x'}{L}\right)}{\left(\frac{\pi n}{L}\right)^2} = \frac{L}{12}\left[h\left(\frac{x-x'}{L}\right)+h\left(\frac{x+x'}{L}\right)\right], \quad h(\theta)\equiv 3\theta^2-6|\theta|+2. \tag{3.41}$$

The function $H(x;x')$ is the one-dimensional Green's function in the x-direction.

Since $q=\mathrm{e}^{-2L/l}<1$, we can write $(1-q^m)^{-1}=\sum_{n=0}^{\infty}(q^m)^n$ for all $m\ge 1$. In this way, the sum in (3.38) can be written as

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty}(q^n)^m\frac{\left(z_+^m+\overline{z_+}^m+z_-^m+\overline{z_-}^m+\zeta_+^m+\overline{\zeta_+}^m+\zeta_-^m+\overline{\zeta_-}^m\right)}{4\pi m}. \tag{3.42}$$

Notice that when $z_\pm\neq 1$ and $\zeta_\pm \neq 1$ (i.e., when $r_\pm\neq 0$ and $\rho_\pm\neq 0$) this double sum is absolutely convergent, so we can interchange the order of summation in (3.42) and then perform the sum over the index m, yielding

$$-\frac{1}{4\pi}\sum_{n=0}^{\infty}\left(\ln|1-q^n z_+|^2+\ln|1-q^n z_-|^2+\ln|1-q^n\zeta_+|^2+\ln|1-q^n\zeta_-|^2\right) = -\frac{1}{2\pi}\ln|1-z_+||1-z_-||1-\zeta_+||1-\zeta_-|+\mathcal{O}(q). \tag{3.43}$$

The only singularity exhibited by (3.43) in Ω_0 is at $(x, y) = (x', y')$, in which case $z_- = 1$ and $\ln|1 - z_-|$ diverges. Writing $\ln|1 - z_-| = \ln|r_-| + \ln(|1 - z_-|/|r_-|)$ and noting that $\ln|r_-| = \ln|\mathbf{r} - \mathbf{r}'|$ and $\ln(|1 - z_-|/|r_-|)$ is regular, we find that

$$G(\mathbf{r}; \mathbf{r}') = -\frac{1}{2\pi} \ln|\mathbf{r} - \mathbf{r}'| + \mathcal{G}(\mathbf{r}; \mathbf{r}'), \tag{3.44}$$

where the regular part $\mathcal{G}$ of G is given explicitly by

$$\mathcal{G}(\mathbf{r}; \mathbf{r}') = \frac{H(x; x')}{2\pi l} - \frac{1}{2\pi} \ln \frac{|1 - z_+||1 - z_-||1 - \zeta_+||1 - \zeta_-|}{|r_-|} + \mathcal{O}(q). \tag{3.45}$$

This expression for $\mathcal{G}$ is valid in the large aspect ratio limit $L/l \gg 1$ for which $q \ll 1$.

4. Numerics. In this section we compare the asymptotic solution U of (3.29) and (3.31) with the full solution obtained by numerically solving (3.3)–(3.6). In order to implement the boundary conditions (3.6) we use the steady-state solution of R_j given by (3.1), with the mean boundary concentration U_j identified with the corresponding solution of the singular perturbation problem as determined by (3.27) and (3.1). We then check that the resulting numerical solution is self-consistent; that is, the mean receptor concentration around the boundary is well approximated by the assumed value for U_j. The two-dimensional numerical solutions are generated by using the *Partial Differential Equation Toolbox* of MATLAB [30]. In Figure 4.1(A) we plot the steady-state concentration U given by (3.29) and (3.31) for a cable of length $L = 100\mu$m and radius $l = (2\pi)^{-1}\mu$m having 99 identical spines spaced 1μm apart along a single horizontal line ($y = 0\mu$m). In Figure 4.1(B) we plot the corresponding values of U_j, R_j, and S_j, with R_j, S_j given by (3.1) and (3.2). Here the diffusivity is $D = 0.1\mu\text{m}^2\text{s}^{-1}$ [13, 2], the somatic flux is $\sigma_0 = 0.1\mu\text{m}^{-1}\ \text{s}^{-1}$, and all spines are identical with radius $\varepsilon\rho = 0.1\mu$m, $A = 1\mu$m, $\omega = 10^{-3}\mu\text{m}^2\text{s}^{-1}$, $k = 10^{-3}\mu\text{m}^2\text{s}^{-1}$, $\sigma^{rec} = 10^{-3}\text{s}^{-1}$, $\sigma^{deg} = 10^{-4}\text{s}^{-1}$, and $\delta = 10^{-3}\text{s}^{-1}$. We set $\rho = 2\pi l = 1\mu$m so that $\varepsilon = 0.1$. While U decays significantly along the length of the cable, it varies very little around the circumference of the cable. In Figure 4.1(C) we show the results of numerically solving the original steady-state system for U described in (3.3)–(3.6). This numerical solution agrees almost perfectly with the perturbation solution shown in Figure 4.1(A).

We consider the parameter regime of Figure 4.1 physiological in the sense that parameter values were chosen from experimental data [2, 11, 13, 23] in conjunction with previous modeling studies [5, 6]. Our results suggest that in this parameter regime the dendrite can be treated as a quasi-one-dimensional system in which variations in receptor concentration around the circumference of the cable can be neglected. This is further reinforced by the observation that the solutions of the two-dimensional model shown in Figure 4.1(A)–(C) are virtually indistinguishable from the corresponding solution of the reduced one-dimensional model previously introduced in [6] (see Figure 4.1(D)). In the one-dimensional model, (2.1)–(2.4) are replaced by the inhomogeneous diffusion equation

$$\frac{\partial U}{\partial t} = D\frac{\partial^2 U}{\partial x^2} - \sum_{j=1}^{N} \frac{\omega_j}{2\pi l}[U_j - R_j]\delta(x - x_j), \tag{4.1}$$

with boundary conditions $U_j(t) = U(x_j, t)$ and

$$D\left.\frac{\partial U}{\partial x}\right|_{x=0} = -\frac{\sigma_0}{2\pi l D}, \quad D\left.\frac{\partial U}{\partial x}\right|_{x=L} = 0. \tag{4.2}$$

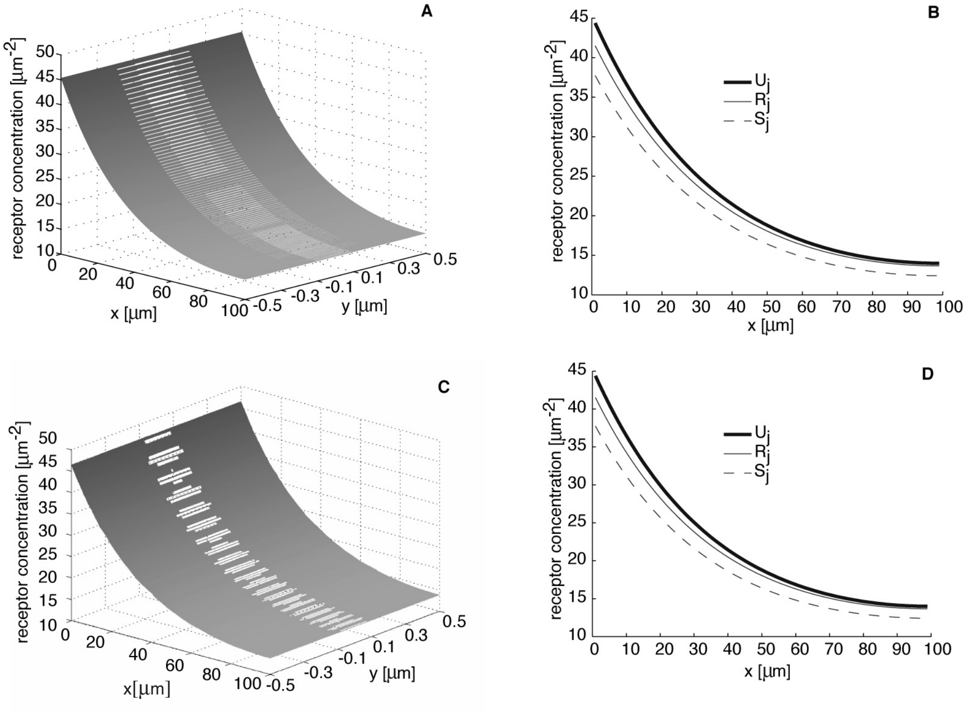


FIG. 4.1. *Plot of receptor concentration profiles along a dendritic cable with a uniform collinear distribution of dendritic spines. Cable and spine parameter values are given in the text.* (A) *Plot of bulk dendritic concentration U given by the outer solution* (3.29) *of the matched asymptotic expansion. Boundaries of spines are indicated by white lines. The dendritic cable is not drawn to scale.* (B) *Corresponding plots of U_j, R_j, and S_j obtained from* (3.27), (3.31), (3.1), *and* (3.2). *(Note that S_j is converted to a concentration by dividing through by the area A_j of a spine, which is taken to be* 1μm.*)* (C) *Numerical solution of* (3.3)–(3.6). (D) *Plots of U_j, R_j, and S_j obtained by solving the corresponding one-dimensional model* [6].

As in the two-dimensional model, R_j evolves according to (2.6) and (2.7). The steady-state solution of (4.1) can be obtained using the one-dimensional Green's function H such that

$$U(x) = \chi - \sum_{j=1}^{N} \frac{\widehat{\omega}_j [U_j - \widehat{R}_j]}{2\pi l D} H(x, x_j) + \frac{\sigma}{2\pi l D} H(x, 0)\,, \tag{4.3}$$

where the constant χ is determined from the self-consistency condition (3.9). The set of concentrations U_j can then be determined self-consistently by setting $x = x_i$ in (4.3) and solving the resulting matrix equation along analogous lines to (3.28) (cf. [6]). It is important to note that the excellent agreement between the two-dimensional and one-dimensional models is not a consequence of taking the spines to all lie along a one-dimensional line. This is illustrated in Figure 4.2 for a configuration of three staggered rows of 33 spines. The receptor concentrations are indistinguishable from the configuration consisting of a single row of 99 spines.

There is a simple heuristic argument to show why our original two-dimensional diffusion problem can be reduced to a corresponding one-dimensional problem, at

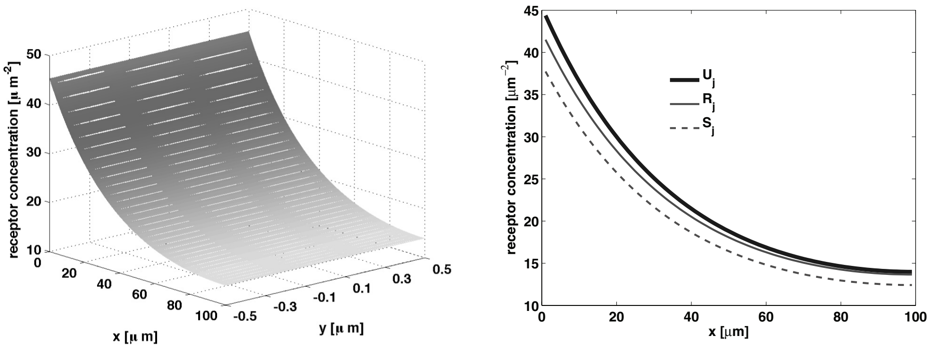


Fig. 4.2. *Plot of receptor concentration profiles along a dendritic cable with three staggered rows of spines. All other parameters are as in Figure* 4.1. *Left: Plot of bulk dendritic concentration U given by the outer solution* (3.29) *of the matched asymptotic expansion. Boundaries of spines are indicated by white lines. The dendritic cable is not drawn to scale. Right: Corresponding plots of U_j, R_j, and S_j obtained from* (3.27), (3.31), (3.1), *and* (3.2). *Concentration plots are indistinguishable from Figure* 4.1.

least in the parameter regime of Figure 4.1. Given a somatic flux $\sigma_0 = 0.1\mu\text{m}^{-1}\text{s}^{-1}$, a cable circumference of 1μm, and $N = 100$ spines, it follows that the mean number of receptors flowing through each spine boundary per second is 10^{-3}. Thus a rough estimate of the mean flux through each spine boundary of circumference $2\pi\varepsilon$ is $\bar{J} = 10^{-3}/(2\pi\varepsilon)\text{s}^{-1}\text{m}^{-1}$. Let ΔU represent the typical size of changes in receptor concentration needed to generate such a flux over a length scale Δx comparable to that of the dendritic circumference. Taking $\bar{J} \sim D\Delta U/\Delta x$ with $\Delta x = 2\pi l = 1\mu$m and $D = 0.1\mu\text{m}^2\text{s}^{-1}$ then gives $\Delta U \sim 10^{-2} l/\varepsilon \approx 1.6 \times 10^{-3}\mu\text{m}^{-2}$. Such a variation is negligible compared to the variation in receptor concentration along the length of the cable due to the source at the soma (see Figure 4.1), thus justifying a reduction to a one-dimensional model.

The above argument shows that the relatively small variation of receptor concentration around the circumference of the cable is a consequence of two factors: the large number of spines and the large aspect ratio ($l \ll L$) of the dendritic geometry; the length scales L, l put upper bounds on the maximum variation of the concentration in the x- and y-directions, respectively, for fixed σ_0. Thus we expect the two-dimensional nature of the spine's surface to become significant in the case of a few spines distributed on a short dendritic cable; such a situation could be relevant in the case of immature neurons. The solution for the receptor concentration will then be sensitive to the size of the spine radius ε. We illustrate this in Figure 4.3 in the case of a single spine centered at $(x, y) = (1, 0)$ on the surface of a dendrite of length $L = 2\mu$m. With this smaller value of L, the aspect ratio of the cable is now $L/l = 4\pi$. We also choose $\omega = k = 1\mu\text{m}^2\text{s}^{-1}$. In Figure 4.3(A)–(C) we show the singular perturbation solution for U when $\varepsilon = 0.01$, 0.1, and 0.4, respectively. In Figure 4.3(D)–(F) we show corresponding plots for the numerical solutions of U. It can be seen that the effect of the logarithmic singularity in the vicinity of the spine becomes prominent as ε decreases. Finally, in Figure 4.3(G) we plot the solution from the one-dimensional model. Although this model contains no information about the radius of the spine neck, and the aspect ratio of the system is now only moderately large, it still provides an approximate solution that agrees quite well with the full two-dimensional solutions.

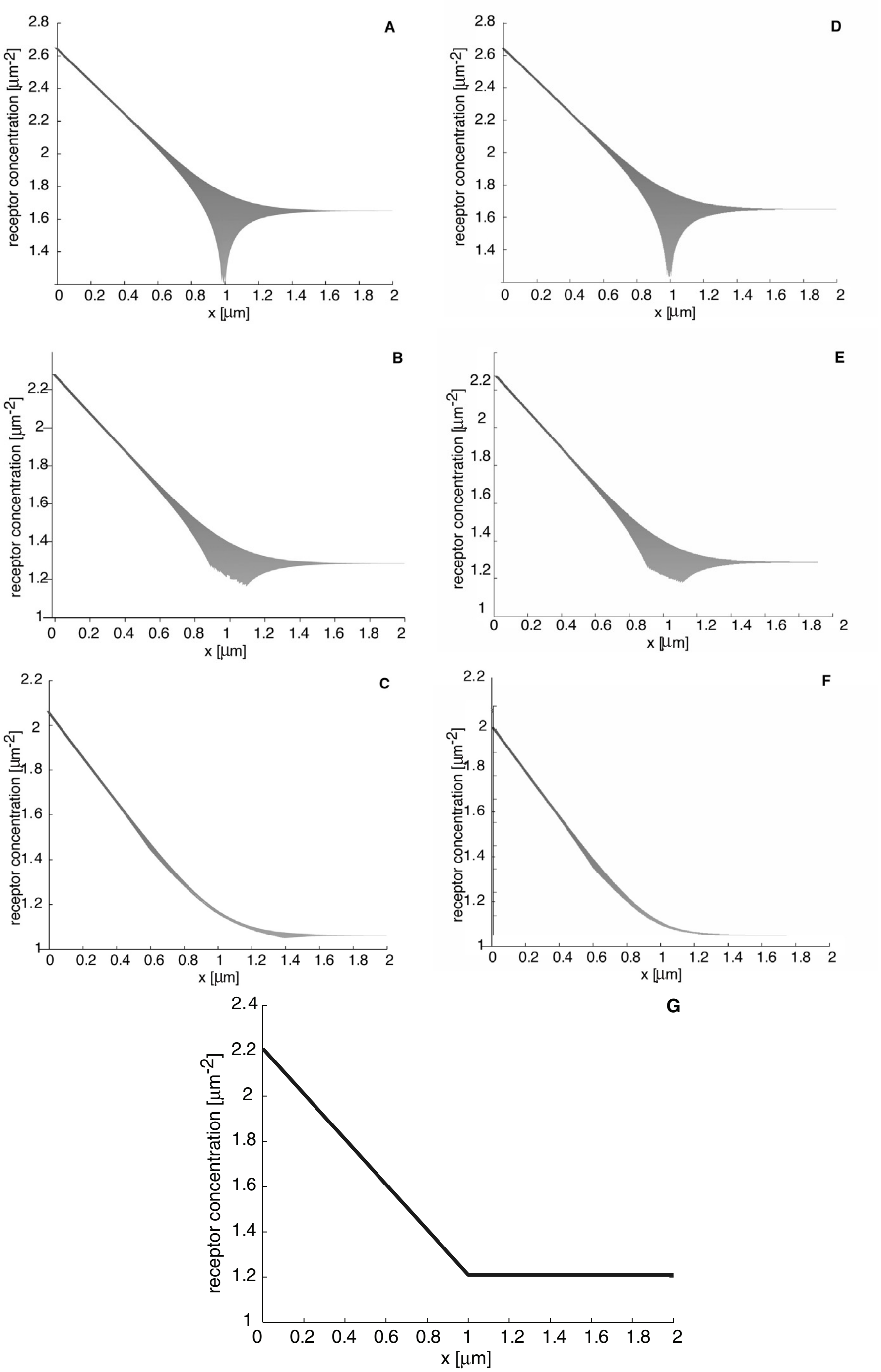


FIG. 4.3. *Effect of spine radius* ε *on the solution* U *for a single spine centered at* $x = 1\mu\mathrm{m}$, $y = 0$ *on a short dendritic cable of length* $L = 2\mu\mathrm{m}$. (A)–(C) *Plots of bulk dendritic receptor concentration* U *along the dendrite given by the outer solution* (3.29) *of the matched asymptotic expansion for parameter values as specified in the text and* $\varepsilon = 0.01$, 0.1, *and* $0.4\mu\mathrm{m}$, *respectively. The shaded region shows the range of values taken by the receptor concentration around the circumference of the cable as a function of distance* x *from the soma.* (D)–(F) *Corresponding plots of numerical solutions for* U. (G) *Plot of the one-dimensional solution* [6].

5. Mean first passage time (MFPT) for a single receptor. In this section we calculate the MFPT for a single tagged receptor to travel an axial distance X from the soma, $X < L$, assuming that the receptor does not undergo degradation. We then use this to determine an effective diffusivity, which takes into account the effects of trapping at spines. We proceed by reinterpreting the dendritic receptor concentration as a probability density and the diffusion equation (2.1) as a Fokker–Planck (FP) equation. The FP equation is defined on a spatial domain Ω_ε^X, where

$$\Omega_\varepsilon^X = \Omega_X \setminus \bigcup_{j=1}^{N_X} \Omega_j, \quad \Omega_j = \{\mathbf{r} : |\mathbf{r} - \mathbf{r}_j| \leq \varepsilon\}.$$

Here $\Omega_X = \{(x, y); 0 < x < X, |y| < \pi l\}$ and N_X is the number of spines within the rectangular domain Ω_X. We impose an absorbing boundary condition at $x = X$ so that the receptor is immediately removed once it reaches this boundary; i.e., we are interested only in the time it takes for a receptor to first reach $x = X$ from the soma. Let $u(\mathbf{r}, t|\mathbf{r}_0, 0)$ denote the probability density that at time $t \geq 0$ the receptor is located at $\mathbf{r} \in \Omega_\varepsilon^X$, given that it started at the point $\mathbf{r}_0 = (0, y_0)$. The probability density u evolves according to the FP equation

$$\frac{\partial u}{\partial t} = D\nabla^2 u, \quad (\mathbf{r}, t) \in \Omega_\varepsilon^X \times \mathbf{R}_+ \,, \tag{5.1}$$

with periodic boundary conditions at the ends $y = \pm\pi l$,

$$u(x, \pi l, t|\mathbf{r}_0, 0) = u(x, -\pi l, t|\mathbf{r}_0, 0), \quad \partial_y u(x, \pi l, t|\mathbf{r}_0, 0) = \partial_y u(x, -\pi l, t|\mathbf{r}_0, 0) \,, \tag{5.2}$$

and with

$$\partial_x u(0, y, t|\mathbf{r}_0, 0) = 0 \,, \quad u(X, y, t|\mathbf{r}_0, 0) = 0. \tag{5.3}$$

At each interior boundary $\partial\Omega_j$ we impose the mixed boundary condition

$$\varepsilon \partial_n u(\mathbf{r}, t|\mathbf{r}_0, 0) = -\frac{\omega_j}{2D\pi}(u(\mathbf{r}, t|\mathbf{r}_0, 0) - r_j(t|\mathbf{r}_0, t)), \quad \mathbf{r} \in \partial\Omega_j \,, \quad j = 1, \ldots, N_X. \tag{5.4}$$

Here $A_j r_j(t|\mathbf{r}_0, t)$ denotes the probability that the receptor is located within the jth spine at time t. Defining $s_j(t|\mathbf{r}_0, t)$ to be the corresponding probability that the receptor is located within the jth intracellular pool, we have

$$A_j \frac{dr_j}{dt} = \omega_j[u_j - r_j] - k_j r_j + \sigma_j^{rec} s_j \,, \tag{5.5}$$

$$\frac{ds_j}{dt} = -\sigma_j^{rec} s_j + k_j r_j. \tag{5.6}$$

Since we are assuming that the tagged receptor has not been degraded over the time interval of interest, we have set $\sigma_j^{deg} = 0$ for all j. We also assume no production of intracellular receptors so that $\delta_j = 0$. The initial conditions are $u(\mathbf{r}, 0|\mathbf{r}_0, 0) = \delta(\mathbf{r} - \mathbf{r}_0)$ and $r_j(0|\mathbf{r}_0, 0) = s_j(0|\mathbf{r}_0, 0) = 0$ for all j.

5.1. MFPT. Let $\tau(X|\mathbf{r}_0)$ denote the time it takes for a receptor starting at $\mathbf{r}_0 = (0, y_0)$ to first reach the boundary $x = X$. The function

$$F(X, t|\mathbf{r}_0) \equiv \int_{\Omega_\varepsilon^X} u(\mathbf{r}, t|\mathbf{r}_0, 0)\, d\mathbf{r} + \sum_{j=1}^{N_X} [A_j r_j(t|\mathbf{r}_0, 0) + s_j(t|\mathbf{r}_0, 0)] \tag{5.7}$$

is the probability that $t < \tau(X|\mathbf{r}_0)$; i.e., the probability that a receptor which was initially at the origin has not yet reached $x = X$ in a time t. Notice that $1 - F$ is the cumulative density function for τ, and hence

$$\frac{\partial(1-F)}{\partial t} = -\frac{\partial F}{\partial t} \tag{5.8}$$

is its probability density function. Thus the MFPT, denoted by T, is

$$T = -\int_0^\infty t\frac{\partial F}{\partial t}\, dt = \int_0^\infty F\, dt. \tag{5.9}$$

The last equality in (5.9) follows by integrating the first integral by parts and recalling that F, being an L^1 function in time, decays more rapidly to zero than t^{-1} as t becomes large. Therefore, integrating (5.7) over time gives us the following expression for $T(X|\mathbf{r}_0)$:

$$T(X|\mathbf{r}_0) = \lim_{z\to 0}\left(\int_{\Omega_\varepsilon^X} \widehat{u}(\mathbf{r}, z|\mathbf{r}_0, 0) d\mathbf{r} + \sum_{j=1}^{N_X} [A_j \widehat{r}_j(z|\mathbf{r}_0, 0) + \widehat{s}_j(z|\mathbf{r}_0, 0)]\right), \tag{5.10}$$

where $\widehat{\cdot}$ denotes the Laplace transform,

$$\widehat{f}(z) \equiv \int_0^\infty \mathrm{e}^{-zt} f(t) dt. \tag{5.11}$$

Upon Laplace transforming (5.1)–(5.6) and using the initial conditions, we can take the limit $z \to 0$ to obtain

$$\widehat{u}_j(\mathbf{r}_0) = \widehat{r}_j(0|\mathbf{r}_0, 0) = \frac{\sigma_j^{rec}}{k_j}\widehat{s}_j(0|\mathbf{r}_0, 0), \tag{5.12}$$

where

$$\widehat{u}_j(\mathbf{r}_0) = \frac{1}{2\pi\varepsilon}\int_{\partial\Omega_j} \widehat{u}(\mathbf{r}; \mathbf{r}_0)\, d\mathbf{r}. \tag{5.13}$$

Here we have set $\widehat{u}(\mathbf{r}; \mathbf{r}_0) = \lim_{z\to 0} \widehat{u}(\mathbf{r}, z|\mathbf{r}_0, 0)$. Hence, we obtain the boundary value problem

$$D\nabla^2 \widehat{u}(\mathbf{r}; \mathbf{r}_0) = -\delta(\mathbf{r} - \mathbf{r}_0), \quad \mathbf{r} \in \Omega_\varepsilon^X, \tag{5.14}$$

with

$$\widehat{u}(x, \pi l; \mathbf{r}_0) = \widehat{u}(x, -\pi l; \mathbf{r}_0), \quad \partial_y \widehat{u}(x, \pi l; \mathbf{r}_0) = \partial_y \widehat{u}(x, -\pi l; \mathbf{r}_0), \tag{5.15}$$

$$\partial_x \widehat{u}(0, y; \mathbf{r}_0) = 0, \quad \widehat{u}(X, y; \mathbf{r}_0) = 0, \tag{5.16}$$

and the mixed boundary condition

$$\varepsilon \partial_n \widehat{u}(\mathbf{r};\mathbf{r}_0) = -\beta_j \left(\widehat{u} - \frac{1}{2\pi\varepsilon} \int_{\partial\Omega_j} \widehat{u}\, d\mathbf{r} \right), \quad j = 1, \ldots, N_X \,, \tag{5.17}$$

where β_j is defined in terms of the hopping rate by $\beta_j \equiv \omega_j/(2D\pi)$.

As in section 3, we have a singularly perturbed boundary value problem, although in a weaker sense than the previous case. Carrying out a matched asymptotic expansion, the details of which are presented in the appendix, we find that there are no logarithmic singularities and the dependence on the spine size is $\mathcal{O}(\varepsilon^2)$. More specifically, the outer solution has the asymptotic expansion

$$\widehat{u} \sim \frac{G_X(\mathbf{r};\mathbf{r}_0)}{D} + 2\pi\varepsilon^2 \sum_{j=1}^{N_X} \left(\frac{\beta_j - 1}{\beta_j + 1} \right) \mathbf{a}_j \cdot \nabla_j G_X(\mathbf{r};\mathbf{r}_j)\,, \tag{5.18}$$

where ∇_j denotes differentiation with respect to the source variable $\mathbf{r}_j$, and $\mathbf{a}_j$ is defined by

$$\mathbf{a}_j \equiv \frac{\nabla G_X(\mathbf{r}_j;\mathbf{r}_0)}{D}. \tag{5.19}$$

Here G_X is the Green's function on the rectangular domain Ω_X with periodic boundary conditions at the ends $y = \pm\pi l$, a reflecting boundary at $x = 0$, and an absorbing boundary at $x = X$. Thus,

$$G_X(\mathbf{r};\mathbf{r}') = \frac{2}{|\Omega_X|} \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{\cos\left(\frac{\pi(2n+1)x}{2X}\right) \cos\left(\frac{\pi(2n+1)x'}{2X}\right) \mathrm{e}^{im(y-y')/l}}{\left(\frac{\pi(2n+1)}{2X}\right)^2 + \left(\frac{m}{l}\right)^2}. \tag{5.20}$$

Since G_X has a logarithmic singularity, it follows that

$$\nabla_j G_X(\mathbf{r};\mathbf{r}_j) \sim -\frac{1}{2\pi} \frac{(\mathbf{r} - \mathbf{r}_j)}{|\mathbf{r} - \mathbf{r}_j|^2} + \nabla_j \mathcal{G}_X \quad \text{as} \quad \mathbf{r} \to \mathbf{r}_j\,, \tag{5.21}$$

where $\mathcal{G}_X$ is the regular part of G_X. The average concentration $\widehat{u}_j$ around the boundary of the jth spine is obtained from the corresponding inner solution given in the appendix; see (A.6) and (A.8). Thus

$$\widehat{u}_j = \frac{G_X(\mathbf{r}_j;\mathbf{r}_0)}{D} + \frac{\varepsilon}{2\pi} \sum_{k=1}^{N_X} \left(\frac{2}{\beta_j + 1} \right) \int_0^{2\pi} \mathbf{a}_j \cdot \mathbf{e}_\rho \, d\theta\,, \tag{5.22}$$

where $\mathbf{e}_\rho$ is the unit normal to the circular boundary $\partial\Omega_j$. Since $\mathbf{a}_j$ is a constant vector, it follows that $\int \mathbf{a}_j \cdot \mathbf{e}_\rho d\theta = 0$, and so the $\mathcal{O}(\varepsilon)$ term vanishes. Finally, substituting (5.12), (5.18), and (5.22) into (5.10) shows that

$$T(X|\mathbf{r}_0) = \int_{\Omega_X} \frac{G_X(\mathbf{r};\mathbf{r}_0)}{D} d\mathbf{r} + \sum_{j=1}^{N_X} \eta_j \frac{G_X(\mathbf{r}_j;\mathbf{r}_0)}{D} + \varepsilon^2 \mathcal{J} + \cdots\,, \tag{5.23}$$

where $\eta_j = A_j + k_j/\sigma_j^{rec}$ and

$$\mathcal{J} = -\pi \sum_{j=1}^{N_X} \frac{G_X(\mathbf{r}_j;\mathbf{r}_0)}{D} + 2\pi \sum_{j=1}^{N_X} \left(\frac{\beta_j - 1}{\beta_j + 1} \right) \mathbf{a}_j \cdot \nabla_j \int_{\Omega_X} G_X(\mathbf{r};\mathbf{r}_j)\, d\mathbf{r}. \tag{5.24}$$

In the following we will determine the zeroth-order expression for the MFPT by dropping the $\mathcal{O}(\varepsilon^2)$ terms.

5.2. Evaluation of Green's function. We wish to evaluate the Green's function G_X in (5.20). We begin by expressing the double sum as

$$G_X(\mathbf{r};\mathbf{r}') = \frac{2}{|\Omega_X|}\sum_{n=0}^{\infty}\frac{\cos\left(\frac{\pi(2n+1)x}{2X}\right)\cos\left(\frac{\pi(2n+1)x'}{2X}\right)}{\left(\frac{\pi(2n+1)}{2X}\right)^2} + \frac{4}{|\Omega_X|}\sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\frac{\cos\left(\frac{\pi(2n+1)x}{2X}\right)\cos\left(\frac{\pi(2n+1)x'}{2X}\right)\cos\left(\frac{m(y-y')}{l}\right)}{\left(\frac{\pi(2n+1)}{2X}\right)^2+\left(\frac{m}{l}\right)^2}. \tag{5.25}$$

Upon using the identity (derived from p. 46 of [12])

$$\sum_{k=0}^{\infty}\frac{\cos((2k+1)\theta)}{(2k+1)^2+b^2} = \frac{\pi}{4b}\left[\frac{\cosh(b(\pi-|\theta|))}{\sinh(\pi b)} - \frac{\cosh(b|\theta|)}{\sinh(\pi b)}\right], \quad |\theta|\le\pi, \tag{5.26}$$

we can perform the sum over the index n in (5.25), yielding

$$\frac{1}{2\pi}\sum_{m=1}^{\infty}\frac{\cos\left(\frac{m(y-y')}{l}\right)\left[\cosh\left(\frac{m(2X-|x-x'|)}{l}\right)+\cosh\left(\frac{m(2X-|x+x'|)}{l}\right)\right]}{m\sinh\left(\frac{2Xm}{l}\right)} + \mathcal{E}\,, \tag{5.27}$$

where $\mathcal{E}$ is defined by

$$\mathcal{E} \equiv -\frac{1}{2\pi}\sum_{m=1}^{\infty}\frac{\cos\left(\frac{m(y-y')}{l}\right)\left[\cosh\left(\frac{m|x-x'|}{l}\right)+\cosh\left(\frac{m|x+x'|}{l}\right)\right]}{m\sinh\left(\frac{2Xm}{l}\right)}. \tag{5.28}$$

Following arguments similar to those used in section 3.3, together with the identity (3.37), the infinite sums in (5.27) and (5.28) can be represented as infinite sums of logarithmic terms. Our calculations are greatly simplified if X is not too small (e.g., by assuming that $X \gg l/2$). In this large aspect ratio limit, the identity (3.37) yields that

$$\frac{\cosh\left(\frac{m(2X-|x-x'|)}{l}\right)+\cosh\left(\frac{m(2X-|x+x'|)}{l}\right)}{\sinh\left(\frac{2Xm}{l}\right)} \approx \mathrm{e}^{-m|x-x'|/l}+\mathrm{e}^{-m|x+x'|/l}+\mathcal{O}(q_X)\,, \tag{5.29}$$

where $q_X \equiv \mathrm{e}^{-2X/l}$. In addition, $\mathcal{E}=\mathcal{O}(q_X)\ll 1$ and can be neglected to a first approximation. Using these approximations for the large aspect ratio limit, we readily derive that

$$G_X(\mathbf{r};\mathbf{r}') = \frac{H_X(x;x')}{2\pi l} - \frac{1}{2\pi}\ln|1-z_+||1-z_-| + \mathcal{O}(q_X)\,, \tag{5.30}$$

where (cf. p. 46 of [12])

$$H_X(x;x') = \frac{2}{X}\sum_{n=0}^{\infty}\frac{\cos\left(\frac{\pi(2n+1)x}{2X}\right)\cos\left(\frac{\pi(2n+1)x'}{2X}\right)}{\left(\frac{\pi(2n+1)}{2X}\right)^2} = \frac{X}{2}\left[h_X\left(\frac{x-x'}{X}\right)+h_X\left(\frac{x+x'}{X}\right)\right], \quad h_X(\theta)=1-|\theta|\,, \tag{5.31}$$

is the one-dimensional Green's function in the x-direction, and $z_{\pm}$ is as defined in (3.39).

Suppose that $\mathbf{r}' = \mathbf{r}_0 = (0, y_0)$. Since $x_0 = 0$,

$$G_X(\mathbf{r};\mathbf{r}_0) = \frac{X-x}{2\pi l} - \frac{1}{2\pi}\ln\left|1 - \mathrm{e}^{-x/l}\mathrm{e}^{i(y-y_0)/l}\right|^2 + \mathcal{O}(q_X). \tag{5.32}$$

If x is sufficiently large (e.g., $x \geq l$), then the contribution of the logarithmic term in (5.32) is of order $q_x = \mathrm{e}^{-2x/l}$, and hence

$$G_X(\mathbf{r};\mathbf{r}_0) = \frac{X-x}{2\pi l} + \mathcal{O}(q_x). \tag{5.33}$$

Since $\mathcal{O}(q_x)$ is exponentially small, this term can be dropped from (5.33), yielding the one-dimensional Green's function used in [6]. The fact that these results are effectively one-dimensional is again due to the large aspect ratio of our system.

5.3. Effective diffusivity and anomalous diffusion. Let us now evaluate the zeroth-order contributions to the MFPT $T(X|\mathbf{r}_0)$ given in (5.23). First, it follows from integrating (5.20) that $\int_{\Omega_X} G_X(\mathbf{r};\mathbf{r}_0)d\mathbf{r} = X^2/2$. Following the discussion of the previous paragraph, we will assume that all x_j are sufficiently large so that $G_X(\mathbf{r}_j;\mathbf{r}_0)$ is well approximated by the one-dimensional Green's function $(X - x_j)/(2\pi l)$. Since we are dropping any explicit dependence on y, y_0, we simply denote the MFPT $T(X|\mathbf{r}_0)$ by T. In the case of a large number of identical spines uniformly distributed along the length of the cable with spacing d (i.e., $N_X = X/d \gg 1$ and $x_j = jd$ for all j), we can compute an effective diffusivity D_{eff}. That is, substituting our one-dimensional approximation for G_X into (5.23) and dropping $\mathcal{O}(\varepsilon^2)$ terms gives

$$\begin{aligned} T &\approx \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\sum_{j=1}^{N_X}(X - jd) = \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\left(N_X X - \frac{(N_X+1)N_X d}{2}\right) \\ &\approx \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\left(N_X X - \frac{N_X^2 d}{2}\right) = \frac{X^2}{2D}\left(1 + \frac{\eta}{2\pi l d}\right) = \frac{X^2}{2D_{eff}}, \end{aligned} \tag{5.34}$$

where $\eta \equiv A + k/\sigma^{rec}$. In (5.34), the effective diffusivity D_{eff} is

$$D_{eff} = D\left(1 + \frac{\eta}{2\pi l d}\right)^{-1} = D\left(1 + \frac{A + k/\sigma^{rec}}{2\pi l d}\right)^{-1}. \tag{5.35}$$

As one would expect, the presence of traps reduces the effective diffusivity of a receptor. In particular, the diffusivity is reduced by increasing the ratio k/σ^{rec} of the rates of endocytosis and exocytosis, by increasing the surface area A of a spine, or by decreasing the spine spacing d. Interestingly, D_{eff} does not depend on the hopping rate ω, at least to lowest order in the spine size ε. At first sight this might seem counterintuitive, since a smaller ω implies that a receptor finds it more difficult to exit a spine. However, this is compensated by the fact that it is also more difficult for a receptor to enter a spine in the first place. (For a more detailed analysis of entry/exit times of receptors with respect to spines see [14, 15]).

In (5.34) the MFPT T is proportional to X^2. This relationship is the hallmark of Brownian diffusion, and here it is due to the fact that the spacing between spines is independent of the index j. Now suppose that the spacing varies with j according to $x_j = d(\ln(j)+1)$. In this case $N_X = \mathrm{e}^{X/d-1}$, and hence N_X grows exponentially with X [19]. Therefore, upon summing the series and using Stirling's formula, we get

(5.36)
$$\begin{aligned} T &\approx \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\sum_{j=1}^{N_X}(X - d(\ln(j)+1)) = \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\left(N_X X - d(\ln(N_X!) + N_X)\right) \\ &\approx \frac{X^2}{2D} + \frac{\eta}{2\pi l D}\left(N_X X - d N_X \ln(N_X)\right) = \frac{X^2}{2D} + \frac{\eta d}{2\pi l D}\mathrm{e}^{X/d-1} = \frac{X^2}{2D_{eff}(X)}, \end{aligned}$$

where the effective diffusivity is

$$D_{eff}(X) = D\left(1 + \frac{A + k/\sigma^{rec}}{2\pi l d}\frac{\mathrm{e}^{X/d-1}}{\frac{(X/d)^2}{2}}\right)^{-1}. \tag{5.37}$$

The fact that the effective diffusivity is a function of X indicates anomalous diffusion, which is to say that the relationship $T \propto X^2$ does not hold. Moreover, because $\mathrm{e}^{X/d-1}$ grows faster than $(X/d)^2$, the anomalous behavior is subdiffusive.

Note that the above analysis reproduces results obtained previously for a simplified one-dimensional model [6]. However, our asymptotic analysis shows that there are $\mathcal{O}(\varepsilon^2)$ corrections to the one-dimensional results given by (5.24). In particular, these higher-order corrections introduce a weak dependence of the MFPT on the size of the spines and the hopping rates ω_j via the parameters β_j.

6. Discussion.

In this paper we have used singular perturbation theory to determine the steady-state receptor concentration on the cylindrical surface of a dendritic cable in the presence of small dendritic spines, which act as partially absorbing traps. In the case of long, thin dendrites we have shown that the variation of the receptor concentration around the circumference of the cable is negligible so that the concentration profile along the cable can be determined using a simpler one-dimensional model [6]. We have also shown that the MFPT for a single tagged receptor to travel a certain distance along the cable is well approximated by considering a random walk along a one-dimensional cable. In both cases, our perturbation analysis provides details regarding corrections to the one-dimensional results that depend on the size ε of spines. Such corrections would be significant in the case of short dendrites with few spines, which may occur in immature neurons. An important extension of our work would be to consider a much more detailed model of receptor trafficking within each spine [10]. This would take into account the fact that the spine is not a homogeneous medium but contains a protein rich subregion known as the postsynaptic density where receptors can bind and unbind to various scaffolding proteins [4]. Interestingly, the coupling between the spine and the dendritic cable would not be affected by such details so that our solutions for the dendritic receptor concentration within the cable would carry over to more complex models.

The analysis presented in this paper provides a general mathematical framework for taking into account the effects of the size of spines on the surface diffusion of receptors (and other proteins) within the cell membrane. In the particular case of spiny dendrites it allows us to establish rigorously the validity of a one-dimensional reduction. This is important from a biological modeling perspective since the reduced model provides a relatively simple system in which to explore the role of diffusion in protein receptor trafficking along a dendrite. For example, one important biological issue is whether or not diffusion is sufficient as a mechanism for delivering protein receptors to distal parts of the dendrite [1]. If one ignores the effects of trapping in spines, then an estimate for the mean time a receptor takes to travel a distance X from the soma via surface diffusion along a uniform cable is $T = X^2/2D$. Even for a relatively large diffusivity $D = 0.45\mu\text{m}^2\text{s}^{-1}$, the mean time to reach a proximal synapse at 100μm from the soma is approximately 3 hrs., whereas the time to reach a distal synapse at 1mm from the soma is around 300 hrs. The latter is much longer than the average lifetime of a receptor, which is around 1 day. The one-dimensional formulae for the MFPT in the presence of traps (see section 5) establishes that trapping within spines increases the delivery time of receptors to synapses even further due to an effective reduction in the diffusivity. Indeed, if the density of spines grows sufficiently fast towards distal ends of the dendrite, then this increase in the MFPT could be significant due to the emergence of anomalous subdiffusive behavior. Interestingly, there is experimental evidence for an enhanced spine density at distal locations [18].

Finally, it would be interesting to consider protein receptor trafficking across a population of synapses with other geometric configurations. In this study we focused on synapses located within dendritic spines that are distributed along a dendritic cable, since most excitatory neurons in the central nervous system have such structures. However, there are some classes of neurons that have synapses located directly on the cell body or soma. One striking example is the chick ciliary ganglion, which supplies motor input to the iris of the eye; the ganglion has nicotinic receptors that are distributed across the surface of the cell body within somatic spines [3]. Thus the basic mathematical approach presented here could be extended to other biologically relevant examples of surface diffusion in the presence of partially absorbing traps, including diffusion on the surface of a spherical cell body, where a reduction to a one-dimensional problem would not be possible.

Appendix. In this appendix we present the singular perturbation analysis used to obtain the outer solution (5.18). First, let $\widehat{u}_0$ be the solution to the boundary value problem without traps given by (5.14), (5.15), and (5.16). Then

$$\widehat{u}_0(\mathbf{r}) = \frac{G_X(\mathbf{r};\mathbf{r}_0)}{D}\,, \tag{A.1}$$

where G_X is the Green's function of (5.20). For the problem with traps, we write the outer expansion as

$$\widehat{u} = \frac{G_X(\mathbf{r};\mathbf{r}_0)}{D} + \sigma(\varepsilon)\widehat{u}_1 + \cdots\,, \tag{A.2}$$

where $\sigma(\varepsilon)$ is to be found. In order to determine the inner solution near the jth hole, we introduce the scaled coordinates $\mathbf{s} = \varepsilon^{-1}(\mathbf{r} - \mathbf{r}_j)$ and set $V(\mathbf{s}) = \widehat{u}(\mathbf{r}_j + \varepsilon\mathbf{s})$. Then V satisfies (omitting the far-field condition)

$$\nabla_{\mathbf{s}}^2 V = 0\,, \quad s \equiv |\mathbf{s}| \geq 1\,, \tag{A.3}$$

$$\partial_s V = \beta_j \left(V - \frac{1}{2\pi}\int_0^{2\pi} V\, d\theta \right) \quad \text{on} \quad s \equiv |\mathbf{s}| = 1. \tag{A.4}$$

Notice that any constant V_0 satisfies this problem. We therefore write $V = V_0 + \mu(\varepsilon)V_1 + \cdots$. The inner and outer solutions must satisfy the matching condition

$$\frac{1}{D}\left[G_X(\mathbf{r}_j;\mathbf{r}_0) + \nabla G_X(\mathbf{r}_j;\mathbf{r}_0)\cdot(\mathbf{r}_j - \mathbf{r}) + \cdots\right] + \sigma(\varepsilon)\widehat{u}_1 \sim V_0 + \mu(\varepsilon)V_1 + \cdots\,. \tag{A.5}$$

This implies that $\mu(\varepsilon) = \varepsilon$ and that the constant V_0 is given by

$$V_0 = \frac{G_X(\mathbf{r}_j; \mathbf{r}_0)}{D}. \tag{A.6}$$

In addition, V_1 is the solution to the inner problem (A.3) and (A.4) with the far-field behavior

$$V_1 \sim \mathbf{a}_j \cdot \mathbf{s}\,, \qquad \mathbf{a}_j \equiv \frac{\nabla G_X(\mathbf{r}_j; \mathbf{r}_0)}{D}. \tag{A.7}$$

A simple separation of variables calculation gives the exact solution

$$V_1 = \mathbf{a}_j \cdot \mathbf{s} - \left(\frac{\beta_j - 1}{\beta_j + 1}\right) \mathbf{a}_j \cdot \frac{\mathbf{s}}{|\mathbf{s}|^2}. \tag{A.8}$$

Substituting this into the matching condition (A.5) gives $\sigma(\varepsilon) = \varepsilon^2$ and that $\widehat{u}_1$ satisfies the asymptotic singularity conditions

$$\widehat{u}_1 \sim -\left(\frac{\beta_j - 1}{\beta_j + 1}\right) \mathbf{a}_j \cdot \frac{(\mathbf{r} - \mathbf{r}_j)}{|\mathbf{r} - \mathbf{r}_j|^2} \quad \text{as} \quad \mathbf{r} \to \mathbf{r}_j. \tag{A.9}$$

The function $\widehat{u}_1$ is to satisfy Laplace's equation, the boundary conditions (5.15) and (5.16), and the singularity conditions (A.9) for $j = 1, \dots, N$.

Since the two-dimensional Green's function G_X has the logarithmic singularity $\frac{1}{2\pi} \log |\mathbf{r} - \mathbf{r}_j|$ for $\mathbf{r} \to \mathbf{r}_j$, it follows that $\nabla G_X(\mathbf{r}; \mathbf{r}_j)$ has the dipole singularity

$$\frac{1}{2\pi} \frac{(\mathbf{r} - \mathbf{r}_j)}{|\mathbf{r} - \mathbf{r}_j|^2} \quad \text{as} \quad \mathbf{r} \to \mathbf{r}_j. \tag{A.10}$$

Unfortunately, $\nabla G_X(\mathbf{r}; \mathbf{r}_j)$ does not satisfy the boundary conditions (5.15) and (5.16), so it cannot be used to construct an outer solution with the correct near-field behavior. On the other hand, we can construct a solution using $\nabla_j G_X(\mathbf{r}; \mathbf{r}_j)$, where ∇_j is the gradient operator with respect to the singular point $\mathbf{r}_j$. That is, $-\nabla_j G_X(\mathbf{r}; \mathbf{r}_j)$ has the same singular behavior as $\nabla G_X(\mathbf{r}; \mathbf{r}_j)$ and also satisfies the boundary conditions (5.15) and (5.16). The latter follows from the observation that the boundary conditions for G_X do not involve $\mathbf{r}_j$ so that the boundary and Laplace operators commute with ∇_j. Finally, using linearity and superposition over the N_X holes, we readily obtain that the outer approximation is given explicitly by

$$\widehat{u} \sim \frac{G_X(\mathbf{r}; \mathbf{r}_0)}{D} + 2\pi\varepsilon^2 \sum_{j=1}^{N_X} \left(\frac{\beta_j - 1}{\beta_j + 1}\right) \mathbf{a}_j \cdot \nabla_j G_X(\mathbf{r}; \mathbf{r}_j). \tag{A.11}$$

REFERENCES

[1] H. Adesnik, R. A. Nicoll, and P. M. England, *Photoinactivation of native AMPA receptors reveals their real-time trafficking*, Neuron, 48 (2005), pp. 977–985.

[2] M. C. Ashby, S. R. Maier, A. Nishimune, and J. M. Henley, *Lateral diffusion drives constitutive exchange of AMPA receptors at dendritic spines and is regulated by spine morphology*, J. Neurosci., 26 (2006), pp. 7046–7055.

[3] D. K. Berg and W. G. Conroy, *Nicotinic $\alpha 7$ receptors: Synaptic options and downstream signaling in neurons*, J. Neurobiol., 53 (2002), pp. 512–523.

[4] D. S. Bredt and R. A, Nicoll, *AMPA receptor trafficking at excitatory synapses*, Neuron, 40 (2003), pp. 361–379.

[5] P. C. Bressloff, *Stochastic model of protein receptor trafficking prior to synaptogenesis*, Phys. Rev. E (3), 74 (2006), 031910.

[6] P. C. Bressloff and B. A. Earnshaw, *Diffusion–trapping model of receptor trafficking in dendrites*, Phys. Rev. E (3), 75 (2007), 041915.

[7] L. Chen, T. Tracy, and C. I. Nam, *Dynamics of postsynaptic glutamate receptor targeting*, Curr. Opin. Neurobiol., 17 (2007), pp. 53–58.

[8] D. Choquet and A. Triller, *The role of receptor diffusion in the organization of the postsynaptic membrane*, Nat. Rev. Neurosci., 4 (2003), pp. 251–265.

[9] G. L. Collingridge, J. T. R. Isaac, and Y. T. Wang, *Receptor trafficking and synaptic plasticity*, Nat. Rev. Neurosci., 5 (2004), pp. 952–962.

[10] B. A. Earnshaw and P. C. Bressloff, *A biophysical model of AMPA receptor trafficking and its regulation during LTP/LTD*, J. Neurosci., 26 (2006), pp. 12362–12373.

[11] M. D. Ehlers, *Reinsertion or degradation of AMPA receptors determined by activity-dependent endocytic sorting*, Neuron, 28 (2000), pp. 511–525.

[12] I. M. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[13] L. Groc, M. Heine, L. Cognet, K. Brickley, F. A. Stephenson, B. Lounis, and D. Choquet, *Differential activity-dependent regulation of the lateral mobilities of AMPA and NMDA receptors*, Nat. Neurosci., 7 (2004), pp. 695–696.

[14] D. Holcman and Z. Schuss, *Escape through a small opening: Receptor trafficking in a synaptic membrane*, J. Statist. Phys., 117 (2004), pp. 975–1014.

[15] D. Holcman and A. Triller, *Modeling synaptic dynamics driven by receptor lateral diffusion*, Biophys. J., 91 (2006), pp. 2405–2415.

[16] M. J. Kennedy and M. D. Ehlers, *Organelles and trafficking machinery for postsynaptic plasticity*, Ann. Rev. Neurosci. 29 (2006), pp. 325–362.

[17] C.-H. Kim and J. E. Lisman, *A labile component of AMPA receptor-mediated synaptic transmission is dependent on microtubule motors, actin, and N-ethylmaleimide-sensitive factor*, J. Neurosci., 21 (2001), pp. 4188–4194.

[18] S. Konur, D. Rabinowitz, V. L. Fenstermaker, and R.Yuste, *Regulation of spine sizes and densities in pyramidal neurons*, J. Neurobiol., 56 (2003), pp. 95–112.

[19] M. Marin-Padilla, *Number and distribution of the apical dendritic spines of the layer V pyramidal cells in man*, J. Comp. Neurol., 131 (1967), pp. 475–489.

[20] M. Setou, D. H. Seog, Y. Tanaka, Y. Kanai, Y. Takei, and N. Hirokawa, *Glutamate-receptor-interacting protein GRIP*1 *directly steers kinesin to dendrites*, Nature, 417 (2002), pp. 83–87.

[21] M. Sheng and M. J. Kim, *Postsynaptic signaling and plasticity mechanisms*, Science, 298 (2002), pp. 776–780.

[22] I. Song and R. L. Huganir, *Regulation of AMPA receptors during synaptic plasticity*, Trends Neurosci., 25 (2002), pp. 578–588.

[23] K. E. Sorra and K. M. Harris, *Overview on the structure, composition, function, development, and plasticity of hippocampal dendritic spines*, Hippocampus, 10 (2000), pp. 501–511.

[24] R. Straube, M. J. Ward, and M. Falcke, *Reaction rate of small diffusing molecules on a cylindrical membrane*, J. Stat. Phys., 129 (2007), pp. 377–405.

[25] M. Titcombe and M. J. Ward, *Summing logarithmic expansions for elliptic equations in multiply-connected domains with small holes*, Canad. Appl. Math. Quart., 7 (1999), pp. 313–343.

[26] A. Triller and D. Choquet, *Surface trafficking of receptors between synaptic and extrasynaptic membranes*, Trends Neurosci., 28 (2005), pp. 133–139.

[27] M. J. Ward, W. D. Henshaw, and J. B. Keller, *Summing logarithmic expansions for singularly perturbed eigenvalue problems*, SIAM J. Appl. Math., 53 (1993), pp. 799–828.

[28] M. J. Ward, *Diffusion and bifurcation problems in singularly perturbed domains*, Natur. Resource Modeling, 13 (2000), pp. 271–302.

[29] P. Washbourne, X.-B. Liu, E. G. Jones, and A. K. McAllister, *Cycling of NMDA receptors during trafficking in neurons before synapse formation*, J. Neurosci., 24 (2004), pp. 8253–8264.

[30] MATLAB, *Partial Differential Equation Toolbox, User's Guide*, The MathWorks, Inc., Natick, MA, 1996.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1247–1267

CONVERGENCE OF DUAL ALGORITHM WITH ARBITRARY COMMUNICATION DELAYS[*]

RICHARD J. LA[†] AND PRIYA RANJAN[‡]

Abstract. We study the issue of convergence of user rates and resource prices under a family of rate control schemes called dual algorithms with arbitrary communication delays. We first consider a case where a single resource is shared by many users. Then we study a general network shared by heterogeneous users and derive sufficient conditions for convergence. We show that in the case of a single user utilizing a single resource, our condition is also necessary. Using our results we derive a sufficient condition for convergence with a family of popular utility and resource price functions. We present numerical examples to validate our analysis.

Key words. feedback control, attractivity

AMS subject classifications. 37C75, 37N35

DOI. 10.1137/050637716

1. Introduction.

A communication network, e.g., the Internet or a telephone network, comprises networking equipment that enables exchange of information between multiple parties or end user equipment (e.g., personal computers or laptops). Networking equipment includes switches, routers, and links (e.g., fiber optics or twisted copper wires) that connect switches and routers and have finite bandwidth or capacity, i.e., the number of bits that can be transmitted by a link per unit time. This networking equipment, which we call network elements in this paper, communicates using a set of rules called *network protocols*.

Since the links in the Internet have finite capacity, excessive demands brought on by the users can cause severe congestion or even a collapse of the Internet (e.g., the congestion collapse of 1986). Hence, in order to prevent any unexpected collapse of the Internet from severe congestion, it is imperative to control the congestion level inside the network by regulating the rates at which packets are injected into the network by the users, called the packet transmission rates or simply rates of the users. With the increasing size and complexity of the Internet, the problem of computing a fair share of network bandwidth for every user and allocating their rates is becoming a challenging task. To this end, in his seminal work [13] Kelly suggested that the problem of allocating fair shares of available network bandwidth to elastic traffic users,[1] which we call a *rate allocation* problem, can be posed as an optimization problem.

Under the proposed optimization framework each user receives a utility as a function of its rate, i.e., its share of bandwidth. The utility of a user can represent either the true utility or benefit the user receives or a utility function enforced by the end user protocol that adjusts the packet transmission rate on behalf of the user. An

[*] Received by the editors August 5, 2005; accepted for publication (in revised form) January 7, 2008; published electronically April 2, 2008. This work was supported in part by the National Science Foundation under grant ANI02-37997.

http://www.siam.org/journals/siap/68-5/63771.html

[†] Department of Electrical and Computer Engineering, A. V. Williams Bldg., University of Maryland, College Park, MD 20742 (hyongla@isr.umd.edu).

[‡] Institute for Systems Research, A. V. Williams Bldg., University of Maryland, College Park, MD 20742 (priya@isr.umd.edu).

[1] A user or a connection created by the user for information exchange is said to be *elastic* if its packet transmission rate can be adjusted based on feedback from the network (e.g., packet losses).

example of such a protocol is transmission control protocol (TCP), which is the most popular congestion or rate control protocol in the Internet today. In the latter case the selection of the utility function determines the behavior of the end user protocol and the desired rate allocation [14, 18, 20]. The objective of the optimization problem is to maximize the aggregate utility, i.e., the sum of the utilities of the users, subject to the link capacity constraints. Using the proposed framework Kelly and his colleagues proposed two classes of *distributed* rate control algorithms—primal and dual algorithms [14] (described in subsections 2.1 and 2.2)—and established their convergence to the desired rate allocation in the absence of delays.

In reality signals take time to travel from one end of a link connecting two network elements to the other end. This introduces a communication delay, which equals the length of the link divided by the speed at which the signal travels the medium, when a signal is transmitted over a link. Modeling communication delays over links between network elements is important when the delays are nonnegligible (e.g., intercontinental links) or widely varying with uncertainty; i.e., the delays are not known in advance. An example of such an environment is multihop wireless networks, which are wireless networks formed and maintained by (mobile) users without any infrastructure including physical links [27].

Recently, the convergence of user rates to the desired allocation under the primal algorithm in the presence of communication delays has been studied extensively [4, 12, 24, 30, 33]. The authors of [4, 12, 24] provided sufficient conditions on the gain parameters of the users[2] and communication delays for convergence, whereas Ranjan, La, and Abed [30] studied the case with arbitrary communication delays and provided sufficient conditions on user utility functions and resource price functions for convergence. Throughout this paper a resource refers to a link that connects two network elements. In addition, the authors of [30, 33] provided sufficient conditions for convergence with popular utility and resource price functions [2].

The convergence of user rates and resource prices under the dual algorithm in the presence of communication delays, however, has not been studied as much. Maulloo [23] studied the local convergence of the dual algorithm using a linearized model and provided sufficient conditions on the delay and resource gain parameters for local convergence. Low et al. introduced a family of dual algorithms, which are variants of those proposed by Kelly, Maulloo, and Tan [14], and studied the convergence in the presence of communication delays [1, 26].

In this paper we study the convergence property of the dual algorithm proposed by Kelly, Maulloo, and Tan [14] and derive sufficient conditions for convergence with *arbitrary* communication delays. We use the same technique we employed in [30] for the primal algorithm and demonstrate that the same framework can be used to investigate both the primal algorithm and the dual algorithm. We first consider a simpler scenario where a single resource is shared by a large number of users with heterogeneous round-trip delays. We model the dispersion or spread of heterogeneous round-trip delays of many users utilizing the resource using a family of probability distributions known as gamma kernels. This family of distributions has been used to model delay dispersion in other disciplines (see, e.g., [3, 32]). Then, using MacDonald's linear chain technique [21] we write the system dynamics as higher-order differential equations [7]. We derive a sufficient condition for convergence of user rates and resource prices. We also study the case where the users have the same round-trip delay (i.e., a discrete

[2] The gain parameter of a user or a resource determines how fast the user changes its rate or the resource updates its price in response to a change in network congestion level and is explained in subsections 2.1 and 2.2.

delay). We show that when the derived sufficient condition is violated, the system becomes unstable for sufficiently large delays and exhibits oscillatory behavior. Using a linear analysis we provide an upper bound on the delay for local convergence.

We extend the above results to general network cases where a set of resources is shared by heterogeneous users and derive sufficient conditions for convergence in the presence of arbitrary communication delays. These sufficient conditions are derived based on a simple discrete time map that emerges from the intrinsic market structure that underlies the rate control system and captures the interaction between the demands of the users and supplies of the network resources. We note that a similar approach has been used [10, 11, 22] in the past to study the behavior of delay differential systems. We apply our results to derive (necessary and) sufficient conditions with the same utility and resource price functions studied in [30, 33]. We show that the derived conditions for the dual algorithm are less restrictive than those for the primal algorithm in [30, 33]. In other words, for some choices of users' utility and resource price functions the dual algorithm converges irrespective of the communication delays, while the convergence of the primal algorithm with the same utility and resource price functions can be ensured only for small delays.

The main contributions of this paper can be briefly summarized as follows:

- We provide sufficient conditions for convergence of the dual algorithm, which are robust against the variations in the communication delays and resource gain parameters. This result can be used to provide a guideline and to simplify the design of the rate control system for a communication network that constantly evolves and changes (e.g., the Internet).
- We demonstrate that when the dispersion of round-trip delays can be modeled by a gamma kernel, the effects of the heterogeneous delays can be studied using the model of Hale and Ivanov [7] and are similar to those of introducing low pass filters in the feedback control loop [19].

This paper is organized as follows: Section 2 describes the optimization framework for rate allocation and the primal and dual algorithms proposed by Kelly [13] and Kelly, Maulloo, and Tan [14]. We study the simpler case of a single resource in section 3. This is followed by a study of general network cases in section 4. We apply our results to example utility and resource price functions and derive (necessary and) sufficient conditions for convergence with arbitrary communication delays in section 5. Simulation results are provided in section 6. We conclude in section 7.

2. Background. In this section we briefly describe the rate allocation problem in the proposed optimization framework. Consider a network with a set $\mathcal{L}$ of resources and a set $\mathcal{I}$ of users. Let C_l denote the finite capacity of resource $l \in \mathcal{L}$. Each user $i \in \mathcal{I}$ has a fixed route r_i, which is a set of resources traversed by user i's packets. We define a zero-one matrix A, where $A_{i,l} = 1$ if $l \in r_i$ and $A_{i,l} = 0$ otherwise. When its rate is x_i, user i receives utility $U_i(x_i)$. We take the view that the utility functions of the users are used to select the desired rate allocation among the users. The utility $U_i(x_i)$ is an increasing, strictly concave, and continuously differentiable function of x_i over the range $x_i \geq 0$. Under this setting, the rate allocation problem of interest can be formulated as the following optimization problem [13]:
SYSTEM(U,A,C):

$$\begin{aligned} &\text{maximize} \quad \sum_{i \in \mathcal{I}} U_i(x_i) \\ &\text{subject to} \quad A^T x \leq C, \qquad x \geq 0, \end{aligned} \tag{2.1}$$

where $C = (C_l,\, l \in \mathcal{L})$.[3] The first constraint is the capacity constraint, which states that the sum of the rates of all users utilizing resource l should not exceed its capacity C_l.

With the goal of solving this optimization problem in a *distributed* manner, Kelly, Maulloo, and Tan proposed two classes of rate-based algorithms [14]: Suppose that every user adopts rate-based congestion control in that it adjusts its rate based on the feedback from the network in the form of resource prices. Let $w_i(t)$ and $x_i(t)$ denote the amount user i is willing to pay, which we call its willingness to pay, per unit time and its rate at time t, respectively.[4] At time t each resource $l \in \mathcal{L}$ charges a price per unit flow of $\mu_l(t)$.

2.1. Primal algorithm. In a primal algorithm the end users adjust their rates based on the (shadow) prices per unit time of the resources given by

$$\mu_l(t) = p_l\Big(\sum_{i:l\in r_i} x_i(t)\Big), \qquad l \in \mathcal{L}, \tag{2.2}$$

where $p_l(\cdot)$ is an increasing function of the aggregate rate of the users going through it. Based on the resource prices, each user i adjusts its rate according to the following differential equation:

$$\frac{d}{dt}x_i(t) = \kappa_i\Big(w_i(t) - x_i(t)\sum_{l\in r_i}\mu_l(t)\Big), \qquad i \in \mathcal{I}, \tag{2.3}$$

where $w_i(t) = x_i(t) \cdot U_i'(x_i(t))$ and the user gain parameter $\kappa_i > 0$. The basic idea in (2.3) is to provide a market-based rate control mechanism; each user i constantly attempts to reach a market equilibrium where its willingness to pay $w_i(t)$ equals its total price per unit time charged by the resources given by $x_i(t)\sum_{l\in r_i}\mu_l(t)$. Note that the prices charged by the resources in (2.2) depend on the rates of the users, which can be viewed as users' current demands.

Under (2.3) one can see that both users' utility functions and resource price functions can be utilized to decide a desired allocation of network bandwidth to the end users. Therefore, the design of rate control algorithms is equivalent to selecting users' utility functions and the price functions of the resources that appear in (2.2) and (2.3).

Kelly, Maulloo, and Tan [14] have shown that, under some conditions on $p_l(\cdot)$, $l \in \mathcal{L}$, the user rates $x(t) = (x_i(t),\, i \in \mathcal{I})$ converge to a rate vector that maximizes the following expression:

$$\mathcal{U}(x) = \sum_i U_i(x_i) - \sum_l \int_0^{\sum_{i:l\in r_i} x_i} p_l(y)\,dy. \tag{2.4}$$

The first term in (2.4) is the aggregate utility of the users in our *SYSTEM(U,A,C)* problem in (2.1) which we want to maximize. Thus, the primal algorithm proposed by Kelly, Maulloo, and Tan in (2.3) solves a variation of the *SYSTEM(U,A,C)* problem in that it maximizes (2.4) instead of the aggregate utility in the original *SYSTEM(U,A,C)* problem.

[3] All vectors are assumed to be column vectors.

[4] Throughout the rest of the paper we refer to the willingness to pay per unit time as simply willingness to pay.

2.2. Dual algorithm. In a dual algorithm each resource $l \in \mathcal{L}$ updates its price, $\mu_l(t)$, based on the difference between the observed aggregate rate of the users and its *expected* rate $q_l(\mu_l(t))$ according to

$$\frac{d}{dt}\mu_l(t) = \kappa_l\Big(\sum_{j:l\in r_j} x_j(t) - q_l(\mu_l(t))\Big), \tag{2.5}$$

where the resource gain parameter $\kappa_l > 0$. The user rates are set to

$$x_j(t) = \frac{w_j(t)}{\sum_{l\in r_j}\mu_l(t)} =: D_j\Big(\sum_{l\in r_j}\mu_l(t)\Big), \tag{2.6}$$

where $D_j(\lambda)$ is the solution to $\lambda = U_j'(x)$ with $D_j(\lambda) = 0$ if $\lambda \geq U_j'(0)$ and $D_j(\lambda) = \infty$ if $\lambda \leq U_j'(\infty)$.[5] In other words, $D_j(\lambda)$ denotes the demand of user j as a function of the price per unit flow λ, which is the solution to the user optimization problem $\max_{x\geq 0} U_j(x) - x \cdot \lambda$ (called the USER$(U_j; \lambda)$ problem [13]). We call D_j the demand function of user j throughout the paper. It is easy to see that if $(U_j')^{-1}$ exists, then $D_j(\lambda) = (U_j')^{-1}(\lambda)$.

Here the function q_l can be viewed as the inverse function of the resource price function p_l in the primal algorithm. Hence, $q_l(\mu_l(t))$ gives the expected rate of resource l, given its current price $\mu_l(t)$. From (2.5)–(2.6) one can see that each resource adjusts its price according to the difference between the user demand (given by aggregate rate $\sum_{j:l\in r_j} x_j(t)$) and its desired supply at the current price (given by $q_l(\mu_l(t))$) [14, 31].

Kelly, Maulloo, and Tan [14] have proved that under mild technical conditions on the functions q_l, $l \in \mathcal{L}$, the expression

$$\mathcal{V}(\mu) = \sum_{i\in\mathcal{I}} \int_0^{\sum_{l\in r_i}\mu_l} D_i(\xi)\, d\xi - \sum_{l\in\mathcal{L}} \int_0^{\mu_l} q_l(\eta)\, d\eta \tag{2.7}$$

provides a Lyapunov function for the system of differential equations (2.5)–(2.6). We call a resource price vector μ that maximizes (2.7) a solution to (2.7) in the rest of the paper. Similarly, we call a rate vector x that maximizes (2.4) a solution to (2.4).

It is a simple exercise to show that if $q_l(\cdot) = p_l^{-1}(\cdot)$, at the solution to (2.7) denoted by $\overline{\mu}^\star = (\mu_l^\star,\ l \in \mathcal{L})$, (i) the solutions to USER$(U_i; \sum_{l\in r_i}\mu_l^\star)$ problems are the solution to (2.4), $\overline{x}^\star = (x_i^\star,\ i \in \mathcal{L})$, and (ii) $p_l(\sum_{i:l\in r_i} x_i^\star) = \mu_l^\star$ for all $l \in \mathcal{L}$. In other words, the user rates and resource prices at the equilibrium are the same under both the dual algorithm and the primal algorithm.

3. Single resource case. In this section we first study a simpler case where a single resource is shared by users with the same utility function. This is similar to the model used in [9] for studying the interaction of TCP connections with a random early detection (RED) gateway.[6] However, unlike in [9] we do not assume that the round-trip delays of the users are the same. General network cases will be studied in the following section.

[5] One should interpret $U_j'(\infty)$ to be $\lim_{x\to\infty} U_j'(x)$.

[6] A RED gateway is a queue management scheme that attempts to regulate the rates of the users by either dropping or marking packets with some probability to signal to the users impending congestion. The drop or mark probability is a function of the average number of packets queued at the gateway over a period.

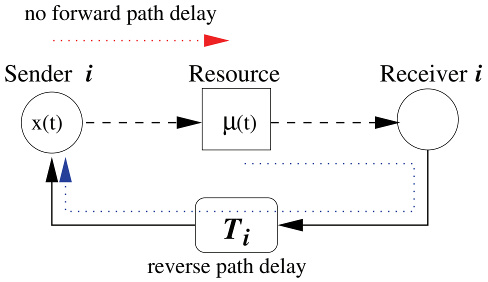


FIG. 3.1. *Network model for a single resource case.*

We assume that there is no forward delay from senders to the resource, and all of user i's round-trip delay, denoted by T_i, lies in the reverse path from the resource back to the sender. This is shown in Figure 3.1. Under this assumption the resource price update rule in (2.5) is given by

$$\frac{d}{dt}\mu(t) = \kappa\Big(\sum_{i\in\mathcal{I}} \frac{w_i(t)}{\mu(t-T_i)} - q(\mu(t))\Big). \tag{3.1}$$

We first focus on the case where users' willingness to pay is fixed, i.e., $w_i(t) = w$, $w > 0$, for all $i \in \mathcal{I}$. The case with user adaptation is discussed in subsection 3.3.

Here we are interested in the case where the resource is shared by a large number of users, e.g., an intercontinental link. In order to facilitate the analysis we assume that we can model the dispersion of heterogeneous round-trip delays of the users using some distribution function $\bar{K}$ as follows: Suppose $T \geq 0$ is the minimum round-trip delay of the users. For every $u \in [0,\infty)$, let $\bar{K}(u)$ be the *fraction* of users whose round-trip delays are less than or equal to $u + T$. We assume $\bar{K}$ is differentiable and yields a density function K, i.e., $K(u) = \bar{K}^{(1)}(u)$. This is reasonable when the number of users is large. Under this assumption, we can approximate the *average* rate of the users seen at the resource at time t using

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \frac{w}{\mu(t-T_i)} \approx \int_0^\infty \frac{w}{\mu(t-T-s)} K(s)\, ds, \tag{3.2}$$

where $|\mathcal{I}|$ is the cardinality of $\mathcal{I}$.

In the rest of this section we normalize both the aggregate rate of the users at the resource and the expected rate of the resource in (3.1) by the number of users $|\mathcal{I}|$, and we replace these terms with (3.2) and the expected rate per user of the resource, respectively:

$$\begin{aligned} \frac{d}{dt}\mu(t) &= \overline{\kappa}\left(\int_0^\infty \frac{w}{\mu(t-T-s)}\, K(s)\, ds - q_N(\mu(t))\right) \\ &= \overline{\kappa}\left(\int_0^\infty f(\mu(t-T-s))\, K(s)\, ds - q_N(\mu(t))\right), \end{aligned} \tag{3.3}$$

where $\overline{\kappa} = \kappa \cdot |\mathcal{I}|$, $q_N(\mu) = q(\mu)/|\mathcal{I}|$, and $f(\mu) = \frac{w}{\mu}$. Hence, the resource adjusts its price based on the average rate of the users and its expected (average) user rate based on the current price $\mu(t)$.

In this paper we consider the case where the delay density function K can be modeled by a family of generic delay kernels also known as gamma kernels, which have the following form:

$$K(u) = \begin{cases} \frac{\alpha^{r+1} u^r}{r!} \exp(-\alpha u) & \text{if } u \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.4}$$

where $\alpha > 0$ is a constant, and $r \in \{0, 1, 2, \ldots\}$. The kernel K with different parameters is plotted in Figure 3.2.

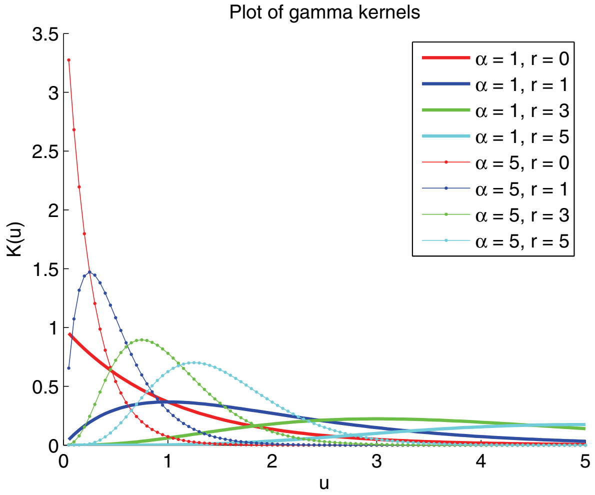


FIG. 3.2. *Gamma kernels for different parameters.*

The parameter r is called the order of gamma kernel K [21], and the mean of K for fixed (α, r) is given by

$$\mathbf{E}[K] = \int_0^\infty u \frac{\alpha^{r+1} u^r}{r!} \exp(-\alpha u)\, du = \frac{r+1}{\alpha}.$$

The kernel K with $r = 0$ and $r = 1$ is called the weak and strong kernel, respectively, and is frequently used to model distributed delay in different disciplines including population biology [3, 32]. These gamma kernels can be used to model a whole class of delay distributions (see, e.g., Figure 3.2), including an exponential distribution (i.e., the weak kernel). The case of discrete delay can be realized by letting r and α go to infinity simultaneously while keeping the mean delay $\frac{r+1}{\alpha}$ fixed.

In the remainder of this section we study the asymptotic behavior of the system in (3.3) under a set of reasonable assumptions on function q. In particular, we adopt the gamma kernels in (3.4) and apply *MacDonald's linear chain technique* to derive sufficient conditions for convergence of the dual algorithm [21].

Define $\mathbb{R}_+ := (0, \infty)$ and $\overline{\mathbb{R}}_+ := [0, \infty)$. The resource price function p is a strictly increasing function that maps $\overline{\mathbb{R}}_+$ to $\overline{\mathbb{R}}_+$. We first introduce the following assumption.

Assumption 1. (i) The function $q : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+$ is strictly increasing with $q(0) = 0$.

(ii) The function q is Lipschitz continuous on any bounded interval $[\mu_{min}, \mu_{max}] \subset \mathbb{R}_+$.

(iii) There exists a unique point $\mu^\star \in \mathbb{R}_+$ such that $f(\mu^\star) = q(\mu^\star)/|\mathcal{I}| = q_N(\mu^\star)$.

It is clear from (3.3) that $\mu^\star$ is the unique equilibrium of the dual algorithm in (2.5), at which the total demand of the users equals the expected rate of the resource given by $q(\mu^\star)$. Note that under Assumption 1, if the initial function is nonnegative, i.e., $\mu(s) \geq 0$ for all $s \in (-\infty, 0)$, the resource price $\mu(t)$ remains nonnegative because when $\mu(t) = 0$, the right-hand side of (3.3) is nonnegative if $\mu(t') \geq 0$ for all $t' < t$.

The existence of a unique solution of (3.3) is guaranteed by Theorem 2.3 in [8, p. 44] under Assumption 1 if there exist bounds μ_{min} and μ_{max}, where $0 < \mu_{min} < \mu_{max} < \infty$, such that $\mu(t)$ remains in $[\mu_{min}, \mu_{max}]$ for all $t \geq 0$, starting with an appropriate initial function that lies in $[\mu_{min}, \mu_{max}]$. In the following section we assume the existence of a unique positive solution of (3.3) and the bounds μ_{min} and μ_{max} such that $\mu_{min} \leq \mu(t) \leq \mu_{max}$ for all $t \geq 0$. We will provide an assumption under which this is true in subsection 3.2.

3.1. MacDonald's linear chain technique. Let $\mu(t)$ be a positive solution of (3.3) with some positive initial function $\mu(s)$ for all $s < 0$. We define

$$\begin{aligned} \omega_i(t) &= \int_0^\infty \frac{w}{\mu(t-T-s)}\, G^i(s)\, ds \\ &= \int_{-\infty}^t f(\mu(\theta - T))\, G^i(t-\theta)\, d\theta, \qquad i = 0, 1, \ldots, r, \end{aligned} \tag{3.5}$$

where $\theta = t - s$, $d\theta = -ds$, and $G^i(u) = \frac{\alpha^{i+1} u^i}{i!} \exp(-\alpha u)$, $u \geq 0$. Note that for any $i \in \{1, 2, \ldots\}$,

$$\frac{d}{du} G^i(u) = -\alpha G^i(u) + \alpha G^{i-1}(u) \quad \text{and} \quad \frac{d}{du} G^0(u) = -\alpha G^0(u). \tag{3.6}$$

Suppose that the delay kernel K is given by G^r for some $\alpha > 0$ and $r \in \{0, 1, \ldots\}$. From (3.6) we see that $(\mu(t), \omega_r(t), \omega_{r-1}(t), \ldots, \omega_0(t))$ satisfies

$$\frac{d}{dt}\mu(t) = \overline{\kappa}\left(\omega_r(t) - q_N(\mu(t))\right), \tag{3.7}$$

$$\frac{d}{dt}\omega_i(t) = -\alpha\left(\omega_i(t) - \omega_{i-1}(t)\right), \qquad i = 1, \ldots, r, \tag{3.8}$$

$$\frac{d}{dt}\omega_0(t) = -\alpha\left(\omega_0(t) - f(\mu(t-T))\right). \tag{3.9}$$

Define $\eta(t) := q_N(\mu(t))$. We have

$$\frac{d}{dt}\eta(t) = q_N'(q_N^{-1}(\eta(t)))\frac{d}{dt}\mu(t) \quad \text{and, hence,} \quad \frac{d}{dt}\mu(t) = \frac{1}{q_N'(q_N^{-1}(\eta(t)))}\frac{d}{dt}\eta(t),$$

where the inverse q_N^{-1} exists from Assumption 1. Note that, from Assumption 1, if $\mu(t)$ is positive for all $t \geq 0$, so is $\eta(t)$.

We can rewrite (3.7) in terms of $\eta(t)$ and $\omega_r(t)$ defined in (3.5):

$$\frac{1}{\overline{\kappa} \cdot q_N'(q_N^{-1}(\eta(t)))}\frac{d}{dt}\eta(t) = \omega_r(t) - \eta(t). \tag{3.10}$$

Using the definition of $\eta(t)$, we can rewrite (3.9) as

$$\begin{aligned} \frac{d}{dt}\omega_0(t) &= -\alpha\,\omega_0(t) + \alpha\, f(q_N^{-1}(\eta(t-T))) \\ &= -\alpha\,\omega_0(t) + \alpha\, \check{F}(\eta(t-T)), \end{aligned}$$

where the map $\check{F} : \overline{\mathbb{R}}_+ \to \mathbb{R}_+ \cup \{\infty\}$ is defined by

$$\check{F}(\eta) := f(q_N^{-1}(\eta)) = \frac{w}{q_N^{-1}(\eta)} \quad \forall\, \eta \in \overline{\mathbb{R}}_+. \tag{3.11}$$

Therefore, we have the following set of differential equations for describing the dynamics of the rate control system:

$$\begin{aligned} &\frac{1}{\kappa \cdot q_N'(q_N^{-1}(\eta(t)))} \frac{d}{dt}\eta(t) = -\eta(t) + \omega_r(t), \\ &\frac{d}{dt}\omega_i(t) = -\alpha\, \omega_i(t) + \alpha\, \omega_{i-1}(t), \qquad i = 1, \ldots, r, \\ &\frac{d}{dt}\omega_0(t) = -\alpha\, \omega_0(t) + \alpha\, \check{F}(\eta(t-T)). \end{aligned} \tag{3.12}$$

We note that the delay differential system in (3.12) is similar to that in [19] for a variant of the primal algorithm where the price charged by a resource is a function of the averaged or low pass filtered version of the aggregate rate of the users through the resource, as opposed to the users' instantaneous rates in the original primal algorithm. Hence, this suggests that the dynamical effect of heterogeneous delays in the dual algorithm, summarized by the averaging in (3.2), is similar to that of low pass filtering (or averaging) of the rate seen by the resource in the primal algorithm [19].

In the rest of this section we will show that, similarly as in the primal algorithm case [30], the convergence of the dual algorithm in (3.3) (or in (3.12)) can be studied using the map $\check{F}$ defined in (3.11).

3.2. Convergence. We denote by $C([-T, 0], A)$ the set of continuous functions from the interval $[-T, 0]$ to some interval $A \subset \mathbb{R}_+$ with topology of uniform convergence [8]. Suppose that there exists an interval $J := [a, b] \subset \mathbb{R}_+$, which is invariant under the map $\check{F}$, i.e., $\check{F}(J) \subset J$. Let $Y_J := C([-T, 0], J)$, and the initial function of $\eta(s)$, $s \in [-T, 0]$, is given by ϕ.

THEOREM 3.1. *If the initial function $\phi \in Y_J$ and $\omega_0(0), \ldots, \omega_r(0) \in J$, then for all $t \geq 0$ we have $(\eta(t; \phi), \omega_0(t), \ldots, \omega_r(t)) \in J^{r+2}$.*

Proof. A proof is provided in Appendix A. □

Theorem 3.1 implies that, under the assumption stated in the theorem, since $\eta(t; \phi)$ remains in J, from the definition of $\eta(t)$ and Assumption 1 the resource price $\mu(t)$ lies in a compact interval not including zero and, hence, stays away from zero for all $t \geq 0$. Therefore, the existence of a unique solution is guaranteed by Theorem 2.3 in [8, p. 44] under the assumption, as mentioned earlier.

For any interval A let $\mathrm{int}(A)$ denote its interior. We establish the convergence of (3.12) under the following assumption.

Assumption 2. There is a sequence of closed intervals $J_k \subset \mathbb{R}_+$, $k = 0, 1, \ldots$, such that (i) $\check{F}(J_k) \subset \mathrm{int}(J_{k+1}) \subset J_{k+1} \subset \mathrm{int}(J_k)$ for all $k = 0, 1, \ldots$, and (ii) $\cap_{k \geq 0} J_k = \{q_N(\mu^\star)\}$, where $\mu^\star$ is the unique point that satisfies $f(\mu^\star) = q_N(\mu^\star)$ in Assumption 1.

An example of utility and resource price functions that satisfy this assumption will be given in section 5.

THEOREM 3.2. *Suppose that Assumption 2 holds. If $\phi \in Y_{J_0}$ and $\omega_i(0) \in J_0$ for all $i = 0, 1, \ldots, r$, then $(\eta(t; \phi), \omega_0(t), \ldots, \omega_r(t)) \to (q_N(\mu^\star), \ldots, q_N(\mu^\star))$ as $t \to \infty$.*

Proof. A proof is given in Appendix B. □

3.3. User adaptation. In the previous subsection we have assumed that the willingness of the users to pay is fixed at w. This describes a case where the users' utility function is given by $w \cdot \log(x)$ [12, 14]. Suppose that the users' utility function is not of the form $w \cdot \log(x)$. If the user can accurately track the price per unit flow $\mu(t)$ and solve the USER$(U; \mu(t))$ problem, it should select a rate $x^\star(\mu(t))$ that satisfies $U'(x) = \mu(t)$ and set its willingness to pay to $w(t) = \mu(t) \cdot x^\star(\mu(t))$. This rate $x^\star(\mu(t))$ is given by the demand function D of the user defined in subsection 2.2.

We assume that the demand function D is (i) strictly decreasing in μ on the interval $[U'(\infty), U'(0))$, (ii) differentiable on $(U'(\infty), U'(0))$, and (iii) Lipschitz continuous on every bounded interval $[U'(\mu_{max}), U'(\mu_{min})]$, where $[\mu_{min}, \mu_{max}] \subset \mathbb{R}_+$. An example of utility functions that satisfy these assumptions is provided in section 5. Under these assumptions one can show that if there exists a unique $\mu^\star$ such that $D(\mu^\star) = q_N(\mu^\star)$ and the map $\check{F}(\eta) := D(q_N^{-1}(\eta))$ satisfies Assumption 2, then the theorems in subsection 3.2 hold with the same proofs [17]. Note that the function $f(\mu) = w/\mu$ plays the role of the demand function D when the willingness to pay w is constant.

3.4. Local stability with a homogeneous delay. In this subsection we consider the case where the resource is utilized by a single user with a fixed delay T. This user can be viewed as the aggregate of many users with the same round-trip delay T [9]. Recall that the case of discrete delay T can be modeled by gamma kernels by letting r and α go to ∞ simultaneously with a fixed mean delay $\frac{r+1}{\alpha} = T$. Using a linear analysis, we study the local stability of the system around the equilibrium $\mu^\star$.

When a single user utilizes the resource, the resource price is updated according to

$$\frac{d}{dt}\mu(t) = \kappa \left(x(t) - q(\mu(t))\right) = \kappa \left(D(\mu(t-T)) - q(\mu(t))\right). \tag{3.13}$$

We rewrite (3.13) in terms of $\eta(t) = q(\mu(t))$ as

$$\begin{aligned} \frac{d}{dt}\eta(t) &= \kappa \cdot q'(q^{-1}(\eta(t))) \left(D(q^{-1}(\eta(t-T))) - \eta(t)\right) \\ &= \zeta(\eta(t)) \left(\check{F}(\eta(t-T)) - \eta(t)\right), \end{aligned} \tag{3.14}$$

where $\zeta(\eta(t)) := \kappa \cdot q'(q^{-1}(\eta(t)))$ and $\check{F}(\eta) = D(q^{-1}(\eta))$. Note that $\zeta(\eta(t)) > 0$ from the assumptions on function q. Following the similar steps in the proof of Theorem 3.2 one can show that the resource price generated by (3.14) converges if Assumption 2 holds and $\phi \in Y_{J_0}$.

Assuming that the map $\check{F}$ is locally smooth around $\eta^\star = q(\mu^\star)$, one can find the conditions for local stability of the fixed point $\eta^\star$ for the delay differential equation in (3.14). Proposition 4 in [25, p. 17] tells us that the linearized system

$$\begin{aligned} \frac{d}{dt}Z(t) \; &= \; \zeta(\eta(t))\check{F}'(\eta(t-T))\Big|_{\eta=\eta^\star} Z(t-T) \\ &\quad + \left(\zeta'(\eta(t)) \left[\check{F}(\eta(t-T)) - \eta(t)\right] - \zeta(\eta(t))\right)\Big|_{\eta=\eta^\star} Z(t) \\ &= \; \zeta(\eta^\star)\check{F}'(\eta^\star)Z(t-T) - \zeta(\eta^\star)Z(t) \\ &:= -B \cdot Z(t-T) - A \cdot Z(t), \end{aligned} \tag{3.15}$$

where $A = \zeta(\eta^\star)$ and $B = -\zeta(\eta^\star)\check{F}'(\eta^\star)$,[7] is stable *if and only if*

[7] This is because $\eta^\star$ is a fixed point of the map $\check{F}$, i.e., $\check{F}(\eta^\star) = \eta^\star$.

(i) $A + B > 0$ and $A \geq |B|$, or

(ii) $B > |A|$ and $T \leq T^\star := \cos^{-1}(-A/B)/\sqrt{B^2 - A^2}$.

Note that in our problem the above conditions tell us that the linearized system in (3.15) is stable if and only if (i) the map $\check{F}$ is locally stable, i.e., $|\check{F}'(\eta^\star)| < 1$, or (ii) $\check{F}'(\eta^\star) < -1$ and

$$T \leq \frac{\cos^{-1}\left((\check{F}'(\eta^\star))^{-1}\right)}{\zeta(\eta^\star)\sqrt{\left(\check{F}'(\eta^\star)\right)^2 - 1}}.$$

Since local stability is required for global stability, these conditions tell us that the local stability of the map $\check{F}$ is both *necessary* and *sufficient* for convergence of the dual algorithm in (3.13) with an arbitrary delay in the neighborhood of the equilibrium point $\mu^\star$. We use these conditions to establish a *necessary* and *sufficient* condition for convergence with example utility and resource price functions in section 5.

4. General network cases.

In the previous section we studied the case where a single resource is shared by many users. Using MacDonald's linear chain technique we demonstrated that the effects of heterogeneous delays of the users are similar to introducing low pass filters in the feedback loop (see (3.12)). Then we showed that the asymptotic stability of the discrete time map $\check{F}$ is sufficient for convergence of the dual algorithm. In this section we extend these results to the case of a general network shared by multiple heterogeneous users with different delays. We first describe the model used for our analysis and then derive sufficient conditions for convergence of the resource prices and user rates with a general network topology.

4.1. General network model with delays.

In this subsection we describe the network model that captures the communication delays between network elements and end users under the assumption that the delays are *constant.* Recall from section 2 that $\mathcal{I} = \{1, \ldots, N\}$ is the set of users sharing a network consisting of a set $\mathcal{L} = \{1, \ldots, L\}$ of resources. Define $I_l = \{i \in \mathcal{I} \mid l \in r_i\}$ to be the set of users utilizing resource $l \in \mathcal{L}$. For all $i \in \mathcal{I}$ and $l \in r_i$ let $T^r_{i,l}$ and $T^f_{i,l}$ denote the delay of the feedback signal from resource l to sender i and the delay from sender i to resource l, respectively. This is shown in Figure 4.1. If user i packets do not traverse resource l, i.e., $l \notin r_i$, we assume that $T^r_{i,l} = T^f_{i,l} = 0$. Suppose that the resources in $r_i = \{l_{i,1}, \ldots, l_{i,R_i}\}$ are arranged in the order user i packets visit, where $R_i = |r_i|$. Define $T_i = T^f_{i,l_{i,k}} + T^r_{i,l_{i,k}}$, $k = 1, \ldots, R_i$, to be the round-trip delay of user i.

Similarly as in the single resource case, we introduce the following assumptions on the demand functions $D_i(\cdot)$ and $q_l(\cdot)$.

Assumption 3. (i) The demand functions $D_i(\mu)$ are strictly decreasing in price per unit flow μ on the interval $[U'_i(\infty), U'_i(0))$ and differentiable on $(U'_i(\infty), U'_i(0))$. In addition, they are Lipschitz continuous on any bounded interval $A \subset (U'_i(\infty), U'_i(0))$.

(ii) The function $q_l : [0, \infty) \to [0, \infty)$ is strictly increasing with $q_l(0) = 0$ for all $l \in \mathcal{L}$. Moreover, the function q_l is Lipschitz continuous on any bounded interval $[\mu_{l,min}, \mu_{l,max}] \subset \mathbb{R}_+$.

Assumption 3(ii) simply says that the equilibrium price of a resource increases with the aggregate rate traversing the resource, i.e., the total demand from the users. Note that Assumption 3(ii) also ensures that the inverse functions q_l^{-1} exist.

With the communication delays defined above, under Assumption 3, the differential equations in (2.5) and (2.6) become[8]

[8] As explained in the single resource case, under Assumption 3, the resource prices $\mu_l(t)$ remain nonnegative if the initial functions are nonnegative.

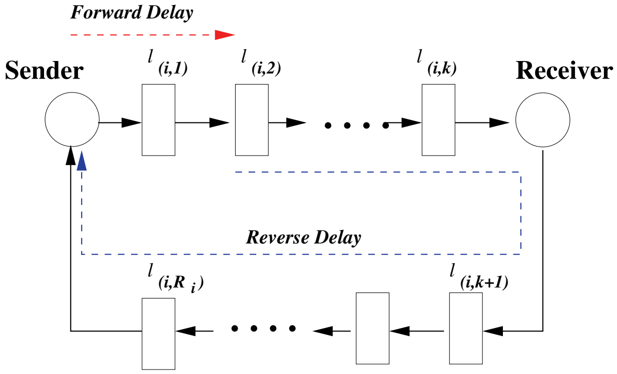


FIG. 4.1. *Network model with delays.*

$$\frac{d}{dt}\mu_l(t) = \kappa_l \left(\sum_{i\in I_l} D_i\Big(\sum_{j\in r_i} \mu_j(t-(T^r_{i,j}+T^f_{i,l}))\Big) - q_l(\mu_l(t))\right) \qquad \forall\, l\in\mathcal{L}. \tag{4.1}$$

We can show that a unique solution of (4.1) exists under an assumption to be stated shortly (i.e., Assumption 4) together with Assumption 3. We will revisit this issue after stating the assumption in subsection 4.2.

Similarly as in the previous section we define $\eta_l(t) := q_l(\mu_l(t))$. Recall that $\eta_l(t)$ denotes the expected rate of resource l at time t as a function of its price $\mu_l(t)$. We rewrite (4.1) in terms of $\eta_l(t)$ as

$$\frac{d}{dt}\eta_l(t) = \kappa_l q_l'(q_l^{-1}(\eta_l(t))) \left(\sum_{i\in I_l} D_i\Big(\sum_{j\in r_i} q_j^{-1}(\eta_j(t-T^r_{i,j}-T^f_{i,l}))\Big) - \eta_l(t)\right). \tag{4.2}$$

This can be put in the following matrix form:

$$\frac{d}{dt}\overline{\eta}(t) = \overline{\zeta}(t)\left[F(\tilde{\eta}(t)) - \overline{\eta}(t)\right], \tag{4.3}$$

where $\overline{\eta}(t) = (\eta_l(t);\ l\in\mathcal{L})$, $\overline{\zeta}(t) = \mathrm{diag}(\kappa_l \cdot q_l'(q_l^{-1}(\eta_l(t)));\ l\in\mathcal{L})$, $\tilde{\eta}(t) = (\eta_{(i,l)}(t);\ l\in\mathcal{L},\ i\in I_l)$, and $\eta_{(i,l)}(t) = (\eta_j(t-T^r_{i,j}-T^f_{i,l});\ j\in\mathcal{L})$. The lth element of the multidimensional map $F : \overline{\mathbb{R}}_+^{L\cdot\Xi} \to \overrightarrow{\mathbb{R}}_+^L$, where $\Xi := \sum_{l\in\mathcal{L}}|I_l|$ and $\overrightarrow{\mathbb{R}}_+ = \overline{\mathbb{R}}_+\cup\{\infty\}$, is defined by

$$F_l(\tilde{\eta}(t)) = \sum_{i\in I_l} D_i\Big(\sum_{j\in r_i} q_j^{-1}(\eta_j(t-T^r_{i,j}-T^f_{i,l}))\Big), \qquad l\in\mathcal{L}. \tag{4.4}$$

Note that Assumption 3 guarantees that the gain matrix $\overline{\zeta}(t)$ is positive definite.

4.2. Convergence. In this subsection we investigate the convergence of resource prices and aggregate rates at the resources generated by (4.3). More specifically, we will provide sufficient conditions for their convergence regardless of the delays $T^f_{i,j}$ and $T^r_{i,j}$.

DEFINITION 1. *A set $D\subset\mathbb{R}^L_+$ is said to be invariant under the map F if $F(\tilde{\eta})\in D$ for all $\tilde{\eta}\in D^\Xi$, i.e., $\tilde{\eta} = (\eta_{(i,l)};\ l\in\mathcal{L},\ i\in I_l)$ and $\eta_{(i,l)}\in D$ for all $l\in\mathcal{L}$ and $i\in I_l$. A vector $\overline{\eta}^\star\in\mathbb{R}^L_+$ is said to be a fixed point of F if $F(\overline{\eta}^\star,\ldots,\overline{\eta}^\star) = \overline{\eta}^\star$.*

The invariance of the map F can be interpreted as follows. Suppose that expected rates $\overline{\eta}(t)$ of the resources based on their current prices at time t as well as time delayed values $\eta_{(i,l)}(t)$, $l \in \mathcal{L}$ and $i \in I_l$, belong to the set D. Then the invariance of the set D implies that $F(\tilde{\eta}(t))$ lies in the set D. Similarly, a fixed point $\overline{\eta}^\star$ means that if $\overline{\eta}(t) = \overline{\eta}^\star$, $\eta_{(i,l)}(t) = \overline{\eta}^\star$ for all $l \in \mathcal{L}$ and $i \in I_l$, then $F(\tilde{\eta}(t)) = \overline{\eta}^\star$. In other words, $\overline{\eta}(t)$ and hence resource prices $\overline{\mu}(t)$ remain constant. One can verify that if $\overline{\eta}^\star$ is a fixed point of F, then $\overline{q}^{-1}(\overline{\eta}^\star) = (q_1^{-1}(\eta_1^\star), \ldots, q_L^{-1}(\eta_L^\star))$ is a solution to (2.7) from (4.3), i.e., $\overline{\eta}^\star = \overline{q}(\overline{\mu}^\star)$, where $\overline{\mu}^\star$ is a solution to (2.7).

We now state the assumption under which the convergence of (4.3) is established.

Assumption 4. Suppose that $\overline{\eta}^\star \in \mathbb{R}_+^L$ is a fixed point of the multidimensional map F. There is a sequence of compact, convex sets $E_k = \times_{l \in \mathcal{L}} E_{k,l} \subset \mathbb{R}_+^L$, $k \geq 0$, such that $F(E_k^\Xi) \subset \text{int}(E_{k+1}) \subset E_{k+1} \subset \text{int}(E_k)$ and $\cap_{k \geq 0} E_k = \{\overline{\eta}^\star\}$.

Define $T_{max} = \max_{l \in \mathcal{L},\, i \in I_l}(\max_{j \in r_i}(T_{i,j}^r + T_{i,l}^f))$. We denote by $C([-T_{max}, 0], E)$ the set of continuous functions mapping the interval $[-T_{max}, 0]$ into E with topology of uniform convergence [8]. Let $Y_{E_0} = C([-T_{max}, 0], E_0)$ be a subset of initial functions. When the initial function ϕ lies in Y_{E_0}, Theorem 2.3 in [8, p. 44] guarantees the existence of a unique solution of (4.3) under Assumptions 3 and 4. Denote by $\overline{\eta}(t;\phi)$ the solution of (4.3) constructed using an initial function $\phi \in Y_{E_0}$.

THEOREM 4.1. *All solutions $\overline{\eta}(t;\phi)$ starting with an initial function $\phi \in Y_{E_0}$ remain in E_0 for all $t \geq 0$ and converge to $\overline{\eta}^\star$ as $t \to \infty$ for all $T_{i,j}^r, T_{i,j}^f \in \mathbb{R}_+$.*

Proof. The basic idea of the proof of the theorem is similar to that of Theorem 4 in [30], and a proof is provided in [17], which is omitted in this paper due to a space constraint. □

Theorem 4.1 tells us that the attracting fixed point of the map F is stable in the set E_0. Since $\overline{\mu}(t) = \overline{q}^{-1}(\overline{\eta}(t))$, this tells us that $\overline{\mu}(t) \to \overline{\mu}^\star$ as $t \to \infty$.

4.3. Comparison with a homogeneous delay system.

In this subsection we investigate how the convergence of the resource prices under the general delay differential system in (4.3) is related to that of a much simpler system where (i) there are no forward delays from the senders to the resources, and (ii) all users have the same homogeneous round-trip delay. In other words, $T_{i,l}^f = 0$ and $T_{i,l}^r = T$ for all $i \in \mathcal{I}$ and $l \in r_i$, where T is some positive constant. Under this assumption the delay differential equations in (4.2) simplify to

$$\text{(4.5)} \quad \frac{d}{dt}\eta_l(t) = \kappa_l q_l'(q_l^{-1}(\eta_l(t))) \left(\sum_{i \in I_l} D_i\Big(\sum_{j \in r_i} q_j^{-1}(\eta_j(t-T)) \Big) - \eta_l(t) \right) \qquad \forall\ l \in \mathcal{L},$$

and the matrix form is given by

$$\frac{d}{dt}\overline{\eta}(t) = \overline{\zeta}(t)\left[\hat{F}(\eta(t-T)) - \overline{\eta}(t)\right],$$

where the map $\hat{F} : \overline{\mathbb{R}}_+^L \to \overrightarrow{\mathbb{R}}_+^L$ is defined by

$$\text{(4.6)} \qquad \hat{F}_l(\overline{\eta}) = \sum_{i \in I_l} D_i\Big(\sum_{j \in r_i} q_j^{-1}(\eta_j) \Big), \qquad l \in \mathcal{L}.$$

Assumption 5. The multidimensional map $\hat{F}$ has a fixed point $\overline{\eta}^\star \in \mathbb{R}_+^L$. In addition, there is a sequence of compact, convex sets $\check{E}_k = \times_{l \in \mathcal{L}} \check{E}_{k,l} \subset \mathbb{R}_+^L$, $k \geq 0$, such that $\hat{F}(\check{E}_k) \subset \mathrm{int}(\check{E}_{k+1}) \subset \check{E}_{k+1} \subset \mathrm{int}(\check{E}_k)$ and $\cap_{k \geq 0} \check{E}_k = \{\overline{\eta}^\star\}$; i.e., the map $\hat{F}$ is stable with a domain of attraction $\check{E}_0$.

We state the following theorem. The proof is provided in [17] due to a space constraint. The basic idea of the proof is essentially identical to that of Theorem 4.1 and is a modification of the proof of Theorem 4 in [30].

THEOREM 4.2. *All solutions* $\overline{\eta}(t;\phi)$ *of* (4.5) *starting with initial function* $\phi \in Y_{\check{E}_0}$ *converge to* $\overline{\eta}^\star$ *as* $t \to \infty$ *for all* $T > 0$.

Theorem 4.2 tells us that if $\check{E}_0$ is a region of attraction of the map $\hat{F}$ in (4.6), then the resource prices under the delay differential system in (4.5) with a homogeneous delay converge, provided that the initial function lies in $\check{E}_0$. It is easy to show that the same sequence of closed, convex sets $\check{E}_k$, $k \geq 0$, in Assumption 5 also satisfies Assumption 4. This follows from the assumed monotonicity properties of the functions q_l and D_i stated in Assumption 3. This in turn implies that the resource prices under the delay differential system in (4.3) converge if the initial function lies in $\check{E}_0$. Hence, the stability of the map $\hat{F}$ is a *sufficient* condition for convergence of resource prices under (4.3) with arbitrary communication delays.

5. Example utility and resource price functions. In this section we adopt a family of well-known users' utility functions and resource price functions studied in [30, 33] and derive a condition for convergence with arbitrary gains κ_l and delays, making use of our results in sections 3 and 4. Users' utility functions are of the following form:

$$U_a(x) = \begin{cases} \frac{1}{a}x^a, & -\infty < a < 1,\ a \neq 0, \\ \log(x), & a = 0. \end{cases} \tag{5.1}$$

In particular, $a = -1$ has been found useful for modeling the utility function of TCP-like algorithms [15]. With the utility functions of the form in (5.1), user i's price elasticity of demand[9] is given by $-1/(1-a)$. Thus, one can see that users become more elastic or responsive with increasing value of a [30]; i.e., the sensitivity of user demand, $1/(1-a)$, increases with a. Since the utility function $U_a(x)$ is strictly concave with $\lim_{x \downarrow 0} U'_a(x) = \infty$ and $\lim_{x \uparrow \infty} U'_a(x) = 0$, the demand function $D_a(\mu)$ is well defined for all $\mu \in \mathbb{R}_+$ and is given by

$$D_a(\mu) = \mu^{-1/(1-a)}. \tag{5.2}$$

The class of resource price functions that we are interested in has the form

$$p(x) = q^{-1}(x) = c \cdot \left(\frac{x}{C}\right)^b, \qquad x \in \overline{\mathbb{R}}_+, \tag{5.3}$$

where $b > 0$, c is some positive constant, and C is the capacity of the resource. However, C can be replaced by a virtual capacity, typically smaller than the real capacity. The use of virtual capacity was first proposed in [6] to reduce packet losses due to buffer overflow at highly utilized resources, at the expense of slightly reduced utilization. Kunniyur and Srikant in [16] proposed dynamically adjusting the virtual capacity and consequently the resource price function, based on the current aggregate rate seen at the resource. The value of c does not affect our convergence results and is assumed to be one unless stated otherwise. From (5.3), the function q is given by

$$q(\mu) = C \cdot \mu^{1/b}, \qquad \mu \in \overline{\mathbb{R}}_+. \tag{5.4}$$

[9] Price elasticity of demand measures the sensitivity of a user's demand to price changes and is defined to be $\frac{\mu}{D(\mu)} \frac{dD(\mu)}{d\mu}$, where $D(\mu)$ is the demand of the user at the price μ.

The parameter b is used to change the shape of the price function. The larger b is, the more convex and responsive the price function is. It is easy to verify that user demand function D_a in (5.2) and resource expected rate function q in (5.4) satisfy the assumptions in sections 3 and 4.

5.1. A single user, single resource case. Suppose that the utility function of the user is given by $U_a(x)$, $a < 1$, and $q(\mu)$ is of the form in (5.4) for some parameter $b > 0$.

Assumption 6. Suppose that $a + b < 1$.

Let σ be a constant that satisfies $\frac{b}{1-a} < \sigma < 1$. Since $a + b < 1$, it is easy to see that one can find such an σ. Fix $\overline{\alpha} > 1$ and choose $\overline{\beta} < 1$ that satisfies

$$\overline{\alpha}^{-\frac{(1-a)}{b}} < \overline{\beta} < \overline{\alpha}^{-\frac{b}{1-a}}. \tag{5.5}$$

Again, the existence of such a $\overline{\beta}$ is guaranteed from the assumption that $\overline{\alpha} > 1$ and $\frac{b}{1-a} < 1$.

Define a sequence of compact intervals I_k, $k \geq 0$, given by

$$I_k = \begin{cases} [\overline{\beta}^{\sigma^k} \mu^\star,\ \overline{\alpha}^{\sigma^k} \mu^\star] & \text{if } k \text{ is even,} \\ [\overline{\alpha}^{-\sigma^k} \mu^\star,\ \overline{\beta}^{-\sigma^k} \mu^\star] & \text{if } k \text{ is odd,} \end{cases} \tag{5.6}$$

where $\mu^\star$ is the unique solution to (2.7) and is given by $C^{-b(1-a)/(1+b-a)}$. Note that since $0 < \sigma < 1$, $\sigma^k \to 0$ and the interval I_k decreases to $\{\mu^\star\}$ as $k \to \infty$.

We define a map $\overline{F} : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+ \cup \{\infty\}$, where $\overline{F}(\mu) = q^{-1}(D_a(\mu))$.

LEMMA 5.1. *If Assumption* 6 *holds, then we have*

$$\overline{F}(I_k) \subset \text{int}(I_{k+1}) \subset I_{k+1} \subset \text{int}(I_k) \qquad \forall\ k \geq 0.$$

Proof. A proof is provided in Appendix C. □

THEOREM 5.2. *Suppose that $a + b < 1$ and the initial function $\phi \in C([-T, 0],$ $I_0 = [\overline{\beta}\mu^\star, \overline{\alpha}\mu^\star])$ with any $\overline{\alpha} > 1$ and $\overline{\beta} < 1$ satisfying* (5.5). *Then the solution $\mu(t; \phi)$ of* (3.13) *converges to $\mu^\star$ as $t \to \infty$.*

Proof. The theorem follows directly from Theorem 3.2 and Lemma 5.1. □

Note that as $\overline{\alpha} \uparrow \infty$, $\overline{\beta} \downarrow 0$. Since the above is true for any arbitrary $\overline{\alpha} > 1$ and $\overline{\beta}$ that satisfies (5.5), the resource price $\mu(t)$ converges to $\mu^\star$ starting from any arbitrary positive value because I_0 can be made large enough to contain the initial function.

We now show that if $a + b > 1$, then the system is unstable for sufficiently large T. Note that the map $\check{F}$ is given by

$$\check{F}(\eta) = D_a(q^{-1}(\eta)) = C^{b/(1-a)} \eta^{-b/(1-a)}.$$

Hence,

$$\check{F}'(\eta)\Big|_{\eta=\eta^\star} = \frac{-b \cdot C^{b/(1-a)}}{1-a} (\eta^\star)^{-(1+\frac{b}{1-a})}.$$

Here the fixed or equilibrium point $\eta^\star$ is given by $C^{b/(1+b-a)}$. Substituting this for $\eta^\star$ we obtain

$$\check{F}'(\eta)\Big|_{\eta=\eta^\star} = \frac{-b}{1-a}.$$

Therefore, if $a + b > 1$, then $\check{F}'(\eta^\star) < -1$ and the system is unstable for sufficiently large T from the linear stability analysis in subsection 3.4. A numerical example illustrating this is provided in subsection 6.2.

5.2. General network with multiple heterogeneous users case. Suppose that the utility functions of the users are of the form given by (5.1) and the resource price functions are of the type described by (5.3). The utility function of user i is parametrized by $a_i \in (-\infty, 1)$, and the price function of resource l has parameter $b_l > 0$. Making use of the fact stated in subsection 4.3 that the stability of the map $\hat{F}$ defined in (4.6) is sufficient for convergence with both a homogeneous delay and heterogeneous delays, we consider only the case described in subsection 4.3 with a homogeneous delay T in the reverse path only. In this case the map $\hat{F}$ is given by

$$\hat{F}_l(\overline{\eta}) = \sum_{i\in I_l} \left(\sum_{j \in r_i} \left(\frac{\eta_j}{C_j} \right)^{b_j} \right)^{-\frac{1}{1-a_i}}, \qquad l \in \mathcal{L}. \tag{5.7}$$

Note that $\hat{F}_l(\overline{\eta})$ is strictly decreasing in each of η_j, $j \in \cup_{i\in I_l} r_i$.

We define $b^i_{max} = \max_{l\in r_i} b_l$ for all $i \in \mathcal{I}$. Fix some finite positive constant $\overline{\alpha}$ larger than one. Suppose that $E_0 = \times_{l\in\mathcal{L}} E^l_0$, where $E^l_0 = [\overline{\beta}\eta^\star_l, \overline{\alpha}\eta^\star_l]$, with $\overline{\beta}$ being a positive constant that satisfies the following componentwise inequalities:

$$\hat{F}(\overline{\beta}\cdot\overline{\eta}^\star) < \overline{\alpha}\cdot\overline{\eta}^\star \quad \text{and} \quad \overline{\beta}\cdot\overline{\eta}^\star < \hat{F}(\overline{\alpha}\cdot\overline{\eta}^\star). \tag{5.8}$$

LEMMA 5.3. *Suppose that $a_i + b^i_{max} < 1$ for all $i \in \mathcal{I}$. Define $\sigma = -\max_{i\in\mathcal{I}} \left\{ \frac{b^i_{max}}{1-a_i} \right\} - \varepsilon$, where $0 < \varepsilon < 1 - \max_{i\in\mathcal{I}} \frac{b^i_{max}}{1-a_i}$. Then any $\overline{\beta}$ such that $\overline{\alpha}^{1/\sigma} < \overline{\beta} < \overline{\alpha}^\sigma$ satisfies* (5.8).

Proof. The proof is given in Appendix D of [17] due to a space constraint. □

We assume that $\overline{\beta}$ satisfies the condition in Lemma 5.3. Now, for $k = 1, 2, \ldots,$ we define

$$E_k = \begin{cases} \prod_{l\in\mathcal{L}} [\overline{\alpha}^{\sigma^k}\eta^\star_l \ , \ \overline{\beta}^{\sigma^k}\eta^\star_l], & k \text{ odd}, \\ \prod_{l\in\mathcal{L}} [\overline{\beta}^{\sigma^k}\eta^\star_i \ , \ \overline{\alpha}^{\sigma^k}\eta^\star_l], & k \text{ even}. \end{cases}$$

LEMMA 5.4. *Suppose that $a_i + b^i_{max} < 1$ for all $i \in \mathcal{I}$. Then $\hat{F}(E_{k-1}) \subset \text{int}(E_k) \subset E_k \subset \text{int}(E_{k-1})$, and $\cap_{k=0}^\infty E_k = \{\overline{\eta}^\star\}$.*

Proof. The proof is provided in Appendix E of [17] due to a space constraint. □

THEOREM 5.5. *Suppose that $a_i + b^i_{max} < 1$ for all $i \in \mathcal{I}$. If the initial function ϕ lies in $C([-T, 0], E_0)$, then $\overline{\eta}(t;\phi)$ produced by* (4.5) *converges to $\overline{\eta}^\star$ as $t \to \infty$ for all $T > 0$ and $\kappa_l > 0$, $l \in \mathcal{L}$.*

Proof. The theorem follows from Lemma 5.4 and Theorem 4.2. □

Now note that as $\overline{\alpha} \uparrow \infty$, $\hat{F}(\overline{\alpha}\cdot\overline{\eta}^\star) \to \underline{0} = [0, \ldots, 0]^T$. Hence, we can see that starting from any positive continuous initial function, the resource prices converge to $\overline{\mu}^\star$ as $t \to \infty$ from the above results because we can select a sufficiently large $E_0 \subset \mathbb{R}^L_+$ that contains the initial function.

5.3. Comparison with the primal algorithm. In this subsection we comment on the difference in the conditions for convergence under the primal algorithm [29, 30] and the dual algorithm studied in this paper. The convergence condition of the primal algorithm with a single user and a single resource is first studied in [29], and a necessary and sufficient condition for convergence with an arbitrary delay using the utility and resource price functions of (5.1) and (5.3) is provided. The derived convergence condition states that the user rate converges if and only if $a + b < -1$.[10]

[10] In [29] we considered only the utility functions with $a < 0$ because when $a \geq 0$ the system is unstable for sufficiently large delays.

Since $b > 0$, this implies that the parameter of the utility function needs to be strictly smaller than -1. Clearly, this is a more restrictive condition than the one provided for the dual algorithm in this paper (i.e., $(a + b < 1)$, which allows positive values of a for $b < 1$).

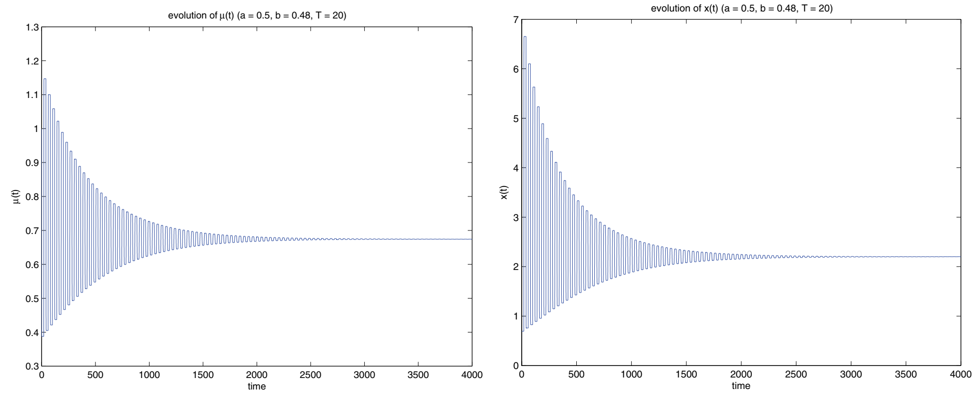
Similarly, in a general network case the sufficient conditions for convergence under the primal algorithm are given by $a_i + b^i_{max} < -1$ for all $i \in \mathcal{I}$ [30, 33], whereas the conditions that $a_i + b^i_{max} < 1$ for all $i \in \mathcal{I}$ suffice in the dual algorithm. Hence, the derived sufficient conditions for the primal algorithm in [30, 33] are more restrictive than those for the dual algorithm.

6. Numerical result.

In this section we present a numerical example to validate our results in the previous sections. We consider a single user, single resource case with the utility and resource price functions given in section 5 with $a = 0.5$. The capacity of the resource is set to $C = 5$. We vary the resource price function parameter b to create both a stable scenario and an unstable scenario according to our condition in Theorem 5.2.

6.1. Stable system.

In the first case we set $b = 0.49$. Since $a + b = 0.99 < 1$, Theorem 5.2 tells us that the resource price and user rate converge irrespective of the gain κ and the delay T. For the numerical example the delay is set to $T = 200$ and the gain is set to $\kappa = 1$. The initial value $\mu(t)$ is set to 1.2 for all $t \in [-T, 0]$. The evolution of $\mu(t)$ and $x(t)$ is plotted in Figure 6.1. As one can see both $\mu(t)$ and $x(t)$ converge to their equilibrium values of 0.6715 and 2.218, respectively.



(a) (b)

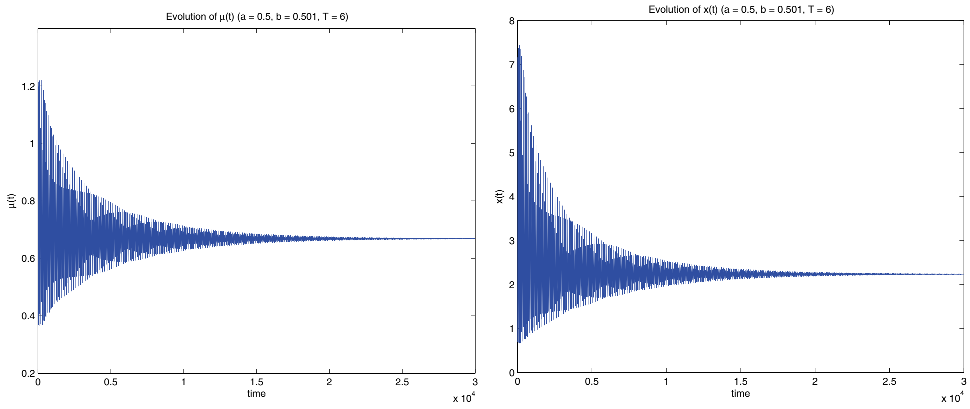
FIG. 6.1. *Evolution of $\mu(t)$ and $x(t)$ ($a = 0.5$, $b = 0.49$, $T = 200$).* (a) $\mu(t)$, (b) $x(t)$.

6.2. Unstable system.

In the second example we have increased the resource price function parameter b to 0.501. Since $a + b = 1.001 > 1$, the system loses its stability for sufficiently large delay T. The linear stability analysis provided in subsection 3.4 tells us that the linearized system in (3.14) is stable if and only if

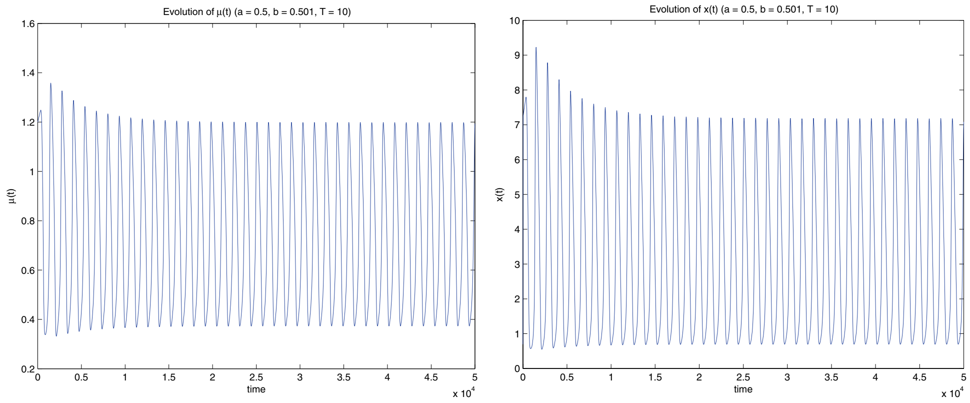
$$T \leq \frac{\cos^{-1}\left((\check{F}'(\eta^\star))^{-1}\right)}{\zeta(\eta^\star)\sqrt{(\check{F}'(\eta^\star))^2 - 1}} = 7.28.$$

Figures 6.2 and 6.3 plot the evolution of $\mu(t)$ and $x(t)$ for $T = 6$ and $T = 10$, respectively, sampled at every 20 unit times. As one can easily see, the system with



(a) (b)

FIG. 6.2. *Evolution of* $\mu(t)$ *and* $x(t)$ *(*$a = 0.5$*,* $b = 0.501$*,* $T = 6$*).* (a) $\mu(t)$, (b) $x(t)$.



(a) (b)

FIG. 6.3. *Evolution of* $\mu(t)$ *and* $x(t)$ *(*$a = 0.5$*,* $b = 0.501$*,* $T = 10$*).* (a) $\mu(t)$, (b) $x(t)$.

$T = 6$ is stable, and $\mu(t)$ and $x(t)$ converge to the equilibrium points of 0.6685 and 2.2379, respectively. However, when the delay T is increased to 10, the system loses stability and exhibits oscillatory behavior, as shown in Figure 6.3.

7. Conclusions.

We studied the issue of convergence of user rates and resource prices under a dual algorithm in the presence of communication delays. Using the same framework first employed in [30] for the primal algorithm, we derived sufficient conditions for convergence with arbitrary delays. In addition, we showed that these sufficient conditions can be obtained from a simple underlying discrete time system. We applied our result to an example of popular utility and resource price functions and derived sufficient conditions for convergence. In the simpler case of a single user utilizing a resource, we derived the necessary and sufficient condition for convergence. In addition, we studied the case when the convergence condition is violated and, using a linear stability analysis, provided an upper bound on the delay for convergence. Numerical examples are presented to validate our analysis. We believe that the framework used in this paper as well as in [30] is quite general and will prove to be useful for studying the convergence property of a variety of distributed control systems, in particular in the context of networking and networked control systems.

Appendix A. Proof of Theorem 3.1. The proof of the theorem is based on the following lemma.

Lemma A.1 (see [7, p. 507]). *Suppose that $I^\star = [a^\star, b^\star] \subset \mathbb{R}$, where $a^\star < b^\star$, is a compact interval and $\xi : \overline{\mathbb{R}}_+ \to I^\star$ is a continuous function. If $\sigma : \overline{\mathbb{R}}_+ \to \mathbb{R}_+$ is a bounded, continuous, strictly positive function and $u(t)$ is a solution of equation*

$$\sigma(t)\dot{u}(t) + u(t) = \xi(t), \tag{A.1}$$

with $u(0) \in I^\star$, then $u(t) \in I^\star$ for all $t \geq 0$.

Proof. The existence of a unique solution of (A.1) is guaranteed by the theorem in [28, p. 74]. We will prove this lemma by contradiction. Suppose that the lemma is not true. Define

$$t_0 = \inf\{t \geq 0 \mid u(t) \notin I^\star\}.$$

First, suppose that $u(t_0) = b^\star$. Then every interval $(t_0, t_0 + \delta)$, $\delta > 0$, contains a point τ such that $u(\tau) > b^\star$ and $\dot{u}(\tau) > 0$ under the assumption $a^\star < b^\star$ stated in the lemma. However, if $u(\tau) > b^\star$, (A.1) tells us that $\dot{u}(\tau) < 0$ because $\xi(t) \leq b^\star$, which is a contradiction. The case $u(t_0) = a^\star$ can be shown to lead to a similar contradiction. This completes the proof. □

We proceed with the proof of Theorem 3.1. Apply Lemma A.1 to $\frac{d}{dt}\omega_0(t) = -\alpha\omega_0(t) + \alpha\check{F}(\eta(t - T))$. Clearly, if $\omega_0(0) \in J$ and initial function $\phi \in Y_J$, then $\omega_0(t) \in J$ for all $0 \leq t \leq T$. By applying Lemma A.1 to $\frac{d}{dt}\omega_1(t) = -\alpha\omega_1(t) + \alpha\omega_0(t)$, we can argue that $\omega_1(t) \in J$ for all $0 \leq t \leq T$. Following this recursive argument, we can show that $\omega_i(t) \in J$ for $i = 0, \ldots, r$ and $\eta(t) \in J$ for all $0 \leq t \leq T$. Now by an induction argument on time (called the method of steps [5]) the same can be argued for all $t \geq 0$.

Appendix B. Proof of Theorem 3.2. The proof of the theorem is a simple application of the following lemma.

Lemma B.1. *Consider the same setup in Lemma* A.1. *Assume that $\bar{I} = [\bar{a}, \bar{b}]$ is a compact interval whose interior contains $I^\star$, i.e., $I^\star \subset \mathrm{int}(\bar{I})$. Then, for any $u(0) = u_0 \in \mathbb{R}_+$, there exists finite $t_0 := t_0(u_0, \bar{I})$ such that $u(t) \in \bar{I}$ for all $t \geq t_0$.*

Proof. First, note that if $u(t^*) \in \bar{I}$ for some $t^* \geq 0$, then from Lemma A.1 $u(t) \in \bar{I}$ for all $t \geq t^*$. Thus, we need only show that there exists some finite t_0 such that $u(t_0) \in \bar{I}$. Suppose that this is not true. First, assume that $u(t) > \bar{b}$ for all $t \geq 0$. Then from (A.1) we have $\dot{u}(t) = \frac{\xi(t) - u(t)}{\sigma(t)} \leq \frac{b^\star - \bar{b}}{\sigma(t)}$. Since $\sigma(t)$ is bounded, we can find a positive ε such that $\dot{u}(t) \leq -\varepsilon$ for all $t \geq 0$. However, this implies that $u(t) \downarrow -\infty$, which contradicts the assumption that $u(t) > \bar{b}$ for all $t \geq 0$. The other case, $u(t) < \bar{a}$ for all $t \geq 0$, can be shown to lead to a similar contradiction, and the lemma follows. □

We now proceed with the proof of the theorem. First, since $\check{F}(J_0) \subset \mathrm{int}(J_1)$, we can find a set of compact intervals $\{L_1^i,\, i = 0, \ldots, r\}$ such that

$$\check{F}(J_0) \subset \mathrm{int}(L_1^0) \subset L_1^0 \subset \mathrm{int}(L_1^1) \subset \cdots \subset \mathrm{int}(L_1^r) \subset L_1^r \subset \mathrm{int}(J_1). \tag{B.1}$$

Using the property in (B.1), we can repeatedly apply Lemma B.1, starting with the third equation in (3.12) for $\omega_0(t)$ and then the second equation with $\omega_i(t)$, $i = 1, \ldots, r$, to find finite t_1^i, $i = 0, \ldots, r$, where $0 \leq t_1^0 \leq t_1^1 \leq \cdots \leq t_1^r$, such that $\omega_i(t) \in L_1^i$ for all

$t \geq t_1^i$. Finally, applying Lemma B.1 to the first equation in (3.12) we can find finite $t_1^\star \geq t_1^r$ such that $\eta(t) \in J_1$ for all $t \geq t_1^\star$.

Now, by an induction argument for each $k = 2, 3, \dots$, one can find an increasing sequence $t_k^\star$, $k = 1, 2, \dots$, such that, for all $t \geq t_k^\star$, $\eta(t) \in J_k$ and $\omega_i(t) \in J_k$, $i = 0, 1, \dots, r$. Now the theorem follows from the assumption that $\mathrm{diam}(J_k) \to 0$ as $k \to \infty$ and $\cap_{k \geq 1} J_k = \{q_N(\mu^\star)\}$.

Appendix C. Proof of Lemma 5.1. In this proof we consider only the case of even k. Proof for the case with odd k follows in a similar manner. From the monotonicity of the map $\overline{F}(\mu) = C^{-b}\mu^{-b/(1-a)}$, it suffices to show that $\overline{F}(\overline{\beta}^{\sigma^k}\mu^\star) \in I_{k+1}$ and $\overline{F}(\overline{\alpha}^{\sigma^k}\mu^\star) \in I_{k+1}$.

First,

$$\begin{aligned}\overline{F}(\overline{\alpha}^{\sigma^k}\mu^\star) &= C^{-b}(\overline{\alpha}^{\sigma^k}\mu^\star)^{-b/(1-a)} = C^{-b}\mu^{\star -b/(1-a)}\overline{\alpha}^{\sigma^k(-b/(1-a))}\\ &= \mu^\star\overline{\alpha}^{\sigma^k(-b/(1-a))} > \mu^\star\overline{\alpha}^{-\sigma^{k+1}},\end{aligned}$$

where the last equality follows from the fact that $\mu^\star$ is a fixed point of the map $\overline{F}$, and the inequality follows from the assumption that $0 < \frac{b}{1-a} < \sigma < 1$ and $\overline{\alpha} > 1$. Clearly, $\mu^\star\overline{\alpha}^{\sigma^k(-b/(1-a))} < \mu^\star < \mu^\star\overline{\beta}^{-\sigma^{k+1}}$ because $\overline{\beta} < 1 < \overline{\alpha}$.

Similarly,

$$\begin{aligned}\overline{F}(\overline{\beta}^{\sigma^k}\mu^\star) &= C^{-b}(\overline{\beta}^{\sigma^k}\mu^\star)^{-b/(1-a)} = C^{-b}\mu^{\star -b/(1-a)}\overline{\beta}^{\sigma^k(-b/(1-a))}\\ &= \mu^\star\overline{\beta}^{\sigma^k(-b/(1-a))} < \mu^\star\overline{\beta}^{-\sigma^{k+1}}.\end{aligned}$$

Therefore, $\overline{F}(I_k) \subset \mathrm{int}(I_{k+1})$, and the lemma follows.

REFERENCES

[1] S. Athuraliya, V. H. Li, S. H. Low, and Q. Yin, *Rem: Active queue management*, IEEE Network, 15 (2001), pp. 48–53.

[2] O. Blanchard and S. Fisher, *Lectures on Macroeconomics*, MIT Press, Cambridge, MA, 1989.

[3] J. M. Cushing, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomath. 20, Springer-Verlag, Berlin, New York, 1977.

[4] S. Deb and R. Srikant, *Global stability of congestion controllers for the internet*, IEEE Trans. Automat. Control, 48 (2003), pp. 1055–1060.

[5] R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, Berlin, 1977.

[6] R. J. Gibbens and F. Kelly, *Resource Pricing and the Evolution of Congestion Control*, http://www.statslab.cam.ac.uk/~frank (1999).

[7] J. K. Hale and A. F. Ivanov, *On a high order differential delay equation*, J. Math. Anal. Appl., 173 (1993), pp. 505–514.

[8] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[9] C. Holot, V. Misra, D. Towsley, and W. Gong, *A control theoretic analysis of RED*, in Proceedings of IEEE INFOCOM, Anchorage, AK, 2001, pp. 1510–1519.

[10] A. F. Ivanov, M. A. Pinto, and S. I. Trofimchuk, *Global behavior in nonlinear systems with delayed feedback*, in Proceedings of the IEEE Conference on Decision and Control (CDC), Sydney, Australia, 2000.

[11] A. F. Ivanov and A. N. Sharkovsky, *Oscillations in singularly perturbed delay equations*, in Dynamics Reported, Dynam. Report. Expositions Dynam. Systems (N. S.) 1, C. K. R. T. Jones, U. Kirchgraber, and H. O. Walther, eds., Springer-Verlag, Berlin, 1992, pp. 164–224.

[12] R. Johari and D. Tan, *End-to-end congestion control for the internet: Delays and stability*, IEEE/ACM Trans. Networking, 9 (2001), pp. 818–832.

[13] F. Kelly, *Charging and rate control for elastic traffic*, Eur. Trans. Telecommun., 8 (1997), pp. 33–37.

[14] F. Kelly, A. Maulloo, and D. Tan, *Rate control for communication networks: Shadow prices, proportional fairness and stability*, J. Oper. Res. Soc., 49 (1998), pp. 237–252.

[15] S. Kunniyur and R. Srikant, *End-to-end congestion control schemes: Utility functions, random losses and ECN marks*, in Proceedings of IEEE INFOCOM, Tel-Aviv, Israel, 2000.

[16] S. Kunniyur and R. Srikant, *Analysis and design of an adaptive virtual queue algorithm for active queue management*, in Proceedings of ACM SIGCOMM, San Diego, CA, 2001.

[17] R. La and P. Ranjan, *Stability of the Dual Algorithm with Arbitrary Network Delays*, Tech. report, 2006, available online from http://www.ece.umd.edu/~hyongla/publication.htm.

[18] R. J. La and V. Anantharam, *Utility-based rate control in the Internet for elastic traffic*, IEEE/ACM Trans. Networking, 10 (2002), pp. 272–286.

[19] R. J. La and P. Ranjan, *Global stability with averaged feedback and network delay*, Automatica, 42 (2006), pp. 1817–1820.

[20] S. Low and D. Lapsley, *Optimization flow control*, IEEE/ACM Trans. Networking, 7 (1997), pp. 861–874.

[21] N. MacDonald, *Time Lags in Biological Models*, Lecture Notes in Biomath. 27, Springer-Verlag, Berlin, New York, 1997.

[22] J. Mallet-Paret and R. D. Nussbaum, *Global continuation and asymptotic behaviour for periodic solutions of a differential-delay equation*, Ann. Mat. Pura Appl. (4), 145 (1986), pp. 33–128.

[23] A. K. Maulloo, *Stability of Communication Networks*, Tech. report, Ph.D. dissertation, University of Mauritius, Reduit, Mauritius, 2000.

[24] F. Mazenc and S.-I. Niculescu, *Remarks on the stability of a class of TCP-like congestion control models*, in Proceedings of the IEEE Conference on Decision and Control (CDC), Maui, HI, 2003.

[25] S. Niculescu, E. I. Verriest, L. Dugard, and J. Dion, *Stability and robust stability of time delay systems: A guided tour*, in Stability and Control of Time-Delay Systems, Lecture Notes in Control and Inform. Sci. 228, Springer-Verlag, London, 1997, pp. 1–71.

[26] F. Paganini, Z. Wang, J. C. Doyle, and S. H. Low, *Congestion control for high performance, stability, and fairness in general networks*, IEEE/ACM Trans. Networking, 13 (2005), pp. 43–56.

[27] C. Perkins, *Ad Hoc Networking*, Addison–Wesley, Reading, MA, 2001.

[28] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 2001.

[29] P. Ranjan, R. J. La, and E. H. Abed, *Rate control with communication delay and trade-offs*, in Proceedings of the Annual Allerton Conference, 2003.

[30] P. Ranjan, R. J. La, and E. H. Abed, *Global stability conditions for rate control with arbitrary communication delays*, IEEE/ACM Trans. Networking, 14 (2006), pp. 94–107.

[31] H. R. Varian, *Microeconomic Analysis*, 3rd ed., W. W. Norton and Company, New York, 1992.

[32] G. S. K. Wolkowicz, H. Xia, and S. Ruan, *Competition in the chemostat: A distributed delay model and its global asymptotic behavior*, SIAM J. Appl. Math., 57 (1997), pp. 1281–1310.

[33] L. Ying, G. E. Dullerud, and R. Srikant, *Global stability of Internet congestion controllers with heterogeneous delays*, in Proceedings of the American Control Conference (ACC), 2004.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1268–1281

ANALYTICAL APPROXIMATION FOR THE GENERALIZED LAPLACE EQUATION WITH STEP FUNCTION COEFFICIENT[*]

R. F. SVIERCOSKI[†], C. L. WINTER[‡], AND A. W. WARRICK[§]

Abstract. Many problems in science and engineering require the solution of the steady-state diffusion equation with a highly oscillatory coefficient. In this paper, we propose an analytical approximation $\tilde{u}(x) \in L^p(\Omega)$, $1 \leq p \leq \infty$, for the generalized Laplace equation $\nabla \cdot (K(x) \nabla u(x)) = 0$ in $\Omega \subset R^n$, with prescribed boundary conditions and the coefficient function $K(x) \in L^p(\Omega)$ defined as a step function, not necessarily periodic. The proposed solution can be regarded as an approximation to the weak solution belonging to $W^{1,p}(\Omega)$, the Sobolev space. When the coefficient function describes inclusions in a main matrix, then $K(x)$ is a periodic function, and such formulation leads to an approximation, in $L^p(\Omega)$, to the solution of the periodic cell-problem, $\nabla \cdot (K(\varepsilon^{-1}x)\nabla \mathbf{w}(\varepsilon^{-1}x)) = \nabla \cdot (K(\varepsilon^{-1}x)\mathbf{1})$. The solution to the cell-problem is the key information needed to obtain the upscaled coefficient and therefore the zeroth-order approximation for a generalized elliptic equation with highly oscillating coefficient in $\Omega \subset R^n$. Our numerical computation of the error between the proposed analytical approximation for the cell-problem, $\tilde{w}(\varepsilon^{-1}x)$ in $L^p(\Omega)$, and the solution $w(\varepsilon^{-1}x)$ in $W^{1,2}(\Omega)$, demonstrates to converge in the L^2-norm, when the scale parameter ε approaches zero. The proposed approximation leads to the lower bound of the generalized Voigt–Reiss inequality, which is a more accurate two-sided estimate than the classical Voigt–Reiss inequality. As an application, we compute our approximate value for the homogenized coefficient when the heterogeneous coefficients are inclusions such as squares, circles, and lozenges, and we demonstrate that the results underestimate the effective coefficient with an error of 10% on average, when compared with published numerical results.

Key words. Laplace's equation, L^p-approximation, homogenization, cell-problem, generalized Voigt–Reiss inequality

AMS subject classifications. 35J25, 35R05, 35B27, 74Q20, 76S05

DOI. 10.1137/070683465

1. Introduction.

A major problem in natural porous media is providing an accurate description of the flow's behavior, given the intrinsic heterogeneity of geological formations. By specifying the flow over the boundary of the given domain Ω, one way of quantifying the steady-state flow behavior in such a medium is by computing the solution $u(x) \in W^{1,p}(\Omega)$ of the generalized Laplace equation:

$$\nabla \cdot (K(x) \nabla u) = 0 \quad \text{in} \quad \Omega, \tag{1.1}$$

where the coefficient function $K(x)$ describes the heterogeneity of the medium over Ω. In this paper, we study the case when the coefficient belongs to $L^p(\Omega)$, $1 \leq p \leq \infty$, and can be described as the step function

$$K(x) = \begin{cases} \xi_1 & \text{if } x \in \Omega_c, \\ \xi_2 & \text{if } x \in \Omega \backslash \Omega_c \end{cases} \tag{1.2}$$

and when (1.1) has some particular prescribed boundary conditions, depending on whether Ω_c is defined such that $K(x)$ is either a nonperiodic or a periodic function.

[*] Received by the editors February 22, 2007; accepted for publication (in revised form) July 19, 2007; published electronically April 9, 2008.
http://www.siam.org/journals/siap/68-5/68346.html

[†] Los Alamos National Laboratory, Los Alamos, NM 87545 (rsvier@lanl.gov). This author's research was supported by LANL-IGPP.

[‡] NCAR, Boulder, CO 80305 (lwinter@ucar.edu).

[§] SWES and Program in Applied Mathematics, University of Arizona, Tucson, AZ 85721 (aww@ag.arizona.edu).

The first goal of this paper is to develop an analytical approximation $\tilde{u}(x) \in L^p(\Omega)$, $1 \leq p \leq \infty$, to the weak solution $u(x) \in W^{1,p}(\Omega)$ of the boundary value problem (BVP) (1.1), when the coefficient $K(x)$ is given as (1.2). By considering (1.2) an ε-periodic function, $K\left(\varepsilon^{-1}x\right)$, and using $\tilde{u}(\varepsilon^{-1}x) \in L^p(\Omega)$, satisfying (1.1) almost everywhere (a.e.), the second goal is to propose an analytical approximate value for the homogenized coefficient, $\tilde{K}^0$, and therefore an approximation as $\varepsilon \to 0$, for the zeroth-order solution of the generalized BVP:

$$\left\{ \begin{aligned} \nabla \cdot \left(K\left(\varepsilon^{-1}x\right) \nabla u\left(\varepsilon^{-1}x\right)\right) &= f(x) \qquad \text{in} \quad \Omega, \\ u\left(\varepsilon^{-1}x\right) &= g(x), \qquad x \in \partial\Omega. \end{aligned} \right. \tag{1.3}$$

We present an approximation and its properties without discussing its uniqueness, as this is a subject that requires mathematical tools beyond the scope of this paper.

It is expected that a solution to (1.1) will depend on the type and smoothness of the coefficient function. In the n-dimensions, for particular types of coefficient and boundary conditions, analytical solutions for this problem are known. The solutions follow for the cases when $K(x)$ is a separable function $K(x) = \prod_{i=1}^n k_i\left(x_i\right)$ and when $K(x)$ describes a layered medium in one of the directions. By allowing the coefficient (1.2), it may happen that the solution and its gradient are not differentiable. Then one needs the definition of a weak solution, where (1.1) is replaced by its variational formulation, namely

$$\left\{ \begin{array}{c} \text{Find} \quad u \in W^{1,p} \quad \text{such that} \\ \int_\Omega K(x) \nabla u \nabla v dx = 0 \qquad \forall\, v \in W^{1,q}, \end{array} \right. \tag{1.4}$$

where $W^{1,p}$ is the appropriate Sobolev space taking into account the boundary conditions, and such that $1/p + 1/q = 1$. Existence and uniqueness of the weak solution, under the condition that $K(x)$ is an elliptic and bounded operator, follow from the Lax–Milgram theorem (see [6], for example).

Equations (1.1)–(1.4) have a wide application in mathematics, physics, and engineering. For example, in water flow systems, $u(x)$ is the hydraulic head, $K(x)$ is the hydraulic conductivity, and the flux is defined from Darcy's law $q = -K(x)\nabla u(x)$; for solute diffusion, Fick's law follows the same formulation with $u(x)$ a concentration and $K(x)$ the diffusion coefficient. Other analogous systems can be found, for example, in Warrick [18]. The reader can further refer to Gilbarg and Trudinger [6] for the mathematical formalism and the respective historical contributions. For the applications of (1.1)–(1.4) to flow and transport in porous media, one can consult references such as Bear [2] and Warrick [18].

The approximation to (1.3) is, on its own, a major problem in the subject of multiscale analysis and therefore in many areas of applied mathematics. In particular, this is so in porous media when $K^\varepsilon(x)$ characterizes a periodic inclusion, with size l, in a main matrix with size L and $\varepsilon = l/L$ defining the period. In such a case, the analysis follows by considering the coefficient defined in $\Omega = \bigcup \Omega^\varepsilon$ so that over each Ω^ε one has the step function

$$K\left(\varepsilon^{-1}x\right) = \left\{ \begin{array}{ll} \xi_1 & \text{if } x \in \Omega_c^\varepsilon, \\ \xi_2 & \text{if } x \in \Omega^\varepsilon \backslash \Omega_c^\varepsilon. \end{array} \right. \tag{1.5}$$

In such a context, one proposes that the solution to (1.3) can be approximated by the two-scale asymptotic expansion

$$u\left(\varepsilon^{-1}x\right) = u^0\left(x, \varepsilon^{-1}x\right) + \varepsilon u^1\left(x, \varepsilon^{-1}x\right) + \varepsilon^2 u^2\left(x, \varepsilon^{-1}x\right) + \cdots . \tag{1.6}$$

The question becomes how to obtain the terms in such an expansion and, in particular, the zeroth-order approximation $u^0\left(x, \varepsilon^{-1}x\right)$. The zeroth-order term describes the averaged or macroscopic behavior of the flow, whereas the other terms add microscopic features to the approximation. This is the main scope of homogenization theory. To find $u^0(x, \varepsilon^{-1}x)$ one needs to obtain the homogenized or effective coefficient K^0, which can be accomplished by solving, say for a given ε, the cell-problem

$$\left\{ \begin{array}{lr} \nabla \cdot (K\left(x\right) \nabla \mathbf{w}\left(x\right)) = -\nabla \cdot \left(K\left(x\right) \nabla \mathbf{x}\right) & \text{in} \quad \Omega = (0,1)^n, \\ \mathbf{w}(x) = \sum_{i=1}^n w_i(x), & x \in \partial\Omega, \end{array} \right. \tag{1.7}$$

where $\mathbf{x} = \sum_{i=1}^n x_i$ and $x_i = x \cdot e_i$, with $x = (x_1, x_2, \ldots, x_n)$.

Note that (1.7) is a particular case of (1.1), by setting $u = \mathbf{w} + \mathbf{x}$.

Historical works in homogenization theory include Tartar [17], Keller [8], [9], Bensoussan, Lions, and Papanicolau [3], and Sanchez-Palencia [13]. A complete review of homogenization theory and its application, including the cases of random coefficients, has been compiled in Jikov, Kozlov, and Oleinik [7]. There is also a vast literature on obtaining the effective coefficient by other methods; for example, in Milton [10] one can follow a thorough review and application to a variety of phenomena in composite media. The work by Renard and De Marsily [12] contains a review specifically for upscaling Darcy's law.

We follow the convention that the analytical approximations to (1.1) and (1.7) have the symbol $\tilde{u}$, in contrast to the true solution. In addition, an analytical approximation in L^p means that it solves the respective equation a.e. The results are constrained neither to the periodicity of the coefficient nor to $\Omega = [0,1]^n$, as can be verified shortly.

The presentation is outlined as follows: In section 2, we first look for the approximation $\tilde{w}(x) \in L^p(\Omega)$ to the BVP (1.7), which is obtained by the superposition of $\tilde{w}_i(x) \in L^p(\Omega)$ for each $i = 1, \ldots, n$ and $\partial\Omega = \bigcup_{i=1}^n \Gamma_i$, the approximate solution of the BVP

$$\left\{ \begin{array}{lr} \nabla \cdot \left(K\left(x\right) \nabla w_i\left(x\right)\right) = -\nabla \cdot \left(K\left(x\right) e_i\right) & \text{in} \quad \Omega, \\ w_i\left(x\right) = 0, & x \in \Gamma_i. \end{array} \right. \tag{1.8}$$

In section 3, we consider the case when $K(x)$ is periodic, leading (1.7) and (1.8) to be periodic BVPs. In this case, some properties of interest in homogenization theory are obtained. In section 4, we apply the results to the homogenization and analyze the numerical convergence between our solution in L^2 and the numerical solution in H_0^1 for particular shapes of interest in porous media applications. We also compute the approximate value for the effective coefficient and compare it with some numerical values reported in the literature for inclusions such as squares, circles, and lozenges. This comparison demonstrates an error of about 10% on average.

2. Approximate solutions.

Throughout this section, we consider the two-dimensional version of BVP (1.8), without loss of generality.

THEOREM 2.1. *Let $K(x)$ be defined as in* (1.2) *with e_1 the unit vector in R^2. If $\tilde{w}_1(x_1, x_2) \in L^p(\Omega)$, $i = 1, 2$, can be written as the product $\tilde{w}_1(x_1, x_2) = f_1(x_1, x_2) f_2(x_2)$ for $f_1(x)$, a linear function on x_1, and some $f_2(x_2)$, then the expression*

$$\tilde{w}_1\left(x\right) = \int_0^{x_1} \frac{d\tau}{K\left(\tau, x_2\right)} \left(\int_0^1 \frac{d\tau}{K\left(\tau, x_2\right)}\right)^{-1} - x_1 \tag{2.1}$$

satisfies the BVP

$$\begin{cases} \nabla \cdot (K(x)\, \nabla \tilde{w}_1(x)) = -\nabla \cdot (K(x)\, e_1) & \textit{in} \quad \Omega = (0,1)^2, \\ \tilde{w}_1(x) = 0, & x \in \Gamma_1, \end{cases} \tag{2.2}$$

a.e. in Ω.

Proof. ($\Rightarrow$) Suppose that $\tilde{w}_1(x)$ can be written as $\tilde{w}_1(x) = f_1(x_1, x_2)\, f_2(x_2)$. The boundary conditions on f_1 and f_2 on the $x_1 = 0$ and $x_1 = 1$ faces of Ω will be such that

$$\tilde{w}_1(0, x_2) = f_1(0, x_2)\, f_2(x_2) = \tilde{w}_1(1, x_2) = f_1(1, x_2)\, f_2(x_2) = 0$$

$$\Leftrightarrow f_1(0, x_2) = f_1(1, x_2) = 0$$

for all $f_2(x_2) \neq 0$. Now, we make the substitution of $\tilde{w}_1(x)$, and the equality

$$\nabla \cdot \left(K(x) \left(\frac{\partial f_1}{\partial x_1} f_2(x_2), \frac{\partial \tilde{w}_1}{\partial x_2} \right) \right) = -\nabla \cdot (K(x)\, e_1)$$

must be satisfied a.e. in Ω. The last relationship can be rearranged to give

$$\frac{\partial}{\partial x_1} \left(K(x)\, f_2(x_2) \left(\frac{\partial f_1}{\partial x_1} + \frac{1}{f_2(x_2)} \right) \right) + \frac{\partial}{\partial x_2} \left(K(x) \frac{\partial \tilde{w}_1}{\partial x_2} \right) = 0. \tag{2.3}$$

Setting the first term everywhere in Ω to zero, i.e.,

$$\frac{\partial}{\partial x_1} \left(K(x)\, f_2(x_2) \left(\frac{\partial f_1}{\partial x_1} + \frac{1}{f_2(x_2)} \right) \right) = 0, \tag{2.4}$$

and letting $v(x) = \frac{\partial f_1}{\partial x_1} + \frac{1}{f_2(x_2)}$, we have

$$\frac{\partial}{\partial x_1} (K(x)\, v(x)) = 0 \Rightarrow v(x) = \frac{h_1(x_2)}{K(x)}.$$

This implies the following:

$$\frac{\partial f_1}{\partial x_1} = \frac{h_1(x_2)}{K(x)} - \frac{1}{f_2(x_2)}. \tag{2.5}$$

Furthermore, integrating this last equation over x_1 gives

$$f_1(x_1, x_2) = \int_0^{x_1} \frac{h_1(x_2)\, d\tau}{K(\tau, x_2)} - \frac{x_1}{f_2(x_2)} + d.$$

In order to have $f_1(0, x_2) = 0$, d must be zero. Therefore

$$f_1(x_1, x_2) = \int_0^{x_1} \frac{h_1(x_2)\, d\tau}{K(\tau, x_2)} - \frac{x_1}{f_2(x_2)}. \tag{2.6}$$

The condition $f_1(1, x_2) = 0$ will be satisfied if

$$h_1(x_2) = \frac{1}{f_2(x_2) \int_0^1 \frac{1}{K(\tau, x_2)} d\tau}. \tag{2.7}$$

Now, going back to the expression for $f_1(x_1, x_2)$ above, one ends up with

$$(2.8)\quad f_1(x_1,x_2) = \frac{\int_0^{x_1} \frac{d\tau}{K(\tau,x_2)}}{f_2(x_2)\int_0^1 \frac{1}{K(\tau,x_2)}d\tau} - \frac{x_1}{f_2(x_2)} = \frac{1}{f_2(x_2)}\left[\frac{\int_0^{x_1}\frac{d\tau}{K_1(\tau,x_2)}}{\int_0^1 \frac{1}{K_1(\tau,x_2)}d\tau} - x_1\right].$$

The multiplication by $f_2(x_2)$, just now defined to be $f_2(x_2) = \left(\int_0^1 \frac{1}{K_1(\tau,x_2)}d\tau\right)^{-1}$, leads to $\tilde{w}_1(x_1,x_2)$ in (2.1), which can then be written as

$$(2.9)\quad \tilde{w}_1(x_1,x_2) = \left(\int_0^{x_1}\frac{d\tau}{K(\tau,x_2)} - x_1\int_0^1 \frac{1}{K_1(\tau,x_2)}d\tau\right)\left(\int_0^1 \frac{1}{K_1(\tau,x_2)}d\tau\right)^{-1}.$$

Note also that $f_1(x_1,x_2) = \left(\int_0^{x_1}\frac{d\tau}{K(\tau,x_2)} - x_1\int_0^1 \frac{1}{K_1(\tau,x_2)}d\tau\right)$ is indeed a linear function of x_1. To finalize, we look back at the sum (2.3), and since the first term is zero and the sum is zero a.e., it follows that the second term is zero a.e.

($\Leftarrow$) Let $\tilde{w}_1(x)$ as in (2.1) be given as

$$(2.10)\qquad \tilde{w}_1(x) = \frac{\int_0^{x_1}\frac{d\tau}{K(\tau,x_2)}}{\int_0^1 \frac{d\tau}{K(\tau,x_2)}} - x_1 = \tilde{u}_1(x) - x_1.$$

We make the substitution into (2.2) to get

$$\nabla\cdot(K(x)\nabla\tilde{w}_1(x)) = \nabla\cdot(K(x)\nabla\tilde{u}_1(x)) - \nabla\cdot(K(x)e_1) = -\nabla\cdot(K(x)e_1) \quad \text{a.e.}$$

This is true since the equation in the middle can be written as

$$(2.11)\quad \frac{\partial}{\partial x_1}\left(K(x)\frac{\partial\tilde{u}_1}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(K(x)\frac{\partial\tilde{u}_1}{\partial x_2}\right) - \frac{\partial}{\partial x_1}(K(x)e_1) = -\frac{\partial}{\partial x_1}(K(x)e_1).$$

On the left-hand side, the first term is zero, and the second term may be extended continuously as zero, even though it is not zero at the discontinuity points of $K(x)$ in the x_2-direction, say x_2^*. The extension can be formalized as follows:

$$(2.12)\qquad \lim_{x_2\to x_2^*-}\frac{\partial\tilde{u}_1}{\partial x_2}(x) = \lim_{x_2\to x_2^*+}\frac{\partial\tilde{u}_1}{\partial x_2}(x) = 0.$$

Moreover, this is true since $K(x)$ is defined as in (1.2).

The same procedure applies to obtaining $\tilde{w}_2(x) = \tilde{u}_2(x) - x_2$, which corresponds to (2.2) in the e_2-direction. In two dimensions, one can think of $\tilde{w}_2(x)$ as the transpose of $\tilde{w}_1(x)$. □

A natural question regards the estimate of the error of this approximation, which requires us to analyze the properties of $\frac{\partial\tilde{u}_1}{\partial x_2}$, such as boundedness, and how it relates to $\frac{\partial u_1}{\partial x_2}$. One possible way to address this problem is to identify a mollifier or Green's function $\rho(x) = \rho(K(x)) \in C^\infty(\Omega)$ such that $\int_\Omega \rho(x)dx = 1$ and the convolution

$$(2.13)\qquad w_i^h(x) = h^{-n}\int_\Omega \rho_h(K(x-\zeta))\tilde{w}_i(\zeta)d\zeta$$

with $h < \text{dist}(x,\partial\Omega)$. Then $\tilde{w}_i$ converges to $w_i^h \in W^{1,p}(\Omega)$ in the sense of $L^p(\Omega)$, $p < \infty$ (see Lemma 7.2 of [6]).

However, finding $\rho(x)$ and proving (2.13) is a problem related to the uniqueness of the approximation and would require a deeper analysis beyond the present work. Instead, we present some interesting and relevant results that may be of help when addressing the problem in the future.

Illustrations of the approximation $\tilde{w}_1(x)$, comparing it with the numerical solution in $H_0^1(\Omega)$, are shown in Figures 1 and 2 by using $K(x)$ with square inclusions. Also, an illustration of the functions $\frac{\partial \tilde{u}_1}{\partial x_2}$ and $\frac{\partial \tilde{u}_2}{\partial x_1}$ is given in Figure 3 for the respective functions. Observe also that $\rho(x)$, from (2.13), would smooth out the sharp edges presented in the figures.

The generalization to $[0,1]^n \subset R^n$ follows by solving each subproblem as in (2.2). In such a case, the product form for the ith subproblem is given as $\tilde{w}_i(x) = f_i(x) f_j(x_j)$, where $f_i(x)$ is a function linear in x_i and $f_j(x_j)$ is a function of all the variables but x_i. Applying Theorem 2.1, we find that

(2.14)

$$\begin{aligned}\tilde{w}_i\left(x\right) &= \int_0^{x_i} \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)} \left(\int_0^1 \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)}\right)^{-1} - x_i = \tilde{u}_i(x) - x_i \\ &= \left[\int_0^{x_i} \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)} - x_i \int_0^1 \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)}\right] \\ &\quad \cdot \left(\int_0^1 \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)}\right)^{-1} \\ &= f_i(x) f_j(x_j)\end{aligned}$$

is the approximate solution in $L^p(\Omega)$ for $w_i(x)$, the solution to (1.8). Moreover, we have the following consequences relating the solutions of (1.8) to those of (1.7) and (1.1), for particular cases of prescribed boundary conditions.

Corollary 2.2. *Let $\tilde{w}_i(x)$ be given as* (2.14)*; then $\tilde{u}_i(x) \in L^p(\Omega)$ approximates the solution of the BVP*

$$\left\{ \begin{array}{rl} \nabla \cdot \left(K\left(x\right) \nabla u_i\left(x\right)\right) = 0 & \mathit{in} \quad \Omega = (0,1)^n, \\ u_i(x) = x_i, & x \in \partial\Omega. \end{array} \right. \tag{2.15}$$

The next results are derived by using the superposition principle.

Corollary 2.3. *Let $\tilde{w}_i(x)$ be given as* (2.14)*; then $\tilde{\mathbf{w}}(x) = \sum_{i=1}^n \tilde{w}_i(x) \in L^p(\Omega)$ approximates the solution to the BVP* (1.7).

Corollary 2.4. *Let $\tilde{w}_i(x)$ be given as* (2.14)*; then $\tilde{u}(x) = \sum_{i=1}^n \tilde{u}_i(x) \in L^p(\Omega)$ approximates the solution of the BVP*

$$\left\{ \begin{array}{rl} \nabla \cdot \left(K\left(x\right) \nabla u\left(x\right)\right) = 0 & \mathit{in} \quad \Omega = (0,1)^n, \\ u(x) = \sum_{i=1}^n x_i, & x \in \partial\Omega. \end{array} \right. \tag{2.16}$$

The integration procedure used to obtain (2.14) can be applied to general bounded rectangular domains. In particular, consider $\Omega_\varepsilon = [0,\varepsilon]^n$; then $h_1(x_j)$, $j \neq i$, in (2.7) becomes

$$h_1\left(x_j\right) = \frac{\varepsilon}{f_j\left(x_j\right) \int_0^\varepsilon \frac{1}{K(\tau, x_j)} d\tau}, \tag{2.17}$$

leading to

$$\tilde{w}_i\left(x\right) = \varepsilon \int_0^{x_i} \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)} \left(\int_0^\varepsilon \frac{d\tau}{K\left(x_1, \ldots, \tau, \ldots, x_n\right)}\right)^{-1} - x_i. \tag{2.18}$$

The x variable can be normalized to $\Omega = [0,1]^n$. The last equation then becomes

$$\tilde{w}_i^\varepsilon(x) = \varepsilon \left[\frac{\int_0^{\varepsilon^{-1}x_i} \frac{d\xi}{K(\varepsilon^{-1}x_1,\ldots,\xi,\ldots,\varepsilon^{-1}x_n)}}{\int_0^1 \frac{d\xi}{K(\varepsilon^{-1}x_1,\ldots,\xi,\ldots,\varepsilon^{-1}x_n)}} - \varepsilon^{-1}x_i \right], \tag{2.19}$$

where ε may represent the period of Ω_ε in the i-direction of Ω. One has the following result.

Corollary 2.5. *Let $\tilde{w}_i^\varepsilon(x)$ be as defined above; then*

$$\lim_{\varepsilon\to 0} \tilde{w}_i^\varepsilon(x) = 0 \quad in \quad L^p(\Omega). \tag{2.20}$$

Proof. The proof follows by the definition of $\tilde{w}_i^\varepsilon(x)$ and by noting that the ratio in brackets is a bounded function for all ε, since $x_i \le \varepsilon$. □

This result is also in agreement with the numerical $w_i(\varepsilon^{-1}x)$, as will be illustrated in section 4.1 (Tables 1 and 2) for particular cases of periodic inclusion.

The additional condition on the geometry of a periodic coefficient, outlined in the next corollary, allows one to obtain results that are relevant in homogenization theory presented in section 4.

3. Periodic coefficients.

By considering $K(x)$ a periodic function, it can be verified that the solutions to (1.7)–(1.8) are periodic. Moreover, if $K(x)$ is constant throughout the boundary, then the approximation to (1.8) leads to $w_i(x) = 0$, $x \in \partial\Omega$, and $\mathbf{w}(x) = 0$, $x \in \partial\Omega$, in (1.7). However, some additional properties are presented for the particular case when $K(x)$ has its center of mass at half of the period of the unit cell. This symmetry leads to properties of the functions $f_i(x)$ and $f_j(x)$ defined in (2.14).

Corollary 3.1. *Let $\Omega = [0,1]^n$. If $K(x)$ is periodic and has its center of mass at $x = \left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$, then the following relationships are true:*

(i) $\int_\Omega \tilde{w}_i(x)\,dx = 0$.

(ii) $\int_\Omega \tilde{w}_i\tilde{w}_j dx = 0$ *for $j \ne i$, where $\tilde{w}_j(x)$ is the respective approximation to* (1.8) *in the jth direction.*

(iii) $\int_\Omega K(x) \frac{\partial \tilde{w}_i}{\partial x_j} dx = 0$ *for $j \ne i$.*

Proof. (i) There are two proofs (a) and (b) of this result. (a) First, this is true by defining an appropriate constant of integration c in (2.1). Moreover, if $N(x) = \frac{1}{K(x)}$ has its center of mass at $x_i = \frac{1}{2}$, then c is zero. Indeed, $\tilde{w}_i(x)$ in (2.14) becomes

$$\tilde{w}_i(x) = \frac{\int_0^{x_i} \frac{d\tau}{K(x_1,\ldots,\tau,\ldots,x_n)} - x_i \int_0^1 \frac{d\tau}{K(x_1,\ldots,\tau,\ldots,x_n)}}{\int_0^1 \frac{d\tau}{K(x_1,\ldots,\tau,\ldots,x_n)}} = \frac{\int_0^{x_i} N(\tau)d\tau - x_i \int_0^1 N(\tau)d\tau}{\int_0^1 N(\tau)d\tau}. \tag{3.1}$$

By exploiting the triangular domain of integration, recall that

$$\int_0^1 \int_0^{x_i} N(\tau)d\tau dx_i = \int_0^1 (1-\xi)N(\xi)d\xi.$$

In particular, $\int_0^1 x_i dx_i = \int_0^1\int_0^{x_i} d\tau dx_i = \int_0^1 (1-\xi)d\xi$. Therefore, computing the average of $\tilde{w}_i(x)$ and performing iterated integration in the numerator of (3.1), we solve the ith integral to give

$$\int_0^1 \tilde{w}_i\,(x)\,dx_i = \int_0^1 (1-\xi)N(\tau)d\tau - \int_0^1 (1-\xi)\int_0^1 N(\xi)d\xi d\xi \tag{3.2}$$
$$= \int_0^1 \left[(1-\xi)N(\xi) - (1-\xi)\int_0^1 N(\tau)d\tau\right] d\xi$$
$$= -\int_0^1 \xi N(\xi)d\xi + \int_0^1 \xi d\xi \int_0^1 N(\tau)d\tau,$$

and (3.2) is zero if and only if

$$\frac{\int_0^1 \xi N(\xi)d\xi}{\int_0^1 N(\tau)d\tau} = \frac{1}{2}.$$

Thus, in order for $\tilde{w}_i(x)$ to have zero mean, we need $N(x)$ to have its center of mass at half of the period in the x_i variable. Equivalently, $N(x)$ has its first moment at half of the period.

(b) Note that $\tilde{w}_i\,(x)$ is an odd function w.r.t. x_i and an even function w.r.t. the x_j variables. Indeed, considering $\tilde{w}_i\,(x) = f_i(x)f_j(x_j)$ in (2.14), we observe that $f_i(x)$ is odd and symmetric w.r.t. $x_i = 1/2$, and $f_j(x_j)$ is an even function because $K(x)$ is an even function.

(ii) By having $K(x)$ with its center of mass at $x = \frac{1}{2}$, it follows that $K(x)$ is an even function w.r.t. $x = \frac{1}{2}$. By using the argument from (b) applied to $\tilde{w}_j(x) = g_j(x)g_i(x_i)$, the approximation to the jth problem, $\tilde{w}_j(x)$ is an even function w.r.t. x_i. Thus $\tilde{w}_i(x)$ is orthogonal to $\tilde{w}_j(x)$.

(iii) The proof of (iii) follows since $\frac{\partial \tilde{w}_i}{\partial x_j}$ is an odd function and $K(x)$ is an even function. $\square$

Figures 1, 2, and 3 illustrate the properties (i)–(iii) proved in Corollary 3.1. It also illustrates that they are common properties between $\tilde{w}_1(x)$ and the numerical solution $w_1(x) \in H_0^1(\Omega)$. Another common property is discussed in section 4.2.

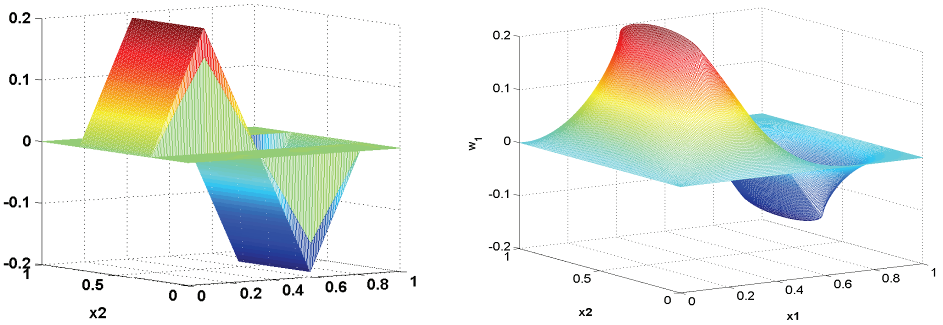


FIG. 1. $\tilde{w}_1\,(x)$ *(left) and numerical* $w_1\,(x)$ *(right), using* $K(x)$ *as in* (1.2)*, with* $\xi_1 = 10$ *and* $\xi_2 = 1$ *and one square inclusion, centered at* $(0.5, 0.5)$ *with an area equal to* $\frac{1}{4}$.

4. Application to homogenization theory. For the purpose of this section we constrain the analysis to L^2 and H^1 spaces and the properties and results obtained from Corollary 3.1 above. In order to illustrate the homogenization procedure, we consider composite materials as media with microstructures on a scale much smaller than the macroscopic scale of interest. The macroscopic length scale, L, is the dimension of a reservoir or a typical wavelength. The characteristic length of the medium

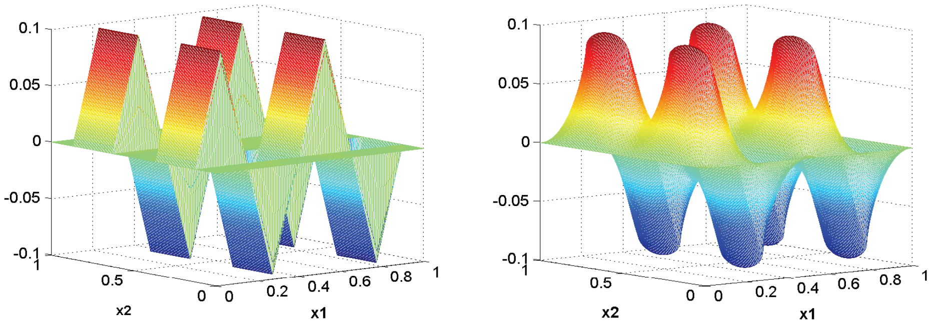


FIG. 2. *$\tilde{w}_1(x)$ (left) and numerical $w_1(x)$ (right), using $K(x)$ as in* (1.2)*, with $\xi_1 = 10$ and $\xi_2 = 1$ and four square inclusions, centered at* $(0.25, 0.25), (0.75, 0.25), (0.25, 0.75)$*, and* $(0.75, 0.75)$*, respectively, with total area equal to $\frac{1}{4}$.*

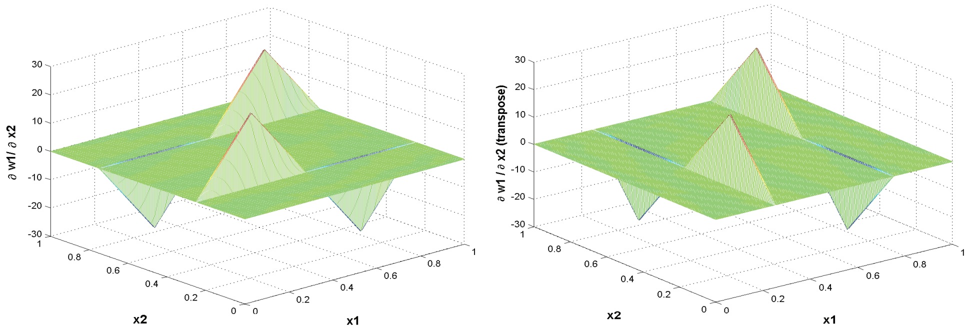


FIG. 3. $\frac{\partial \tilde{w}_1}{\partial x_2}$ *and its transpose, i.e.,* $\frac{\partial \tilde{w}_2}{\partial x_1}$ *from $\tilde{w}_1(x)$ in Figure* 1.

configuration is denoted by l, and the ratio between l and L is denoted by $\varepsilon = \frac{l}{L}$. In the study of physical processes in media with microstructure, known and unknown quantities are dependent on ε, and an asymptotic analysis is used to determine the unknown field quantities. To make precise the fact that the medium varies rapidly on the small scale l and may also vary slowly on the large scale L, we assume that the coefficient is of the form $K^{\varepsilon}(x) = K\left(\varepsilon^{-1}x\right) = K(y)$. One looks for a solution $u(x, y) = u^{\varepsilon}(x)$ given by the two-scale asymptotic expansion

$$u^{\varepsilon}(x) = u^0(x, y) + \varepsilon u^1(x, y) + \varepsilon^2 u^2(x, y) + \cdots, \tag{4.1}$$

where $x = (x_1, x_2, \ldots, x_n)$ is a vector in R^n called the global variable and $y = \varepsilon^{-1}x = (y_1, y_2, \ldots, y_n)$ is the local variable. The two-scale differentiation, $\nabla = \nabla_x + \varepsilon^{-1}\nabla_y$, used in the substitution of $u^{\varepsilon}(x)$ into the BVP,

$$\begin{cases} \nabla \cdot (K^{\varepsilon}(x) \nabla u^{\varepsilon}(x)) = f(x), & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega, \end{cases} \tag{4.2}$$

leads to a rigorous deductive procedure for obtaining the macroscopic equations (in x) based upon solutions of local equations (in y) (examples are given in [3], [5], [7], among others). The equations in y are solvable if the microstructure is locally periodic and the terms $u^i(x, y)$ in the expansion are periodic in the y variable with the same period as the structure. By doing so, homogenization has two objectives:

(i) Determine what equation $u^0(x)$ satisfies, where $u^0(x)$ is

$$u^0(x) = \lim_{\varepsilon \to 0} u^\varepsilon(x).$$

(ii) Determine in what sense the last limit needs to be considered.
To address these questions one has the following definition (from [7]).

Definition 4.1 (H-convergence). *A constant matrix K^0 is said to be the homogenized limit of K^ε if and only if for any bounded domain $\Omega \subset R^n$ and for any $f \in H^{-1}(\Omega)$ the solutions $u^\varepsilon \in H^1(\Omega)$ of the BVP (4.2) possess the properties*

$$u^\varepsilon \rightharpoonup u^0 \quad in \quad H^1(\Omega), \tag{4.3}$$

$$K^\varepsilon(x)\nabla u^\varepsilon(x) \rightharpoonup K^0 \nabla u^0(x) \quad in \quad L^2(\Omega) \tag{4.4}$$

as $\varepsilon \to 0$, where $u^0 \in H^1(\Omega)$ is the solution of the BVP

$$\nabla \cdot \left(K^0 \nabla u^0(x)\right) + f(x) = 0. \tag{4.5}$$

Using the *Kronecker delta* δ_{ij}, the homogenized coefficient K^0 is defined by

$$K^0_{ij} = \int_Y K(y)\left(\delta_{ij} + \partial_{y_i} w_i(y)\right) dy, \tag{4.6}$$

where $w_i(y) \in H^1(\Omega)$ is the solution of the periodic cell-problem

$$\nabla \cdot \left(K(y)\nabla_y w_i(y)\right) = -\nabla \cdot \left(K(y)e_i\right) \quad \text{for} \quad y \in Y \tag{4.7}$$

and e_i is the unit vector in the i-direction. By using the approximation $\tilde{w_i}(y)$ from (2.1) in (4.7) and (iii) from Corollary 3.1, one computes an approximation to (4.6):

$$\tilde{K}^0 = \operatorname{diag} \int_Y (R_1, R_2, \ldots, R_i, \ldots, R_n)\, dY, \tag{4.8}$$

where $R_i = \left(\int_0^1 \frac{d\tau}{K(y_1,\ldots,\tau,\ldots,y_n)}\right)^{-1}$ is the harmonic average in the e_i-direction. Therefore, $\tilde{K}^0$ is the arithmetic average of the harmonic average. One important observation is that (4.8) reduces to three well-known results in the literature: the harmonic average in one dimension, the arithmetic and harmonic averages for layered media, and the arithmetic average of the harmonic average for the case when $K(y)$ is a separable function. Moreover, one has the following result.

Corollary 4.2. *$\tilde{K}^0$ is the lower bound of the generalized Voigt–Reiss inequality:*

$$\tilde{K}^0 \leq K^0 \leq K^u, \tag{4.9}$$

where $K^u = \left(\int_Y \frac{dy_j}{\int_Y K(y)dy_i}\right)^{-1}$, with $j \neq i$.

Proof. The same expression for $\tilde{K}^0$ was obtained in Jikov, Kozlov, and Oleinik [7, eq. (1.74)]; however, they had used a variational principle argument instead of the approximation (2.1). □

This shows how the proposed approximate solution fits into the classical framework of homogenization theory. Inequality (4.9) is a more accurate estimate than the classical Voigt–Reiss inequality, which states that K^0 lies between the harmonic and arithmetic averages of $K(y)$. As will be seen in the next section, $\tilde{K}^0$ indeed always underestimates the numerical values for K^0, and for completeness we also computed K^u. What is surprising, though, is that the error is about 10% on average (see [14] and section 4.2). A corrector is needed in order to obtain K^0, which may arise by finding the mollifier on (2.13) or other analytical means that can take advantage of the proposed approximation. This was proposed in Sviercoski, Travis, and Hyman [15], [16].

4.1. Convergence properties. We have proved in Corollary 2.5 that $\tilde{w}_1^\varepsilon(x) \to 0$ linearly. Tables 1 and 2 illustrate that this is another common property of $\tilde{w}_1^\varepsilon(x) \in L^2(\Omega)$, the approximation to the periodic cell-problem (1.8), and the numerical solution $w_1^\varepsilon(x) \in H^1(\Omega)$, in addition to the symmetries illustrated in Figures 1 and 2. The fact that each of them is going to zero implies that their difference is going to zero in the L^2-norm. The functions $\tilde{w}_1^\varepsilon(x)$ in the tables were obtained by setting up the cell-problem (1.8) considering the $K^\varepsilon(x)$ function as in (1.5) and constructing the sequence by taking its value as $(0.5)^n$, where $n = 1, 2, 3, 4, 5, 6$ and $\Omega = \bigcup \Omega^\varepsilon$ so that over each Ω^ε, Ω_c^ε are taken to be either square or circular inclusions. Furthermore, $K^{0.5}(x)$ refers to having one inclusion over the unit domain with a resulting area of (0.25); $K^{0.25}(x)$ has four inclusions, 2 × 2, in the unit domain with the same total area of (0.25), up to $K^{(0.5)^6}(x)$ with 64×64 inclusions and the same total area of the inclusions. The same is done to obtain the respective numerical solution $w_1^\varepsilon(x)$.

Table 1
Square inclusions with $\xi_1 = 100$ *and* $\xi_2 = 1$.

ε	$\|\|\tilde{w}_1^\varepsilon(x)\|\|_2$	$\|\|w_1^\varepsilon(x)\|\|_2$	$\|\|\tilde{w}_1^\varepsilon(x) - w_1^\varepsilon(x)\|\|_2$
$(0.5)^1$	9.8455e-2	1.0563-1	4.5214e-2
$(0.5)^2$	5.5729e-2	5.4217e-2	2.5796e-2
$(0.5)^3$	2.4305e-2	2.7473e-2	1.3249e-2
$(0.5)^4$	1.1814e-2	1.3904e-2	7.0251e-3
$(0.5)^5$	5.2731e-3	7.2277e-3	4.2687e-3
$(0.5)^6$	2.1386e-3	3.6724e-3	2.4792e-3

Table 2
Circular inclusions with $\xi_1 = 10$ *and* $\xi_2 = 1$.

ε	$\|\|\tilde{w}_1^\varepsilon(x)\|\|_2$	$\|\|w_1^\varepsilon(x)\|\|_2$	$\|\|\tilde{w}_1^\varepsilon(x) - w_1^\varepsilon(x)\|\|_2$
$(0.5)^1$	8.0369e-2	7.8305e-2	2.2019e-2
$(0.5)^2$	4.0110e-2	3.9828e-2	1.2707e-2
$(0.5)^3$	2.0405e-2	2.0200e-2	6.9455e-3
$(0.5)^4$	1.0203e-2	1.0161e-2	3.7447e-3
$(0.5)^5$	5.1821e-3	5.0776e-3	2.2066e-3

4.2. Comparison between numerical results and $\tilde{K}^0$. We compute the approximate value (4.8) and compare it with some published numerical results. The numerical values have been taken from Bourgat [4], Amaziane, Bourgeat, and Koebbe [1], and Moulton, Dendy, and Hyman [11]. Note that in [1], the homogenization procedure has been applied to a nonlinear two-phase flow equation. For some particular cases of nonlinearity, the value of the effective coefficient can be computed as in the linear case.

4.2.1. Case 1. In [1], each function $K^\varepsilon(x)$ is defined as in (1.5), with values of ξ_1 being in the inclusion and ξ_2 in the main matrix. (Note that this is the opposite of the definition used elsewhere in this paper). The effective value $K^\#$ was obtained by numerically solving (1.3) for four tests:

Test 1. $K^\varepsilon(x)$ with $\xi_1 = 1$ and $\xi_2 = 10$.

Test 2. The same $K^\varepsilon(x)$ as in Test 1. However, a two-phase analysis was used with a change in the viscosity ratios.

Test 3. $K^\varepsilon(x)$ with $\xi_1 = 1$ and $\xi_2 = 100$.

Test 4. $K^\varepsilon(x)$ with $\xi_1 = 1$ and $\xi_2 = 10$, as illustrated in Figure 4 (right).

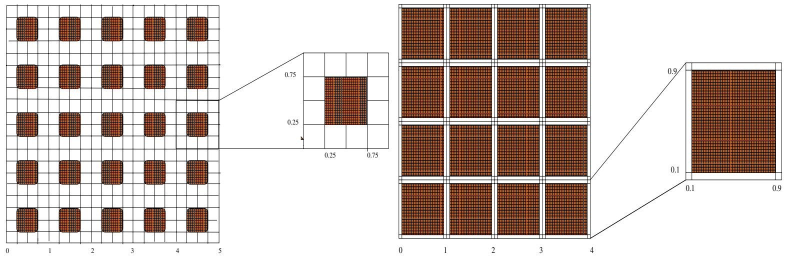


FIG. 4. *Left: Function $K(y)$ used in Tests 1, 2, and 3, respectively. Right: $K(y)$ used in Test 4.*

The results are presented in Table 3, where K^h is the harmonic average, $K^{\#}$ is numerical from [1], $\tilde{K}^0$ results from (4.8), and K^a is the arithmetic average. RD is the relative difference, $RD = \frac{|K^{\#}-\tilde{K}^0|}{\tilde{K}^0}$. The difference between the numerical and analytical results is about 10%, on average. Observe that in Test 4, the error is smaller compared to the others. This coincides with the fact that the medium is "less heterogeneous" than the others. An analogous argument applies for the "highly heterogeneous" Test 3, where the error is the largest.

TABLE 3
Comparison between (4.8) and numerical values from [1].

	K^h	$\tilde{K}^0$	$K^{\#}$	K^u	K^a	RD
Test 1	3.09	**5.91**	6.52	7.09	7.75	10.3 %
Test 2	3.09	**5.91**	6.52	7.09	7.75	10.3 %
Test 3	3.89	**51.0**	59.2	67	75.2	16 %
Test 4	1.48	**2.98**	3.106	3.271	4.24	4 %

4.2.2. Case 2. Table 4 shows the comparison between values of the effective coefficient, K^{bb}, obtained by the black box multigrid algorithm, from [11], the numerical asymptotic value, K^{num}, computed in [4], and the analytical form (4.8). The idea is also to show the dependence of the results on the shape of the inclusions. Three different inclusions are presented with an area equal to $\frac{1}{4}$, as in Figure 5 with $\xi_1 = 10$ and $\xi_2 = 1$, where RD is computed with K^{num} as it is more closely related to our procedure. Table 4 also indicates that $\tilde{K}^0$ is consistent with the others w.r.t. the correspondence between the order of the values and their respective shapes. Indeed, the circular shape renders the smallest effective value, whereas the lozenge gives the largest.

TABLE 4
Comparison between (4.8) and numerical values from [4] and [11].

Shape	K^h	$\tilde{K}^0$	K^{num}	K^{bb}	K^u	K^a	RD
Square	1.292	**1.409**	1.548	1.598	1.695	3.26	9 %
Circle	1.291	**1.403**	1.516	1.563	1.791	3.251	8 %
Lozenge	1.288	**1.417**	1.573	1.608	1.936	3.236	11 %

5. Discussions. The results presented in this paper are one step towards obtaining the solution of a generalized Laplace equation not only for the coefficient functions presented here but also for more general multiscaled geometries. They also provide

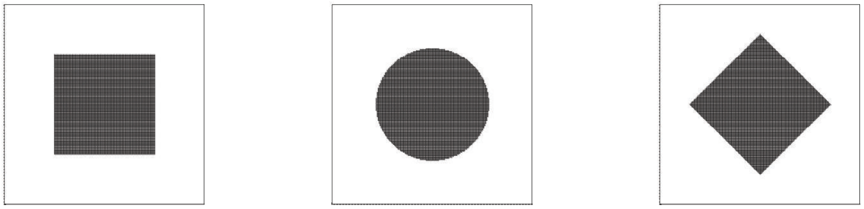


Fig. 5. *Different shape of inclusion.*

insight on deriving, by analytical means, the effective coefficient for the generalized flow equation with a rapidly oscillating coefficient. Other simulations, not presented here, indicate that when using (4.8) and the approximation in Definition 4.1, one gets the boundedness of the sequence, $\left|\left|u^{\varepsilon} - u^{0}\right|\right|_{2} \leq C$, that is, a strong convergence up to a subsequence of u^{ε}. It is possible that, by defining a mollifier in order to smooth out the various approximations presented, the numerical and analytical values for the effective coefficient will be closer. In this sense, we are proposing in Sviercoski, Travis, and Hyman [15], [16] an analytical corrector to (2.14) and (4.8) to overcome the demonstrated difference.

Acknowledgments. The authors are grateful to Mac Hyman, Bryan Travis, and Peter Popov for useful discussions.

REFERENCES

[1] B. Amaziane, A. Bourgeat, and J. Koebbe, *Numerical simulation and homogenization of two-phase flow in heterogeneous porous media*, Transp. Porous Media, 6 (1991), pp. 519–547.

[2] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1972.

[3] A. Bensoussan, J.-L. Lions, and G. Papanicolau, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, New York, 1978.

[4] J. F. Bourgat, *Numerical experiments of the homogenization method for operators with periodic coefficients*, in Computing Methods in Applied Sciences and Engineering (Versailles, 1977), I. R. Glowinski and J.-L. Lions, eds., Springer-Verlag, Berlin, 1979, pp. 330–356.

[5] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math. Appl. 17, The Clarendon Press, Oxford University Press, New York, 1999.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[7] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[8] J. B. Keller, *Effective behavior of heterogeneous media*, in Statistical Mechanics and Statistical Methods in Theory and Application, U. Landman, ed., Plenum, New York, 1977, pp. 631–644.

[9] J. B. Keller, *Darcy's law for flow in porous media and the two-space method*, in Nonlinear Partial Differential Equations in Engineering and Applied Sciences, R. L. Sternberg, ed., Dekker, New York, 1980, pp. 429–443.

[10] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.

[11] J. D. Moulton, J. E. Dendy, and J. M. Hyman, *The black box multigrid numerical homogenization algorithm*, J. Comput. Phys., 142 (1998), pp. 80–108.

[12] Ph. Renard and G. De Marsily, *Calculating equivalent permeability: A review*, Adv. in Water Res., 20 (1997), pp. 253–278.

[13] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer-Verlag, Berlin, 1980.

[14] R. F. Sviercoski, *Multiscale Analytical Solutions and Homogenization of n-Dimensional Generalized Elliptic Equations*, Ph.D. dissertation, University of Arizona, Tucson, AZ, 2005.

[15] R. F. Sviercoski and B. J. Travis, *Analytical effective coefficient for block inclusions: Numerical results for linear flow*, Transp. Porous Media, submitted.

[16] R. F. Sviercoski, B. J. Travis, and J. M. Hyman, *Analytical effective coefficient and first-order approximation for linear flow through block permeability inclusions*, Comput. Math. Appl., 55 (2008), pp. 2118–2133.

[17] L. Tartar, *Quelques remarques sur l'homogeneisation*, in Functional Analysis and Numerical Analysis, Proceedings of the Japan-France Seminar, H. Fujita, ed., Japanese Society for the Promotion of Science, Tokyo, 1978, pp. 469–481.

[18] A. W. Warrick, *Soil Water Dynamics*, Oxford University Press, New York, 2003.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1282–1303

LOCAL TOMOGRAPHY IN ELECTRON MICROSCOPY*

ERIC TODD QUINTO† AND OZAN ÖKTEM‡

Abstract. We present a new local tomographic algorithm applicable to electron microscope tomography. Our algorithm applies to the standard data acquisition method, single-axis tilting, as well as to more arbitrary acquisition methods including double axis and conical tilt. Using microlocal analysis we put the reconstructions in a mathematical context, explaining which singularities are stably visible from the limited data given by the data collection protocol in the electron microscope. Finally, we provide reconstructions of real specimens from electron tomography data.

Key words. X-ray transform, limited data tomography, microlocal analysis, Lambda tomography, electron tomography

AMS subject classifications. Primary, 44A12, 92C55; Secondary, 65R30, 35S30

DOI. 10.1137/07068326X

1. Introduction.

Our goal is to show how singularity detection algorithms can be useful in *electron (microscope) tomography (ET)*. Briefly, given *transmission electron microscope (TEM)* data and using principles of tomography, the goal in ET is to reconstruct the scattering potential of the individual molecules in an in situ (in the cellular environment) or in vitro (in aqueous environment) specimen, each of which can be fairly arbitrary. Because the specimen extends far beyond the area exposed to the electrons, the exposed region covers only a small subregion, which is usually referred to as the *region of interest*. Again, because of the size of the whole specimen, one can rotate it only in a limited range of angles, so the reconstruction problem is a *limited angle* problem. These imply that one has *nonuniqueness* and *severe ill-posedness*. Nonuniqueness, as illustrated in Example 3.1, means that one cannot exactly reconstruct the scattering potential of the specimen even in cases when one assumes exact data (no measurement errors) and disregards the discretization of the set of lines (i.e., one deals with the corresponding continuous problem where data are given over a continuous set of lines). Furthermore, as discussed in section 2.2, the data are very noisy, in particular because of the dose problem—the dose needed to get low-noise data destroys the specimen. Since the limited angle problem leads to severe ill-posedness, the reconstruction problem is unstable and the noise in the data is amplified.

These issues, namely nonuniqueness and ill-posedness, point to using a reconstruction method that regularizes by reconstructing only some information about the specimen that can be stably retrieved, in our case the shape of the boundaries of the molecules in the specimen. Our method is a generalization of Lambda tomography [6, 28].

The article is organized as follows. In section 2, we give the background from physics and state the inverse problem. To provide perspective, we briefly describe

*Received by the editors February 21, 2007; accepted for publication (in revised form) December 26, 2007; published electronically April 9, 2008. This work was supported by the National Science Foundation under grants DMS-0200788 and DMS-0456858 and by the Wenner Gren Center, Tufts University FRAC, the Humboldt Stiftung, and Sidec.

http://www.siam.org/journals/siap/68-5/68326.html

†Department of Mathematics, Tufts University, 503 Boston Avenue, Medford, MA 02155 (todd.quinto@tufts.edu).

‡Sidec Technologies, Torshamnsgatan 28A, SE-164 40 Kista, Sweden (ozan.oktem@sidec.com).

planar Lambda tomography in section 3. Next, we provide our general algorithm for arbitrary data sets in section 4. Then, in section 5 we give the specifics for single-axis tilt ET. In section 6, we describe which singularities of the specimen are stably recoverable from the limited data in ET. We put our results in the context of microlocal analysis as done by [20] for planar CT. This is based on the theory of Fourier integral operators, and the specific results are based on very general theorems in [8] or [2]. Finally, in section 7, we give reconstructions from real data to illustrate the efficiency of our algorithm and demonstrate our characterization of stably recoverable singularities. The appendix includes proofs of our technical theorems.

2. Electron (microscope) tomography (ET).

In what follows we will provide a very brief overview of ET, where our aim is to properly state the inverse problem and show how integral geometry can be used to solve it. The reader is referred to [4] and the references therein for a more detailed account.

2.1. Scientific application and experimental setting.

The problem of recovering the three-dimensional structure of an individual molecule (e.g., a protein or a macromolecular assembly) at the highest possible resolution in situ or in vitro plays a central role in understanding biological processes in time and space. Established approaches, such as X-ray crystallography and nuclear magnetic resonance (NMR), for dealing with this problem cannot recover the structure of an *individual* molecule in a sample. The publication of [3, 29, 11] in 1968 marked the beginning of ET, where the idea of recovering the structure of a sample from TEM data using principles of tomography was first outlined. ET is currently the *only* approach that allows one to reconstruct the three-dimensional structure of *individual* molecules in in situ/in vitro samples. The main drawback of ET when compared to NMR/X-ray crystallography, mentioned earlier, is that it provides only a low-resolution structure due to reasons explained in section 2.2. However, since the ability to study individual molecules is important in order to address many biological problems, ET is nowadays enjoying an increasing interest within life sciences as a technique for low- or medium-resolution structure determination of individual molecules.

A specimen that is to be imaged in a TEM must first be physically immobilized since it is imaged in a vacuum. It also needs to be thin (about 70–100 nm) if enough electrons are to pass through to form an image. The purpose of sample preparation is to achieve this *without interfering with the structure of the specimen.* Data collection in ET is done by mounting the specimen on a holder (goniometer) that allows one to change its positioning relative to the optical axis of the TEM. For a fixed position, the specimen is radiated with an electron beam, and the resulting data, referred to as a *micrograph*, is recorded by a detector. Hence, each fixed orientation of the specimen yields one micrograph, and the procedure is then repeated for a set of different positions. The most common data acquisition geometry is *single-axis tilting*, where the specimen plane is allowed only to rotate around a fixed single axis, called the *tilt axis*, which is orthogonal to the optical axis of the TEM. The rotation angle is called the *tilt angle*, and its angular range is usually contained in a subset of $[-60^\circ, 60^\circ]$.

2.2. Difficulties.

Limitations in instrumentation combined with the unfortunate combination of very noisy data and the severe ill-posedness of the inverse problem have been (and still are) responsible for the slow dissemination of ET as a reliable structure determination technique in life sciences. The former issue is partly addressed by the rapid technological development, so we focus on the latter, which is due to the following reasons.

The dose problem. This is *the* single most important problem in ET. It limits the total number of images that can be taken and arises due to specimen damage during electron exposure. A typical range of dose that can be tolerated by a biological specimen is about 2000–7000 $\mathrm{e}^-/\mathrm{nm}^2$, which translates into about a total of 500–1250 e^-/pixel (at 25000× magnification with pixel size of 14 μm) distributed over 60 or 120 micrographs, so each micrograph is very noisy, and the Poisson randomness of the data (shot noise) has to be accounted for.

Limited range of the tilt angle. Restrictions in the data acquisition geometry for ET, especially the restriction on the range of the tilt angle in single-axis tilting, lead to limited angle data and therefore imply that the conditions for stable reconstruction are not fulfilled.

Region of interest problem. For a given positioning of the specimen, only a subregion of it is subject to electron exposure. Thus, the region of interest then equals (or is a subset of) the intersection of all the exposed parts of the specimen from different positions. Since we have contribution from outside the region of interest,[1] we are dealing with the region of interest problem (local tomography), somewhat similar to the well-known "long object problem" in three-dimensional CT.

2.3. The inverse problem in ET. We therefore confine ourselves to presenting a very brief outline for how one arrives at the expression for the forward operator that occurs in the standard phase contrast model used by the ET community. The interested reader is referred to [4, 10, 23] for a more detailed exposition.

The starting point is to assume that we have perfect coherent imaging; i.e., the incoming electron wave is a monochromatic plane wave (coherent illumination), and electrons scatter only elastically. The scattering properties of the specimen are in this case given by the electrostatic potential, and the electron-specimen interaction is modeled by the scalar Schrödinger equation. The picture is completed by adding a description of the effects of the optics and the detector of the TEM, both modeled as convolution operators. However, inelastic scattering and incoherent illumination introduce partial incoherence, so the basic assumption of perfect coherent imaging must be relaxed. The incoherence that stems from inelastic scattering is accounted for within the coherent framework by introducing an imaginary part to the scattering potential, called the absorption potential. The incoherence that stems from incoherent illumination is modeled by modifying the convolution kernel that describes the effect of the optics. Next, as shown in [4, Theorem 9.5], taking the first order Born approximation and linearizing the intensity enables one to explicitly express the measured intensity in terms of the propagation operator (well known from diffraction tomography [18, p. 48]) acting on the scattering potential of the specimen convolved with point spread functions describing the optics and detector. The *standard phase contrast model* used by the ET community for the image formation in TEM is based on replacing the propagation operator by its high energy limit as the wave number tends to infinity. This yields a model for the image formation that is based on the parallel beam transform (see (2.3) for a definition).

The structure of the specimen is assumed to be fully described by the *scattering potential* $f\colon \mathbb{R}^3 \to \mathbb{C}$, which is defined as

$$f(\boldsymbol{x}) := -\frac{2m}{\hbar^2}\big(V(\boldsymbol{x}) + iV_{\mathrm{abs}}(\boldsymbol{x})\big), \tag{2.1}$$

[1] The exposed part of the specimen is larger than the region of interest.

where m denotes the electron mass at rest, $V\colon \mathbb{R}^3 \to \mathbb{R}^-$ is the potential energy[2] that models elastic interaction, and $V_{\text{abs}}\colon \mathbb{R}^3 \to \mathbb{R}^-$ is the absorption potential that models the decrease in the flux, due to inelastic scattering, of the nonscattered and elastically scattered electrons. Under the assumptions and approximations outlined in the previous paragraph, the expression for the *intensity generated by a single electron* is given as

$$(2.2)\quad \mathcal{I}(f)(\boldsymbol{z},\boldsymbol{\omega}) := \frac{1}{M^2}\Biggl(1-(2\pi)^{-2}\Biggl[\Bigl\{\mathrm{PSF}^{\mathrm{re}}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f^{\mathrm{re}})(\cdot,\boldsymbol{\omega})\Bigr\}\Bigl(\frac{\boldsymbol{z}}{M}\Bigr) + \Bigl\{\mathrm{PSF}^{\mathrm{im}}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f^{\mathrm{im}})(\cdot,\boldsymbol{\omega})\Bigr\}\Bigl(\frac{\boldsymbol{z}}{M}\Bigr)\Biggr]k^{-1}\Biggr)$$

for a unit vector $\boldsymbol{\omega} \in S^2$ and $\boldsymbol{z} \in \boldsymbol{\omega}^\perp$, where $\boldsymbol{\omega}^\perp := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}\cdot\boldsymbol{\omega} = 0\}$. In the above expression, $f^{\mathrm{re}}, f^{\mathrm{im}}\colon \mathbb{R}^3 \to \mathbb{R}^+$ denotes the real and imaginary parts of f in (2.1) and $\mathcal{P}$ denotes the *parallel beam transform (X-ray transform)*, which is defined as the operator taking the line integral of a function, i.e.,

$$(2.3)\qquad \mathcal{P}(f)(\boldsymbol{y},\boldsymbol{\omega}) := \int_{t=-\infty}^{\infty} f(\boldsymbol{y}+t\boldsymbol{\omega})\,dt \quad \text{for } \boldsymbol{\omega} \in S^2 \text{ and } \boldsymbol{y} \in \boldsymbol{\omega}^\perp.$$

Moreover, $\circledast_{\boldsymbol{\omega}^\perp}$ denotes the two-dimensional convolution in the $\boldsymbol{\omega}^\perp$-plane, and the point spread functions $\mathrm{PSF}^{\mathrm{re}}$ and $\mathrm{PSF}^{\mathrm{im}}$ in (2.2) model the effect of the optics and incoherent illumination of the TEM. A precise expression for these can be found, e.g., in [4, section 9.1], [10, Chapter 65], or [23, section 3.3]. Finally, k is the particle wave number[3] w.r.t. the homogeneous background medium (which in our case is a vacuum) and M denotes the magnification.

As already mentioned, (2.2) yields the expression for the intensity generated by a single electron. The expression for the actual data measured on a micrograph needs to account for the detector point spread function (usually a slow-scan CCD camera) as well as the stochasticity in the data. Following [4, section 6.3], the stochasticity in the data is captured by assuming that the actual data delivered by the detector from a pixel should be modeled as a *sample of a random variable*, which in turn implies that the inverse problem in ET must be defined in a probabilistic setting.

Definition 2.1. *We have a fixed finite set S_0 of directions on a smooth curve $S \subset S^2$ that defines our parallel beam data collection geometry. The scattering properties of the specimen are assumed to be fully described by the complex valued scattering potential f defined in* (2.1). *For each direction $\boldsymbol{\omega} \in S_0$, the specimen is probed by a monochromatic wave, and the resulting data on the micrograph at pixel (i,j) is denoted by* $\mathrm{data}[f](\boldsymbol{\omega})_{i,j}$. *The* forward operator *in ET, denoted by $\mathcal{T}$, is defined as the expected value of* $\mathrm{data}[f](\boldsymbol{\omega})_{i,j}$, *i.e.,*

$$\mathcal{T}(f)(\boldsymbol{\omega})_{i,j} := \mathbf{E}\Bigl[\mathrm{data}[f](\boldsymbol{\omega})_{i,j}\Bigr] \quad \textit{for } \boldsymbol{\omega} \in S^2 \textit{ and pixel } (i,j).$$

The inverse problem *is to determine f when a sample of* $\mathrm{data}[f](\boldsymbol{\omega})_{i,j}$ *is known for $\boldsymbol{\omega} \in S_0$ and finitely many pixels (i,j).*

[2] The potential energy is related to the electrostatic potential $U\colon \mathbb{R}^3 \to \mathbb{R}^+$ by $V = -eU$, where e is the charge of the electron.

[3] We use the convention that the relation between the wave number k and the wavelength λ is given by $k = 2\pi/\lambda$.

The full expression for $\mathrm{data}[f](\boldsymbol{\omega})_{i,j}$ (and the corresponding forward operator $\mathcal{T}$) is given in [4, equation (29)]. We will settle for a simplified version, also given in [4, equation (30)], which yields the following expression for the forward operator:

$$\mathcal{T}(f)(\boldsymbol{\omega})_{i,j} = \mathrm{gain}_{i,j}\,|\triangle_{i,j}|\,\mathrm{Dose}(\boldsymbol{\omega})\Big\{\mathrm{PSF}_{\mathrm{det}} \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{I}(f)(\cdot,\boldsymbol{\omega})\Big\}(\boldsymbol{z}_{i,j}) + \epsilon_{i,j}. \tag{2.4}$$

In the above expression, $\mathrm{gain}_{i,j}$ is a detector constant, $|\triangle_{i,j}|$ is the area of the (i,j)th pixel, $\mathrm{Dose}(\boldsymbol{\omega})$ is the incoming dose which gives the number of electrons hitting the specimen per area unit, $\mathrm{PSF}_{\mathrm{det}}$ is the detector point spread function, and $\epsilon_{i,j}$ is the mean value of the stochastic variable representing the *additive noise introduced by the detector.*

2.4. Integral geometric approaches for solving the inverse problem. There are two main assumptions underlying all current integral geometric approaches for solving the inverse problem given in Definition 2.1. The first is to assume that *the forward operator yields the actual measured data*; i.e., the data in pixel (i,j) in the micrograph with tilt $\boldsymbol{\omega}$ equals the expected value of the random variable $\mathrm{data}[f](\boldsymbol{\omega})_{i,j}$. The second is to assume that *the forward operator is given by* (2.4). Next, we shall see that appropriate postprocessing of the measured data allows us to obtain an expression for the values of the parallel beam transform on $(\boldsymbol{z}_{i,j},\boldsymbol{\omega})$ with $\boldsymbol{\omega}\in S_0$ and $\boldsymbol{z}_{i,j}\in\Sigma\subset\boldsymbol{\omega}^\perp$, where Σ is a fixed finite set defined by the pixels in the detector. We have in this way recast the inverse problem in ET (given by Definition 2.1) as the problem of inverting the parallel beam transform.

2.4.1. Generate single electron intensity data. The first step is to generate *single electron intensity data* from the actual measured data. This can be done by deconvolving the effects of the detector point spread function $\mathrm{PSF}_{\mathrm{det}}$ and rescaling the measured data so that it corresponds to the intensity generated by a *single electron.* Let $\mathrm{I}(\boldsymbol{\omega})_{i,j}$ correspond to the intensity generated by a single electron at pixel (i,j). If the rescaling and deconvolution are appropriately[4] done, then we get

$$\mathrm{I}(\boldsymbol{\omega})_{i,j} \approx \mathcal{I}(f)(\boldsymbol{z}_{i,j},\boldsymbol{\omega}) \quad \text{for } \boldsymbol{\omega}\in S_0 \text{ and } \boldsymbol{z}_{i,j}\in\Sigma.$$

By (2.2), for $\boldsymbol{z}_{i,j}\in\Sigma$ we then get that

$$\begin{aligned}\Big\{\mathrm{PSF}^{\mathrm{re}}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f^{\mathrm{re}})(\cdot,\boldsymbol{\omega})\Big\}\Big(\frac{\boldsymbol{z}_{i,j}}{M}\Big)& \\ + \Big\{\mathrm{PSF}^{\mathrm{im}}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f^{\mathrm{im}})(\cdot,\boldsymbol{\omega})\Big\}\Big(\frac{\boldsymbol{z}_{i,j}}{M}\Big) &\approx (2\pi)^2 k\Big(1 - M^2\,\mathrm{I}(\boldsymbol{\omega})_{i,j}\Big).\end{aligned} \tag{2.5}$$

One can now proceed in a number of different ways in order to recast the inverse problem in Definition 2.1 as the problem of inverting the parallel beam transform.

2.4.2. Amplitude contrast only. The easiest approach is to assume that we have *perfect optics (no defocus and no spherical or chromatic aberration) and ignore all apertures.* These assumptions imply that $\mathrm{PSF}^{\mathrm{re}}\equiv 0$ and $\mathrm{PSF}^{\mathrm{im}} = \delta_{\boldsymbol{\omega}^\perp}$ (see, e.g., [4, section 9.3]), so (2.5) reduces to

$$\mathcal{P}(f^{\mathrm{im}})\Big(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\Big) \approx (2\pi)^2 k\Big(1 - M^2\,\mathrm{I}(\boldsymbol{\omega})_{i,j}\Big) \quad \text{for } \boldsymbol{z}_{i,j}\in\Sigma. \tag{2.6}$$

[4] The deconvolution of the detector point spread function $\mathrm{PSF}_{\mathrm{det}}$ needed to create $\mathrm{I}(\boldsymbol{\omega})$ is an ill-posed operation, and therefore it needs to be performed using a regularization scheme. However, the ill-posedness is not severe since the Fourier transform of $\mathrm{PSF}_{\mathrm{det}}$ is positive [4, Remark 6.3], so it should be fairly straightforward to perform this regularization as exemplified in [32].

The inverse problem in Definition 2.1 can now be reformulated as the problem of inverting the parallel beam transform of f^{im} when the data is given by the right-hand side of (2.6).

Note that the real part f^{re} of the scattering potential is absent in (2.6) and thus cannot be recovered. This is to be expected since the phase contrast is visible only due to the imperfections in the optics (with nonzero defocus and/or nonzero aberration). The inability to recover f^{re} is a serious deficiency with this approach since f^{re} is the part of the scattering potential that has a straightforward physical interpretation in terms of the molecular structure of the specimen, whereas f^{im} is a phenomenological construction that accounts for the decrease in the flux, due to inelastic scattering, of the nonscattered and elastically scattered electrons. Assuming only amplitude contrast therefore works well only with strongly scattering specimens where most of the contrast in the micrographs is from amplitude contrast.

2.4.3. Constant amplitude contrast ratio. This is the most common approach in ET. It is based on introducing an additional assumption, namely, that $f^{\mathrm{im}}(\boldsymbol{x}) = Qf^{\mathrm{re}}(\boldsymbol{x})$, where the constant Q is called the *amplitude contrast ratio.* Under this assumption (2.5) reduces to

$$\Big\{\mathrm{PSF}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \mathcal{P}(f^{\mathrm{re}})(\cdot,\boldsymbol{\omega})\Big\}\Big(\frac{\boldsymbol{z}_{i,j}}{M}\Big) \approx (2\pi)^2 k\Big(1 - M^2\,\mathrm{I}(\boldsymbol{\omega})_{i,j}\Big) \quad \text{for } \boldsymbol{z}_{i,j} \in \Sigma, \tag{2.7}$$

where

$$\mathrm{PSF}(\boldsymbol{z},\boldsymbol{\omega}) := \Big\{\mathrm{PSF}^{\mathrm{re}}(\cdot,\boldsymbol{\omega}) + Q\,\mathrm{PSF}^{\mathrm{im}}(\cdot,\boldsymbol{\omega})\Big\}(\boldsymbol{z}).$$

An expression for $\mathcal{P}(f^{\mathrm{re}})\big(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\big)$ can now be obtained by deconvolving the point spread function PSF in the expression (2.7).

There are several problems with this above approach. The first is that it requires a priori knowledge of Q. Second, deconvolving PSF is an ill-posed operation. This ill-posedness is especially pronounced since the Fourier transform of the kernel PSF has multiple zeroes (see, e.g., [4, section 9.1]). Thus, if one wants to use (2.7) in order to retrieve f^{re} (and f^{im} with knowledge of Q), then one needs to regularize the deconvolution operation involved in the right-hand side of (2.7). The most common approach is to again assume perfect optics and ignore all apertures. However, in such a case the criticism raised against the amplitude contrast model (2.6) also applies to this case, and not much is gained.

2.4.4. Phase contrast model with low-resolution amplitude contrast. We now propose a novel approach that does circumvent some of the difficulties raised above. The idea is to recover f^{re} by a hybrid approach. Begin by assuming *perfect optics (no defocus and no spherical or chromatic aberration) and ignore all apertures.* Under these assumptions we know that (2.6) is valid, which gives an expression for $\mathcal{P}(f^{\mathrm{im}})\big(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\big)$ with $\boldsymbol{z}_{i,j} \in \Sigma$. Inserting this expression into (2.5) yields an expression for

$$\Big\{\mathrm{PSF}^{\mathrm{re}}(\cdot,\boldsymbol{\omega}) \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \mathcal{P}(f^{\mathrm{re}})(\cdot,\boldsymbol{\omega})\Big\}\Big(\frac{\boldsymbol{z}_{i,j}}{M}\Big).$$

Finally, by deconvolving the point spread function $\mathrm{PSF}^{\mathrm{re}}$, we obtain the expression for $\mathcal{P}(f^{\mathrm{re}})\big(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\big)$ with $\boldsymbol{z}_{i,j} \in \Sigma$. This deconvolution operation is, however, ill-posed since the Fourier transform of the corresponding kernel $\mathrm{PSF}^{\mathrm{re}}$ has multiple zeroes. Hence, in order to use this approach, one needs to regularize this deconvolution operation.

2.4.5. Phase contrast model with higher order terms. The troublesome deconvolutions with the optics point spread functions can be avoided altogether if one makes an approximation based on the asymptotic expansion of the forward operator that includes higher order terms. More precisely, in [4, equation (40)] it is shown that

$$\begin{aligned}\mathcal{I}(f)(\boldsymbol{z},\boldsymbol{\omega}) = \frac{1}{M^2}\Biggl(1 &- (2\pi)^{-2}\mathcal{P}(f^{\mathrm{im}})\Bigl(\frac{\boldsymbol{z}}{M},\boldsymbol{\omega}\Bigr)k^{-1} \\ &+ (2\pi)^{-2}\biggl\{\Bigl(\frac{\triangle z}{2}+q\Bigr)\triangle_{\boldsymbol{\omega}^\perp}\Bigl[\mathcal{P}(f^{\mathrm{re}})(\cdot,\boldsymbol{\omega})\Bigr]\Bigl(\frac{\boldsymbol{z}}{M}\Bigr) \\ &\qquad + \triangle_{\boldsymbol{\omega}^\perp}\Bigl[\int_{\mathbb{R}} s f^{\mathrm{re}}(s\boldsymbol{\omega}+\cdot)\,ds\Bigr]\Bigl(\frac{\boldsymbol{z}}{M}\Bigr)\biggr\}k^{-2}\Biggr) + O(k^{-3}),\end{aligned}$$

where $\triangle_{\boldsymbol{\omega}^\perp}$ is the two-dimensional Laplacian in the $\boldsymbol{\omega}^\perp$-plane, $\triangle z$ is the defocus, and q is the shortest distance (considering all the tilts) between the specimen and the objective lens in the idealized optical system (the value of q is determined by the magnification M and focal length of the objective lens [4, section 8.5]). Now, note that

$$\int_{\mathbb{R}} s f^{\mathrm{re}}(s\boldsymbol{\omega}+\boldsymbol{z})\,ds \approx q\mathcal{P}(f^{\mathrm{re}})(\boldsymbol{z},\boldsymbol{\omega}),$$

which holds simply because q is much larger than the specimen thickness (where f^{re} has its support). Moreover, $\triangle_{\boldsymbol{\omega}^\perp}\bigl[\mathcal{P}(f^{\mathrm{re}})(\boldsymbol{z},\boldsymbol{\omega})\bigr] = \mathcal{P}(\triangle f^{\mathrm{re}})(\boldsymbol{z},\boldsymbol{\omega})$, where $\triangle$ is the Laplacian in $\mathbb{R}^3$, so we therefore end up with the following replacement of (2.5):

$$\begin{aligned}(2.8)\quad \mathcal{P}(f^{\mathrm{im}})\Bigl(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\Bigr) + \Bigl(\frac{\triangle z}{2}+2q\Bigr)k^{-1}\mathcal{P}(\triangle f^{\mathrm{re}})\Bigl(\frac{\boldsymbol{z}_{i,j}}{M},\boldsymbol{\omega}\Bigr)& \\ \approx (2\pi)^2 k\Bigl(1-M^2\,\mathrm{I}(\boldsymbol{\omega})_{i,j}\Bigr)&\end{aligned}$$

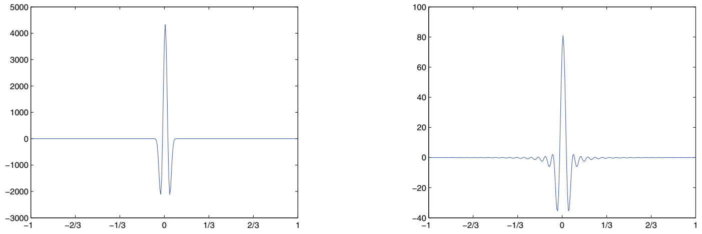
for $\boldsymbol{z}_{i,j}\in\Sigma$. One can now repeat the postprocessing approaches described in sections 2.4.3 and 2.4.4 but this time based on (2.8) instead of (2.5). This would yield postprocessing operations of data where one does not have to go through the ill-posed operation of deconvolving the optics.

2.4.6. Summary. As we have seen in the previous sections, performing a number of approximations allows us to recast the inverse problem in ET (given as in Definition 2.1) as the problem of solving (2.5) for f^{re} and f^{im}. This problem can then by additional assumptions be reduced to the problem of inverting the X-ray transform. Finally, bearing in mind the data collection scheme outlined in Definition 2.1, we are reduced to inverting the parallel beam transform since the line complex where the X-ray transform is sampled consists of lines parallel to a direction (which in turn varies on a curve in S^2).

3. Limited data local tomography.

To help the reader understand our three-dimensional local reconstruction methods, we will first outline planar Lambda tomography and then recall the parameterization of lines for the ET data set in $\mathbb{R}^3$.

Lambda tomography [6, 5, 28] is a very clever algorithm for parallel beam or fan beam tomography in the plane. It allows one to image a function $f(\boldsymbol{x})$ using only line integrals of f for lines near $\boldsymbol{x}$. It is a variant of the standard filtered backprojection inversion algorithm that replaces the standard filter (that has infinite support)



(a) Plot of the convolution kernel we use in place of $-D^2_{\boldsymbol{\sigma}}$ (see (5.8)). The kernel is local because it is zero off of the interval $[-0.1, 0.1]$.

(b) Plot of the standard filtered backprojection kernel with a comparable width.

FIG. 1. *Plot of kernels in Lambda tomography and in filtered backprojection. We see that the Lambda kernel illustrated in Figure* 1(a) *is local, whereas the standard filtered backprojection kernel shown in Figure* 1(b) *has infinite support.*

with a filter that takes a second derivative in the detector variable. Because the numerical derivative filter has small support, just near the line being evaluated, this reconstruction becomes local; see Figure 1.

The formula reads as follows:

$$\Lambda_\mu(f) = \frac{1}{4\pi}\mathcal{P}^*\big(\mu - \mathcal{D}^2_{\boldsymbol{\omega}^\perp}\big)\mathcal{P}(f). \tag{3.1}$$

In the above formula, $\mathcal{P}^*$ is the standard dual parallel beam transform integrating over all lines in the plane through the given point and $\mathcal{D}^2_{\boldsymbol{\omega}^\perp}$ is the second derivative in the $\boldsymbol{\omega}^\perp$ direction, i.e.,

$$\mathcal{D}^2_{\boldsymbol{\omega}^\perp}(g)(\boldsymbol{y}, \boldsymbol{\omega}) := \frac{d^2}{ds^2} g\big(\boldsymbol{y} + s\boldsymbol{\omega}^\perp, \boldsymbol{\omega}\big)\Big|_{s=0} \quad \text{for } \boldsymbol{y} \in \mathbb{R}^2$$

with $\boldsymbol{\omega} := (\sin\theta, \cos\theta)$ and $\boldsymbol{\omega}^\perp := (\cos\theta, -\sin\theta)$. In this section, $\boldsymbol{\omega}^\perp$ is a vector, and for the three-dimensional parallel beam transform, $\boldsymbol{\omega}^\perp$ is a plane. Note that $\boldsymbol{\omega}$ is the unit vector $\pi/2$ radians counterclockwise from $\boldsymbol{\omega}^\perp$.

We subtract $\mathcal{D}^2_{\boldsymbol{\omega}^\perp}$ in (3.1) so that the Fourier transform of the kernel of Λ_μ is positive. *The result is a reconstruction not of f but of a function $\Lambda_\mu(f)$ that has singularities at the same places as f but with the singularities accentuated.* As has been shown in numerous articles (e.g., [6, 16, 21]) and as we will try to show here, Lambda reconstructions can be as useful as reconstructions from filtered backprojection if one does not need actual density values or if one has only local data from which density values cannot be obtained. The constant $\mu \geq 0$ is included in (3.1), as suggested by Smith and coauthors [26, 6], to provide some contour to the reconstruction. That is, the backprojected second derivative

$$-\Delta\mathcal{P}^*\mathcal{P}(f) = \mathcal{P}^*\big(-\mathcal{D}^2_{\boldsymbol{\omega}^\perp}\mathcal{P}\big)(f) = \Lambda_0(f),$$

or "pure" Lambda, emphasizes density changes or boundaries. The μ factor provides "contours" from the smoothed version of the original function since it results in the convolution

$$\mu\mathcal{P}^*\mathcal{P}(f) = f * \frac{2\mu}{\|\cdot\|},$$

where $*$ denotes the convolution in $\mathbb{R}^3$. The sum, (3.1), provides a reconstruction including both contours of f and the boundaries. A more complete rationale and analysis are given in [6, 5].

We now recall the three-dimensional parallel beam complex for ET. Let $S \subset S^2$ be a smooth curve on the sphere, and for $\boldsymbol{\omega} \in S$ let $\boldsymbol{\omega}^{\perp}$ be the plane through the origin perpendicular[5] to $\boldsymbol{\omega}$. Then, following any of the approaches outlined in section 2.4, the inverse problem in ET as stated in Definition 2.1 can be recast as the problem of recovering f given values $\mathcal{P}(f)(\boldsymbol{y}, \boldsymbol{\omega})$, where $\mathcal{P}$ is given as in (2.3), $\boldsymbol{\omega} \in S$, and $\boldsymbol{y} \in \boldsymbol{\omega}^{\perp}$ (or for the local problem, $\boldsymbol{y}$ is in a proper subset of $\boldsymbol{\omega}^{\perp}$). The example below shows that this is an intrinsically ill-posed problem in the single-axis tilting case since the local transform is not injective even in the absence of noise. Thus, singularity detection algorithms such as Lambda tomography are natural methods since they regularize the problem by reconstructing only features that are stably visible (see, e.g., [20, 16, 21]).

Example 3.1. *Assume $f(x) = g(x_3)$, where the specimen is parallel to the (x_1, x_2)-plane. Then, only $\int_{-\infty}^{\infty} g(x_3)\, dx_3$ can be determined from single-axis tilt ET data with tilt angle less than $\pi/2$. This also is a counterexample for any set of lines, as in ET, without horizontal directions.*

4. The algorithm in general.

In this section, we describe our Lambda tomography algorithm for directions (or angles) on an arbitrary smooth curve $S \subset S^2$. Note that we follow an ET convention when we use the word "angle" to describe a point on S^2. This general setup will provide a general framework for the single-axis tilt geometry we use in ET, which will be described in section 5. The algorithm is general enough to take care of other tilting geometries such as dual-axis and conical tilting, which some of the newest electron microscopes can provide. A generalization of algorithm to slant-hole SPECT (with the same geometry as conical tilt) will be given in [22].

The planar Lambda tomography we described in section 3 has two important advantages: it solves the region of interest problem—it is local—and it is easily adaptable to other limited data sets in the plane [16, 21]. As noted in section 1, the inverse problem in ET (as given in Definition 2.1) can be rephrased as a three-dimensional limited angle region of interest reconstruction problem. It is therefore natural to consider a type of singularity detection algorithm related to Lambda tomography. Furthermore, as shown by Example 3.1, *inversion is not possible, so recovering singularities is an appropriate goal.* It also turns out that, despite the severe ill-posedness of the inverse problem, *those singularities that can be recovered are recovered stably* at least in range of Sobolev spaces.[6] Such an algorithm includes two pieces, a backprojection operator and a derivative along the lines.

Let S be a curve on the sphere. The backprojection operator is the *dual parallel beam transform* for directions on the curve S,

$$\mathcal{P}_S^*(g)(\boldsymbol{x}) := \int_{\boldsymbol{\omega} \in S} g\big(\boldsymbol{x} - (\boldsymbol{x} \cdot \boldsymbol{\omega})\boldsymbol{\omega}, \boldsymbol{\omega}\big)\, d\boldsymbol{\omega} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^3, \tag{4.1}$$

where the measure $d\boldsymbol{\omega}$ is the arc length measure on the curve S and the point $\boldsymbol{x} - (x \cdot \boldsymbol{\omega})\boldsymbol{\omega}$ is the projection of $\boldsymbol{x}$ onto the plane $\boldsymbol{\omega}^{\perp}$.

[5] Note that $\boldsymbol{\omega}^{\perp}$ was a direction in S^1 perpendicular to $\boldsymbol{\omega}$ in the planar (two-dimensional) setting, whereas in the three-dimensional setting it is a plane perpendicular to $\boldsymbol{\omega}$.

[6] A stronger type of stability would be a microlocal inverse continuity estimate, and the authors are not aware of such a direct estimate for these operators.

The derivative along lines is defined as follows. Assume the curve S is parameterized by the differentiable function $\boldsymbol{\omega}(\theta)$ with derivative $\boldsymbol{\omega}'(\theta) \neq \mathbf{0}$, and let

$$\boldsymbol{\sigma}(\theta) := \frac{\boldsymbol{\omega}'(\theta)}{\|\boldsymbol{\omega}'(\theta)\|} \tag{4.2}$$

be a unit tangent to the curve S at $\boldsymbol{\omega}(\theta)$. Then, we denote the second derivative in direction $\boldsymbol{\sigma}$ by

$$\mathcal{D}^2_{\boldsymbol{\sigma}} g\big(\boldsymbol{y}, \boldsymbol{\omega}(\theta)\big) := \frac{d^2}{ds^2} g\big(\boldsymbol{y} + s\boldsymbol{\sigma}(\theta), \boldsymbol{\omega}(\theta)\big)\bigg|_{s=0}. \tag{4.3}$$

Our basic reconstruction operator is

$$\mathcal{L}(f) := \mathcal{P}^*_S\big((\mu - \mathcal{D}^2_{\boldsymbol{\sigma}})\mathcal{P}(f)\big). \tag{4.4}$$

This is a natural generalization of the two-dimensional Lambda operator (3.1) since it includes a second derivative along lines, a smoothing term, and a backprojection. We include the factor of μ, as is done for standard Lambda tomography, to provide contour to the reconstruction.

How $\mathcal{L}$ detects singularities can be understood using microlocal analysis, as we do in section 6. We will show that $\mathcal{L}$ is a pseudodifferential operator (PDO) with a mildly singular symbol (Theorem A.1). Moreover, ET data are very noisy, as discussed in section 2.2, so to cope with that we smooth in two ways. First, we evaluate the derivative $\mathcal{D}^2_{\boldsymbol{\sigma}}$ using a kernel that is a smoothed version of the second derivative. Second, we smooth by averaging nearby slices; that is, we also convolve in the $\boldsymbol{\omega}^{\perp}$-plane in the direction perpendicular to $\boldsymbol{\sigma}$. We will describe this smoothing explicitly in the case of single-axis tilting in the next section.

5. Single-axis tilt ET.

In this section, we will describe our algorithm for single-axis tilt ET. In *single-axis tilt* ET, one restricts the directions to a single tilt axis. We use a coordinate system where the electrons come in along the z-axis when $\boldsymbol{\omega} = (0, 0, 1)$, and we assume the *tilt axis* is the x-axis. Let us now write (4.4) in these coordinates.

Expression for S. Because the specimen cannot be fully rotated, this means that the curve of directions, S, is an arc of a circle in the (y, z)-plane and there is a limited angular range of $\pm\theta_{\max}$, where $\theta_{\max} \approx \pi/3$ radians. One appropriate parameterization for the curve S in this setting is

$$\boldsymbol{\omega}(\theta) := (0, \sin\theta, \cos\theta), \quad \theta \in \,]-\theta_{\max}, \theta_{\max}[\,, \tag{5.1}$$

and by (4.2) we get

$$\boldsymbol{\sigma}(\theta) = (0, \cos\theta, -\sin\theta). \tag{5.2}$$

Expression for $\mathcal{P}$. Now, $\boldsymbol{e}_1 := (1, 0, 0)$ and $\boldsymbol{\sigma}(\theta)$ form an orthonormal basis of the plane $\boldsymbol{\omega}(\theta)^{\perp}$ and thereby provide orthonormal coordinates on $\boldsymbol{\omega}(\theta)^{\perp}$:

$$\boldsymbol{y} = (y_1, y_{\boldsymbol{\sigma}}) \mapsto y_1\boldsymbol{e}_1 + y_{\boldsymbol{\sigma}}\boldsymbol{\sigma}(\theta) \in \boldsymbol{\omega}(\theta)^{\perp}. \tag{5.3}$$

In these coordinates the set of lines is parameterized by

$$\begin{aligned} &Y := \Big\{(\boldsymbol{y}, \theta) \;\Big|\; \boldsymbol{y} = (y_1, y_{\boldsymbol{\sigma}}) \in \mathbb{R}^2,\, \theta \in \,]-\theta_{\max}, \theta_{\max}[\,\Big\}, \\ &(\boldsymbol{y}, \theta) \mapsto \ell(\boldsymbol{y}, \theta) := \Big\{y_1\boldsymbol{e}_1 + y_{\boldsymbol{\sigma}}\boldsymbol{\sigma}(\theta) + t\boldsymbol{\omega}(\theta) \;\Big|\; t \in \mathbb{R}\Big\}, \end{aligned} \tag{5.4}$$

and functions on lines will be written $g(\boldsymbol{y},\theta) = g\big((y_1, y_{\boldsymbol{\sigma}}),\theta\big)$, so in particular $\mathcal{P}$ has parameterization

$$\mathcal{P}(f)(\boldsymbol{y},\theta) := \mathcal{P}(f)\big(y_1\boldsymbol{e}_1 + y_{\boldsymbol{\sigma}}\boldsymbol{\sigma}(\theta), \boldsymbol{\omega}(\theta)\big). \tag{5.5}$$

Expression for $\mathcal{P}^*_S$. Before expressing the dual operator $\mathcal{P}^*_S$ in (4.1) in these coordinates, we smooth it slightly in order to make it a classical Fourier integral operator (FIO) and in order to increase the accuracy in the numerical integration. This is done by choosing $0.9\,\theta_{\max} < \theta_{\text{cut}} < \theta_{\max}$ and defining the smooth function

$$\varphi : [-\pi/2, \pi/2] \to [0,1], \quad \operatorname{supp}\varphi = [-\theta_{\text{cut}}, \theta_{\text{cut}}], \tag{5.6}$$

where φ is nonzero on $]{-\theta_{\text{cut}}}, \theta_{\text{cut}}[$ and equal to one on most of this interval. φ is then extended $\mathbb{R}$ by making it π-periodic. The *smoothed limited angle dual parallel beam transform*, which is the version of $\mathcal{P}^*_S$ that we will be using, is

$$\mathcal{P}^*_{\theta_{\text{cut}}}(g)(\boldsymbol{x}) := \int_{-\theta_{\text{cut}}}^{\theta_{\text{cut}}} g\Big(\boldsymbol{x} - \big(\boldsymbol{x}\cdot\boldsymbol{\omega}(\theta)\big)\boldsymbol{\omega}(\theta), \boldsymbol{\omega}(\theta)\Big)\varphi(\theta)\,d\theta \quad \text{for } \boldsymbol{x} \in \mathbb{R}^3. \tag{5.7}$$

A simple trapezoidal rule integration, which corresponds to a specific choice of φ, works well in (5.7).

Expression for $\mathcal{D}^2_{\boldsymbol{\sigma}}$. In our coordinates $(\boldsymbol{y},\theta)$, $\mathcal{D}^2_{\boldsymbol{\sigma}}$ becomes

$$\mathcal{D}^2_{\boldsymbol{\sigma}}(g)(\boldsymbol{y},\theta) := \frac{d^2}{ds^2} g\big((y_1, s),\theta\big)\bigg|_{s=0}. \tag{5.8}$$

Expression for (4.4). Our expression for the Lambda operator $\mathcal{L}$ in (4.4) becomes

$$\mathcal{L}(f) = \mathcal{P}^*_{\theta_{\text{cut}}}\big((\mu - \mathcal{D}^2_{\boldsymbol{\sigma}})\mathcal{P}(f)\big). \tag{5.9}$$

This operator is a two-dimensional limited angle Lambda operator in each fixed plane $\boldsymbol{x} = \text{constant}$ (compare with (3.1)).

Further smoothing. We actually use a smoothed version of the derivative $\mathcal{D}^2_{\boldsymbol{\sigma}}$ in the $\boldsymbol{\sigma}$ direction, and we also smooth between slices in the $\boldsymbol{e}_1$ direction. This can be understood as either a convolution/smoothing of the data in the data plane, $\boldsymbol{\omega}^\perp$, or as a convolution/smoothing of the final reconstruction, as we now explain. Let $\phi_1 \in \mathscr{C}^\infty_c(\mathbb{R})$ be even with $\int_{\mathbb{R}} \phi_1 = 1$ and $\phi_2 \in \mathscr{C}^\infty_c(\mathbb{R}^2)$ be radial with $\int_{\mathbb{R}^2} \phi_2 = 1$. Moreover, let $\widetilde{\phi}_2$ be the two-dimensional parallel beam transform of ϕ_2, and note that $\widetilde{\phi}_2$ is radial and independent of direction. Let $(\phi_1 \otimes \widetilde{\phi}_2)(x_1,x_2,x_3) = \phi_1(x_1)\widetilde{\phi}_2(x_2,x_3)$. Then, for $(\boldsymbol{y},\theta) \in Y$ and data f with compact support,

$$(\phi_1 \otimes \widetilde{\phi}_2) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f)(\boldsymbol{y},\theta) = \mathcal{P}\big((\phi_1 \otimes \phi_2) * f\big)(\boldsymbol{y},\theta), \tag{5.10}$$

where $*$ denotes the convolution in $\mathbb{R}^3$ and $\circledast_{\boldsymbol{\omega}^\perp}$ is the convolution in the detector plane, $\boldsymbol{\omega}(\theta)^\perp$. Equation (5.10) is valid since $\mathcal{P}$ integrates only over lines perpendicular to $\boldsymbol{e}_1$ and ϕ_1 is a function only of x_1. Because $\mathcal{L}$ is a convolution operator (see Theorem A.1), it commutes with the convolution with $\phi_1 \otimes \phi_2$. This means that

$$\big(\phi_1 \otimes \phi_2\big) * \mathcal{L}(f) = \mathcal{P}^*_{\theta_{\text{cut}}}\Big[(\mu - \mathcal{D}^2_{\boldsymbol{\sigma}})\big((\phi_1 \otimes \widetilde{\phi}_2) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(f)\big)\Big] \tag{5.11}$$

$$= \mathcal{P}^*_{\theta_{\text{cut}}}\Big[\phi_1 \underset{\boldsymbol{e}_1}{\circledast} \big((\mu\widetilde{\phi}_2 - \mathcal{D}^2_{\boldsymbol{\sigma}}\widetilde{\phi}_2) \underset{\boldsymbol{\sigma}}{\circledast} \mathcal{P}(f)\big)\Big], \tag{5.12}$$

where $\circledast_{\boldsymbol{e}_1}$ and $\circledast_{\boldsymbol{\sigma}}$ are one-dimensional convolutions in the $\boldsymbol{\omega}^{\perp}$-plane in the respective directions $\boldsymbol{e}_1$ and $\boldsymbol{\sigma}(\theta)$. So, our algorithm can be viewed as a smoothed version of $\mathcal{L}(f)$ (left-hand side of (5.11)), a smoothing of the data before applying $\mathcal{L}$ (right-hand side of (5.11)), or averaging over slices (the $\circledast_{\boldsymbol{e}_1}$ convolution in (5.12)) of a smoothed derivative (the $\circledast_{\boldsymbol{\sigma}}$ convolution in (5.12)).

6. Microlocal analysis applied to ET. We will now use microlocal analysis to analyze which singularities of a specimen are stably visible from single-axis tilt ET data.

Microlocal analysis allows one to rigorously define singularities of functions such as object boundaries. This is made precise by the wavefront set whose definition is our first task. Next, the theory of FIOs describes which singularities of a function are visible from its ET data. This correspondence follows from general theorems of Greenleaf and Uhlmann [8] about geodesic Radon transforms on admissible complexes, and the microlocal properties of this specific transform were examined by Boman and Quinto [2]. Here we will give a basic version of the microlocal regularity theorem which will allow us to characterize visible singularities. The complete version of the theorem will be presented in the appendix along with characterizations of $\mathcal{L}$ as a convolution PDO. In the appendix, we also introduce a generalization, $\mathcal{L}_{\triangle}$ (A.14), of an operator of Louis and Maaß. Our characterization will show the trade-offs between the operators; $\mathcal{L}_{\triangle}$ can add stronger singularities than $\mathcal{L}$. At the end of the section, we give an example that illustrates the predictions.

Before stating the formal definition of the wavefront set we need to deal with a technicality.

REMARK 6.1. *The wavefront set is typically defined as a subset of the cotangent bundle, because in this way it is invariant under diffeomorphisms. Furthermore, this is a natural way to describe wavefronts in general. Here is the identification for* $\mathbb{R}^3$. *For* $\boldsymbol{x} \in \mathbb{R}^3$, *the cotangent space* $T^*_{\boldsymbol{x}}(\mathbb{R}^n)$ *is the set of linear functionals on the tangent space* $T_{\boldsymbol{x}}(\mathbb{R}^3)$, *and* $\boldsymbol{dx}_j$ *is the dual covector to* $\frac{\partial}{\partial x_j}$ $(j = 1, 2, 3)$. *This gives a canonical representation,*

$$\boldsymbol{\xi} \ni \mathbb{R}^3 \to \boldsymbol{\xi}\boldsymbol{dx} := \xi_1\boldsymbol{dx}_1 + \xi_2\boldsymbol{dx}_2 + \xi_3\boldsymbol{dx}_3.$$

The cotangent bundle, $T^*(\mathbb{R}^3)$, *is the set* $T^*(\mathbb{R}^3) := \big\{(\boldsymbol{x}, \boldsymbol{\xi}\boldsymbol{dx}) \;\big|\; \boldsymbol{x} \in \mathbb{R}^3,\ \boldsymbol{\xi} \in \mathbb{R}^3\big\}$, *where* $(\boldsymbol{x}; \boldsymbol{\xi}\boldsymbol{dx}) = (x_1, x_2, x_3; \xi_1\boldsymbol{dx}_1 + \xi_2\boldsymbol{dx}_2 + \xi_3\boldsymbol{dx}_3)$.

Recall that $\mathscr{D}'(\mathbb{R}^3)$ is the set of all distributions, $\mathscr{S}'(\mathbb{R}^3)$ is the set of tempered distributions (dual space of $\mathscr{S}(\mathbb{R}^3)$), and $\mathscr{E}'(\mathbb{R}^3)$ is the space of compactly supported distributions. We are now ready to define the concept of a wavefront set.

DEFINITION 6.2 (see [19, p. 259]). *Let* f *be a distribution,* $\boldsymbol{x}_0 \in \mathbb{R}^n$, *and* $\boldsymbol{\xi}_0 \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. *We then define the following:*

1. f *is in* $\mathscr{C}^\infty$ microlocally near $(\boldsymbol{x}_0, \boldsymbol{\xi}_0\boldsymbol{dx})$ *if and only if there is a cut-off function* $\psi \in \mathscr{C}_c^\infty(\mathbb{R}^n)$ *with* $\psi(\boldsymbol{x}_0) \neq 0$ *and a function* u *homogeneous of degree zero that is smooth on* $\mathbb{R}^n \setminus \{\boldsymbol{0}\}$ *with* $u(\boldsymbol{\xi}_0) \neq 0$ *such that the product* $u(\cdot)\widehat{\psi f}(\cdot)$ *is rapidly decreasing at* ∞.[7] *The* $\mathscr{C}^\infty$ wavefront set of f, $\mathrm{WF}(f)$, *is the complement of the set of* $(\boldsymbol{x}_0; \boldsymbol{\xi}_0\boldsymbol{dx})$ *near which* f *is microlocally smooth.*
2. f *is in* $\mathscr{H}^\alpha$ microlocally near $(\boldsymbol{x}_0; \boldsymbol{\xi}_0\boldsymbol{dx})$ *if and only if there is a cut-off function* $\psi \in \mathscr{C}_c^\infty(\mathbb{R}^n)$ *with* $\psi(\boldsymbol{x}_0) \neq 0$ *and a function* u *homogeneous of*

[7] In our context, a function h is *rapidly decreasing at* ∞ if for each $k \in \mathbb{N}$ there is a $C > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^n$, $\big|h(\boldsymbol{x})\big| \leq C\big(1 + \|\boldsymbol{x}\|\big)^{-k}$. Sometimes one replaces the function u by an open cone U containing $\boldsymbol{\xi}_0$ on which $\widehat{\psi f}$ is rapidly decreasing at ∞.

degree zero and smooth on $\mathbb{R}^n \setminus \{\mathbf{0}\}$ *and with* $u(\boldsymbol{\xi}_0) \neq 0$ *such that the product* $u(\cdot)\widehat{\psi f}(\cdot) \in \mathscr{L}^2\big(\mathbb{R}^n, \big(1+|\boldsymbol{\xi}|^2\big)^\alpha\big)$*. The* $\mathscr{H}^\alpha$ wavefront set of f, $\mathrm{WF}^\alpha(f)$*, is the complement of the set of* $(\boldsymbol{x}_0, \boldsymbol{\xi}_0 \boldsymbol{dx})$ *near which* f *is microlocally in* $\mathscr{H}^\alpha$*.*

For example, if f is one inside the unit disk and zero outside, then $\mathrm{WF}(f) = \mathrm{WF}^1(f)$ and they both consist of the covectors conormal to the boundary of the disk. In general, if $f : \mathbb{R}^3 \to \mathbb{R}$ is $\mathscr{C}^\infty$ except for jump singularities along smooth surfaces, then the $\mathscr{C}^\infty$ wavefront set of f consists of all the conormals to these surfaces of discontinuity.

Having defined the necessary concept of a wavefront set, we now turn our attention to our main theorem, which characterizes the singularities that are visible from single-axis tilt ET data.

Theorem 6.3 (microlocal regularity theorem). *Let* $f \in \mathscr{E}'(\mathbb{R}^3)$*,* $(y_1, y_{\boldsymbol{\sigma}}, \theta_0) = \boldsymbol{y} \in Y$*, and let* $\boldsymbol{\xi}_0 \in \boldsymbol{\omega}(\theta)^\perp$ *be a nonzero vector where we write* $\boldsymbol{\xi}_0 = \xi_1 \boldsymbol{e}_1 + \xi_{\boldsymbol{\sigma}} \boldsymbol{\sigma}(\theta)$*. Finally, let* $\boldsymbol{x}_0 \in \ell(y_1, y_{\boldsymbol{\sigma}}, \theta)$*. If* $\xi_{\boldsymbol{\sigma}} \neq 0$*, then there is a corresponding covector in* $T^*_{\boldsymbol{y}}(Y)$ *such that* $(\boldsymbol{x}_0, \boldsymbol{\xi}_0 \boldsymbol{dx}) \in \mathrm{WF}(f)$ *if and only if this covector is in* $\mathrm{WF}\big(\mathcal{P}(f)\big)$ *(this correspondence is given in Theorem* A.6*). If we also assume that* $\mathcal{P}(f)$ *is* $\mathscr{C}^\infty$ *near* $\boldsymbol{y}$*, then* $(\boldsymbol{x}_0, \boldsymbol{\xi}_0 \boldsymbol{dx}) \notin \mathrm{WF}(f)$*.*

Note that $\boldsymbol{dx}_1$ is conormal to $\boldsymbol{\omega}(\theta)$ (and thus conormal to the line $\ell(\boldsymbol{y}, \theta)$ for all θ since $\boldsymbol{\omega}(\theta)$ is in the (y, z)-plane). So, the restriction $\xi_{\boldsymbol{\sigma}} \neq 0$ just means $\boldsymbol{\xi}_0 \boldsymbol{dx}$ defined in the theorem above is not a multiple of $\boldsymbol{dx}_1$ ($\boldsymbol{\xi}_0$ is not parallel to $\boldsymbol{e}_1$).

In general, Radon transforms (such as the parallel beam transform in this article) detect only singularities perpendicular to the sets of integration, so it is not surprising that the theorem provides information only about singularities of f conormal to $\boldsymbol{\omega}(\theta)$ since these are conormal to the corresponding lines in the data set. However, for this transform, there are two conormal directions that are excluded, $\boldsymbol{\xi}_0 = \pm\boldsymbol{e}_1$; these are *"bad" cotangent directions* because they "should" be visible (they are conormal to lines in the data set), but they cause problems. We will examine these problems in the appendix, and in particular we will show that $\mathcal{L}$ can add singularities in these directions.

Example 6.4. *We now illustrate the implications of Theorem* 6.3 *for ET. Let* D *be the unit disk in space, and let* f *be one inside* D *and zero outside. Assume the region of interest contains* D*, and assume that* φ *satisfies* (5.6)*. If* $\boldsymbol{x} \in \operatorname{bd} D$*, then* $\boldsymbol{x}$ *is normal to* $\operatorname{bd} D$ *at* $\boldsymbol{x}$*, so* $(\boldsymbol{x}; \boldsymbol{x}\boldsymbol{dx}) \in \mathrm{WF}(f)$*. No matter what* θ_{cut} *is, the wavefront* $\pm\boldsymbol{e}_1\boldsymbol{dx}$ *is problematic. This is a conormal at the points* $(\pm 1, 0, 0)$ *on the boundary. Singularities in other conormal directions are visible from the data as long as the direction is perpendicular to a line in the data set. Let* $\boldsymbol{x} = (x_1, x_2, x_3) \in \operatorname{bd} D$*. Because of the geometry of the single-axis tilt* (5.1)*, this means that* $|x_3/x_2| < \tan(\theta_{\mathrm{cut}})$ *in order for the wavefront at* $\boldsymbol{x}$ *to be visible. The part of the sphere that should be visible is illustrated in Figure* 2*. One would expect for numerical reasons that the boundary would get gradually less well defined near the edge of the visible part.*

7. Applications to real data.

We have tested the limited angle Lambda algorithm based on the Lambda operator (5.9) on both in vitro and in situ ET data. The Lambda reconstruction is obtained by applying the limited angle Lambda algorithm directly on the region of interest. This reconstruction is compared to a filtered back-projection (FBP) reconstruction that has been regularized by an additional low-pass filtering (low-pass FBP). This latter filtering, which in our case reduces the resolution to 10 nm, is necessary in order to gain stability, and the value for the low-pass filtering represents the best trade-off between stability and resolution if FBP is to be used on

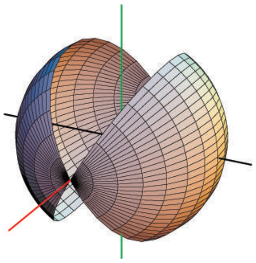


FIG. 2. *Part of sphere with normal vectors normal to lines in the data set with $\theta_{\text{cut}} = 60°$. The x_1-axis is facing out of the page.*

these particular examples. The low-pass FBP is applied to the entire reconstruction region, and the region of interest is then extracted for comparison against the limited angle Lambda reconstruction.

The first case, shown in Figure 3, is the reconstruction of in vitro monoclonal immunoglobulin G (IgG) molecules with a molecular weight of 150 kDa. The ET data was collected from single-axis tilting (see section 2.1) with a uniform sampling of the tilt angle in $[-60°, 60°]$ at 1° step. The pixel size is 0.5241 nm and the total dose is 1820 e^-/nm^2. A detailed account on the background for the study, the experimental setting, and the study objective is given in [24]. The reconstruction region is $256 \times 256 \times 256$ pixels in size, and the local region of interest is centered in the midpoint of the reconstruction region with a size of $128 \times 128 \times 128$ pixels.

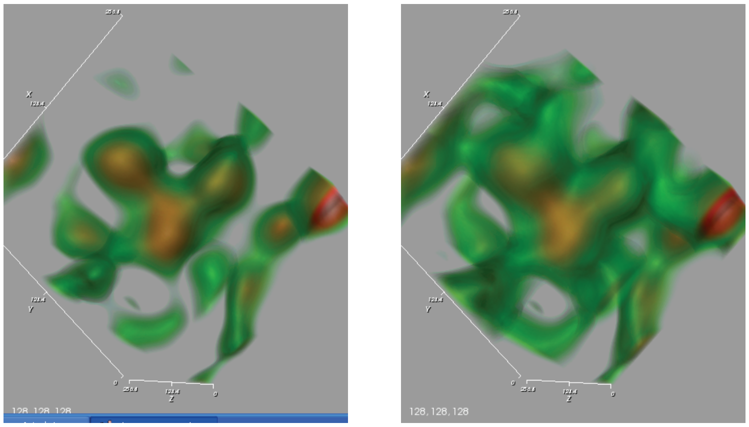
Figure 3 shows how the limited angle Lambda reconstruction emphasizes boundaries better. It also seems to somewhat suppress the background noise outside the molecule, and the IgG molecule (which is in the center) is more visible than in the low-pass FBP reconstruction.

The next case, shown in Figure 4, is the reconstruction of an in situ tissue sample (could be a human, rat, or mouse kidney). The ET data was collected from single-axis tilting (see section 2.1) with a uniform sampling of the tilt angle in $[-60°, 60°]$ at 2° step. The pixel size is 0.5241 nm and the total dose is 1520 e^-/nm^2. A detailed account on the background for the study, the experimental setting, and the study objective is given in [30, 27]. The reconstruction region is $300 \times 300 \times 150$ pixels in size, and the local region of interest is centered in the midpoint of the reconstruction region and is of $200 \times 200 \times 140$ pixels size.

Since the object in Figure 4 is in situ, parts of the object outside the region of interest will affect the FBP reconstruction in the region of interest but not the Lambda reconstruction (since it does not require data from outside the region of interest). The reconstructions in Figure 4 clearly show that the limited angle Lambda reconstruction defines boundaries better since the "V" shaped region containing the slit diaphragm (in the upper right side of the object) is more clearly defined than in the low-pass FBP reconstruction. This also illustrates the microlocal principles of section 6 since the slabs are tangent to lines in the data set.

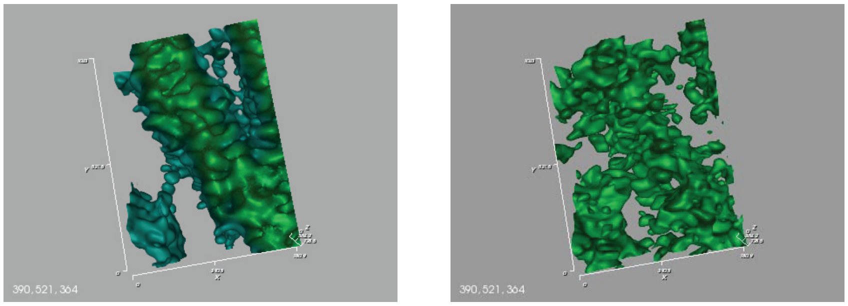
Appendix. The microlocal properties of $\mathcal{P}$ and $\mathcal{L}$. In this section, we will describe the microlocal properties of our transform $\mathcal{P}$ and the reconstruction operator $\mathcal{L}$ (5.9). We will use this information to explain how the transform detects singularities and show the relevance to ET. The properties of the more general operator (4.4) are similar, and the details will be given in a subsequent article [22].

The convolution operator in $\mathbb{R}^n$ is denoted by $*$. For the Fourier transform on



(a) Lambda reconstruction. (b) Low-pass FBP reconstruction.

FIG. 3. *The boundaries are better defined in the Lambda reconstruction when compared to the low-pass (*10 *nm-resolution) FBP reconstruction. The background noise is also suppressed. This makes the analysis of the IgG molecule easier.*



(a) Lambda reconstruction. (b) Low-pass FBP reconstruction.

FIG. 4. *The "V" shaped region containing the slit diaphragm is more clearly defined in the Lambda reconstruction than the low-pass (*10 *nm-resolution) FBP reconstruction.*

$\mathbb{R}^n$, we use the normalization

$$\mathcal{F}(f)(\boldsymbol{\xi}) = \widehat{f}(\boldsymbol{\xi}) := \int_{\boldsymbol{x}\in\mathbb{R}^n} e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}} f(\boldsymbol{x})\, d\boldsymbol{x}.$$

The two-dimensional Fourier transform on the plane $\boldsymbol{\omega}^\perp$ is defined in a similar way, and in the coordinates we chose for single axis tilt (see section 5), it is

$$\mathcal{F}_{\boldsymbol{\omega}^\perp}(g)(\boldsymbol{\eta},\theta) := \int_{(y_1,y_{\boldsymbol{\sigma}})\in\mathbb{R}^2} e^{-i(y_1,y_{\boldsymbol{\sigma}})\cdot(\eta_1,\eta_{\boldsymbol{\sigma}})} g(\boldsymbol{y},\theta)\, d\boldsymbol{y} \quad \text{for } \boldsymbol{\eta}\in\mathbb{R}^2.$$

Our next theorem characterizes the reconstruction operator as a convolution PDO with a symbol that is singular all along the ξ_1-axis. This has specific implications for reconstructions based on $\mathcal{L}$, as we explain in Theorem A.2 and Example A.5.

THEOREM A.1. *Let $\mathcal{P}^*_{\theta_{\text{cut}}}$ be defined by* (5.7)*, where the smooth function φ satisfies the assumptions given in* (5.6)*, $\mathcal{P}$ is defined by* (5.5)*, and $\mu \geq 0$. For $\boldsymbol{\xi} = (\xi_1,\xi_2,\xi_3) \in \mathbb{R}^3$, let $\boldsymbol{\xi}' = (\xi_2,\xi_3)$, and for $\boldsymbol{\xi}' \neq \mathbf{0}$, let $\arg(\boldsymbol{\xi}')$ be one of the* *angles in the plane from the ξ_2-axis to $\boldsymbol{\xi}'$.*[8] *Then, for $\mu \neq 0$ the operator $\mathcal{L}$ is defined for distributions of compact support. For $\mu = 0$, $\mathcal{L}$ is a continuous map from $\mathscr{H}^{\alpha+1}(\mathbb{R}^3)$ to $\mathscr{H}^{\alpha}(\mathbb{R}^3)$ for all $\alpha \in \mathbb{R}$. Moreover, in the coordinates defined above, $\mathcal{L}(f) = f * k$, and its symbol is the Fourier transform of k,*

$$\sigma(\boldsymbol{x}, \boldsymbol{\xi}) = \widehat{k}(\boldsymbol{\xi}) := \left(\frac{\varphi\big(\arg(\boldsymbol{\xi}') + \pi/2\big)}{(2\pi)^2 \|\boldsymbol{\xi}'\|} \right) \big(\mu + \|\boldsymbol{\xi}'\|^2\big). \tag{A.1}$$

Proof. To prove Theorem A.1, we need to show that

$$\mathcal{L}(f)(\boldsymbol{x}) = \frac{1}{(2\pi)^2} \int_{\boldsymbol{\xi} \in \mathbb{R}^3} e^{i\boldsymbol{x}\boldsymbol{\xi}}\, \widehat{k}(\boldsymbol{\xi}) \mathcal{F}(f)(\boldsymbol{\xi})\, d\boldsymbol{\xi}, \quad \text{where } \widehat{k} \text{ is given by (A.1).}$$

Initially, we assume f is a smooth function of compact support, but by continuity in distribution space, the end results will be true for distributions of compact support, as we will explain when needed. We use the convention that if $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, then $\boldsymbol{x}' = (x_2, x_3)$, and we begin the calculations in the plane $x_1 =$ constant. It is straightforward to show using polar coordinates in this plane that

$$\mathcal{P}^*_{\theta_{\text{cut}}} \mathcal{P}(f)(\boldsymbol{x}) = \int_{\boldsymbol{y}' \in \mathbb{R}^2} \frac{\varphi\big(\arg(\boldsymbol{y}')\big)}{\|\boldsymbol{y}'\|} f\big(\boldsymbol{x} + (0, \boldsymbol{y}')\big)\, dy'. \tag{A.2}$$

To write (A.2) as a PDO, we first fix x_1 and take the Fourier transform of (A.2) in $\boldsymbol{x}'$. Then, we use the fact about Fourier transforms of homogeneous functions [25, sect. 4, equation (7), p. 61] that the Fourier transform of $\varphi\big(\arg(\boldsymbol{y}')\big)/\|\boldsymbol{y}'\|$ is given by the first expression in parentheses in (A.1). To finish the proof, we take the inverse Fourier transform in x' and then the Fourier transform and inverse transform in x_1. This shows that

$$\mathcal{P}^*_{\theta_{\text{cut}}} \mathcal{P}(f)(\boldsymbol{x}) = \frac{1}{(2\pi)^2} \int_{\boldsymbol{\xi} \in \mathbb{R}^3} e^{i\boldsymbol{x}\boldsymbol{\xi}} \left(\frac{\varphi\big(\arg(\boldsymbol{\xi}') + \pi/2\big)}{\|\boldsymbol{\xi}'\|} \right) \mathcal{F}(f)(\boldsymbol{\xi})\, d\boldsymbol{\xi}. \tag{A.3}$$

Note that $\mathcal{P}^*_{\theta_{\text{cut}}} : \mathscr{E}'(Y) \to \mathscr{D}'(\mathbb{R}^3)$ is continuous, and a cutoff[9] applied to $\mathcal{P}$ is continuous from $\mathscr{E}'(\mathbb{R}^3)$ to $\mathscr{E}'(Y)$ by duality. These observations explain why $\mathcal{P}^*_{\theta_{\text{cut}}}\mathcal{P}$ and $\mathcal{L}$ are defined and continuous from $\mathscr{E}'(\mathbb{R}^3)$ to $\mathscr{D}'(\mathbb{R}^3)$.

To write $\mathcal{L}$ as a PDO, we first observe that, by an integration by parts, $D^2_\sigma \mathcal{P}(f) = \mathcal{P}\big(\Delta_{\boldsymbol{x}'} f\big)$, where $\Delta_{\boldsymbol{x}'} = \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}$. This is clearly true for functions and true on $\mathscr{E}'$ by continuity. Then, we note that

$$\mathcal{L}(f) = \mathcal{P}^*_{\theta_{\text{cut}}} \mathcal{P}\big((\mu - \Delta_{\boldsymbol{x}'}) f\big), \tag{A.4}$$

so using (A.3) on $(\mu - \Delta_{\boldsymbol{x}'}) f$ gives (A.1).

Finally, let $\mathcal{L}^0$ be defined as $\mathcal{L}$ with $\mu = 0$. The Sobolev continuity of $\mathcal{L}^0$ follows immediately from the calculation of its symbol above since the symbol of $\mathcal{L}^0$ is bounded above by $(1 + \|\xi\|^2)^{1/2}$ since $\|\xi'\| \leq \|\xi\|$ and $|\varphi|$ is bounded above by 1. This proves the Sobolev continuity of $\mathcal{L}^0$. Note that $\mathcal{L}$ is not defined on $\mathscr{H}^\alpha$ because of the singularity of $1/\|\boldsymbol{\xi}'\|$ at $\xi' = 0$. This concludes the proof of Theorem A.1. □

[8] Note that $\varphi\big(\arg(\boldsymbol{\xi}')\big)$ is well defined since φ is π-periodic.

[9] Let ψ be a smooth function that is one on $[-\theta_{\text{cut}}, \theta_{\text{cut}}]$ and supported in $(-\theta_{\max}, \theta_{\max})$; then $\psi\mathcal{P} : \mathscr{E}'(\mathbb{R}^3) \to \mathscr{E}'(Y)$ is continuous, and $\mathcal{P}^*_{\theta_{\text{cut}}} \psi \mathcal{P} = \mathcal{P}^*_{\theta_{\text{cut}}} \mathcal{P}$.

We describe what $\mathcal{L}$ does to wavefront sets in our next theorem and in Example A.5.

THEOREM A.2. *Let $\mathcal{L}$ be as in Theorem* A.1*, and f is a distribution of compact support in the unit disk D. Finally, define*

$$\mathcal{V} := \left\{ \left(\boldsymbol{x}, \boldsymbol{\xi dx}\right) \in T^*(\mathbb{R}^3) \setminus 0 \;\middle|\; \xi_2 \neq 0,\, |\xi_3/\xi_2| < \tan\theta_{\mathrm{cut}} \right\}, \tag{A.5}$$

$$\mathcal{A} := \left\{ (\boldsymbol{x}, \xi_1 \boldsymbol{dx}_1) \;\middle|\; x_1 \in [-1,1],\, \boldsymbol{x}' \in \mathbb{R}^2,\, \xi_1 \in \mathbb{R} \setminus 0 \right\}. \tag{A.6}$$

Then,

$$\mathrm{WF}^{\alpha}\big(\mathcal{L}(f)\big) \cap \mathcal{V} = \mathrm{WF}^{\alpha+1}(f) \cap \mathcal{V}, \tag{A.7}$$

$$\mathrm{WF}^{\alpha}\big(\mathcal{L}(f)\big) \subset \big(\mathrm{WF}^{\alpha+1}(f) \cap \mathrm{cl}(\mathcal{V})\big) \cup \mathcal{A}. \tag{A.8}$$

The set $\mathcal{V}$ in (A.5) is the set of "reliably visible" singularities. Equation (A.7) implies that singularities of f in those codirections are visible in the reconstruction $\mathcal{L}(f)$, and they are one order less smooth in Sobolev scale in the reconstruction than the corresponding singularities of f. Recall that visible covectors have to be conormal to lines in the data set by Theorem 6.3, and directions in $\mathcal{V}$ are all such covectors except for the "bad" cotangent directions, those in the $\pm\boldsymbol{dx}_1$ codirection.

Inclusion (A.8) and Example A.5 demonstrate that $\mathcal{L}$ can give additional singularities in the set $\mathcal{A}$ (in the $\pm\boldsymbol{dx}_1$ codirection). Therefore they do not affect singularities in the visible directions, namely those in $\mathcal{V}$. In Remark A.4, we prove that these added singularities are really a smearing of singularities of f in planes conormal the bad codirections, that is, planes $x_1 = a$.

Proof. In the proof of Theorem A.2 we will follow the conventions in [13, Chapter 8] and allow wavefront directions to be in $\mathbb{R}^n \setminus 0$ rather than in the cotangent space. Let us now give an outline of the proof. If $\mathcal{L}$ were a standard PDO, then the proof would follow from standard results. Our case is complicated by the fact that $\mathcal{L}$ is not a standard PDO since its symbol,

$$\sigma(\boldsymbol{x}, \boldsymbol{\xi}) = \left(\frac{\varphi\big(\arg(\boldsymbol{\xi}') + \pi/2\big)}{\|\boldsymbol{\xi}'\|} \right) \big(\mu + \|\boldsymbol{\xi}'\|^2\big),$$

does not satisfy the decay conditions on the derivatives in the ξ_1 direction. We introduce an operator $\mathcal{M}$ (A.10) that cuts off in the ξ_1 direction and show that $\mathcal{M}$ and $\mathcal{L}$ can be composed to become a standard PDO that detects singularities of f essentially in $\mathcal{V}$. Next, we show that $(1 - \mathcal{M})\mathcal{L}$ contributes to the wavefront set only near the ξ_1 direction. Finally, we put this together to show that the wavefront in directions in $\mathcal{V}$ are visible and that the only added directions are in $\mathcal{A}$. Theorem 8.2.9 and other results in [13, section 8.2] can be used to prove parts of this theorem without introducing the operator $\mathcal{M}$. We include that operator in order to provide an elementary proof of the other properties of $\mathcal{L}$. We begin with a useful lemma.

LEMMA A.3. *Let $f \in \mathscr{S}'(\mathbb{R}^n)$, and let $U \subset \mathbb{R}^n$ be a nonempty open cone containing the vector $\boldsymbol{\xi}_0$. Assume the Fourier transform $\mathcal{F}(f)$ is zero on U except for a compact set (which could be empty). Then, for all $\boldsymbol{x}_0 \in \mathbb{R}^n$, $(\boldsymbol{x}_0, \boldsymbol{\xi}_0) \notin \mathrm{WF}(f)$, where $\mathrm{WF}(f)$ denotes the $\mathscr{C}^\infty$ wavefront set of f.*

The proof follows from [13]. In particular, $\boldsymbol{\xi}_0$ is not in the limit cone at infinity of $\operatorname{supp} \mathcal{F}f$, and so, by [13, Lemma 8.1.7, p. 258], for any point $\boldsymbol{x}_0$, $(\boldsymbol{x}_0, \boldsymbol{\xi}_0) \notin \mathrm{WF}(f)$.

We now define the operator $\mathcal{M}$ such that $\mathcal{ML}$ is a standard PDO that detects most singularities in $\mathcal{V}$. Denote the set of second coordinates in $\mathcal{V}$ by

$$\mathcal{W} := \left\{ \boldsymbol{\xi} \in \mathbb{R}^3 \setminus 0 \;\middle|\; \xi_2 \neq 0,\, |\xi_3/\xi_2| < \tan\theta_{\mathrm{cut}} \right\}. \tag{A.9}$$

Let U be a small conic open neighborhood of $\pm\boldsymbol{e}_1$, and let U' be a conic open subset of U such that $\pm\boldsymbol{e}_1 \in U'$ and $\mathrm{cl}(U') \subset \bigl(U \cup \{0\}\bigr)$. Now let $m(\boldsymbol{\xi})$ be a function that is homogeneous of degree zero in $\mathbb{R}^3$, smooth away from the origin, zero in U', and equal to one off of U. Define

$$\mathcal{M}(f)(\boldsymbol{x}) = \mathcal{F}^{-1}\bigl(m(\cdot)\mathcal{F}(f)(\cdot)\bigr)(\boldsymbol{x}). \tag{A.10}$$

Then, $\mathcal{M}$ is a classical PDO of order zero in Sobolev scale.

It is a straightforward justification using Fourier transforms that one can compose $\mathcal{L}$ with $\mathcal{M}$ (or $(1-\mathcal{M})$) for distributions of compact support, and we will assume this. $\mathcal{ML}$ is a classical PDO because its symbol,

$$m(\boldsymbol{\xi})\left(\frac{\varphi\bigl(\arg(\boldsymbol{\xi}') + \pi/2\bigr)}{\|\boldsymbol{\xi}'\|}\right)\bigl(\mu + \|\boldsymbol{\xi}'\|^2\bigr),$$

is the sum of a term homogeneous of degree (-1) and one homogeneous of degree 1. Since the terms are smooth away from the origin ($m(\xi)$ cuts off the near the nonsmooth ξ_1 direction), $\mathcal{ML}$ is a classical PDO of order one. Since its symbol is elliptic on the open set $\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U')\bigr)$, $\mathcal{L}$ is elliptic on that set. Furthermore, by local Sobolev continuity, the $\mathscr{H}^{\alpha+1}$ wavefront of f in $\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U)\bigr)$ corresponds to the $\mathscr{H}^{\alpha}$ wavefront of $\mathcal{ML}(f)$ on that set,

$$\mathrm{WF}^{\alpha}\bigl(\mathcal{ML}(f)\bigr) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U)\bigr)\Bigr) = \mathrm{WF}^{\alpha+1}(f) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U)\bigr)\Bigr). \tag{A.11}$$

Because $\mathrm{supp}\bigl(1 - m(\boldsymbol{\xi})\bigr) \subset \mathrm{cl}(U)$, $\mathcal{F}\bigl((1-\mathcal{M})\mathcal{L}(f)\bigr)$ has support contained in $\mathrm{cl}(U)$. So, by Lemma A.3,

$$\mathrm{WF}\bigl((1-\mathcal{M})\mathcal{L}(f)\bigr) \subset \mathbb{R}^3 \times \mathrm{cl}(U). \tag{A.12}$$

In addition, since $\mathcal{L} = \mathcal{ML} + (1-\mathcal{M})\mathcal{L}$, the $\mathscr{H}^{\alpha+1}$ wavefront set of $\mathcal{L}(f)$ off of $\mathbb{R}^3 \times \mathrm{cl}(U)$ is the same as that of $\mathcal{ML}(f)$. Using (A.11), we see that

$$\mathrm{WF}^{\alpha}\bigl(\mathcal{L}(f)\bigr) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U)\bigr)\Bigr) = \mathrm{WF}^{\alpha+1}(f) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathcal{W} \setminus \mathrm{cl}(U)\bigr)\Bigr).$$

By making U arbitrarily close to $\boldsymbol{e}_1$, we establish (A.7).

To prove the containment (A.8), we fix U, U', and $\mathcal{M}$ as above. We will now prove

$$\mathrm{WF}^{\alpha}\bigl(\mathcal{ML}(f)\bigr) \subset \left(\mathrm{WF}^{\alpha+1}(f) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathrm{cl}(\mathcal{W}) \setminus \mathrm{cl}(U)\bigr)\Bigr)\right) \cup \bigl(\mathbb{R}^3 \times \mathrm{cl}(U)\bigr). \tag{A.13}$$

First, because the symbol of $\mathcal{ML}$ is supported on the closed set $\mathbb{R}^3 \times \bigl(\mathrm{cl}(\mathcal{W}) \setminus U'\bigr)$,

$$\mathrm{WF}^{\alpha}\bigl(\mathcal{ML}(f)\bigr) \subset \mathbb{R}^3 \times \bigl(\mathrm{cl}(\mathcal{W}) \setminus U'\bigr).$$

Because of (A.11), we need only consider $\boldsymbol{\xi} \in \mathrm{bd}(\mathcal{W}) \setminus \mathrm{cl}(U)$ and $\boldsymbol{x}_0 \in \mathbb{R}^3$ such that $(\boldsymbol{x}_0, \boldsymbol{\xi}) \notin \mathrm{WF}^{\alpha+1}(f)$. Since $\mathcal{ML}$ is a standard PDO of order one, $(\boldsymbol{x}_0, \boldsymbol{\xi}) \notin \mathrm{WF}^{\alpha}\bigl(\mathcal{ML}(f)\bigr)$. This shows (A.13).

Next, the wavefront set of a sum is contained in the union of the wavefront set of the terms. Combining this fact with (A.12) and (A.13) yields

$$\mathrm{WF}\bigl(\mathcal{L}(f)\bigr) \subset \left(\mathrm{WF}^{\alpha+1}(f) \cap \Bigl(\mathbb{R}^3 \times \bigl(\mathrm{cl}(\mathcal{W}) \setminus \mathrm{cl}(U)\bigr)\Bigr)\right) \cup \bigl(\mathbb{R}^3 \times \mathrm{cl}(U)\bigr).$$

The inclusion in (A.8) now follows when we let U shrink to $\pm\boldsymbol{e}_1$.

We claim that if f is supported in B, then $\operatorname{supp}\mathcal{L}(f) \subset [-1,1]\times\mathbb{R}^2$. This is true by a global version of the argument at the end of Remark A.4. This concludes the proof of Theorem A.2. $\square$

Remark A.4. *Note that if $f \in \mathscr{E}'(\mathbb{R}^3)$ is smooth in the $\pm\boldsymbol{dx}_1$ codirection at all points, then, for sufficiently small U, $(1-\mathcal{M})(f)$ is smooth, and so*

$$\mathcal{L}(f) = \mathcal{M}\mathcal{L}(f) + \mathcal{L}(1-\mathcal{M})(f)$$

has no wavefront in the $\pm\boldsymbol{d\xi}_1$ codirection. In other words, for this f there are no added singularities.

A local version of this statement is true: if $a \in \mathbb{R}$ and f is smooth in the $\pm\boldsymbol{dx}_1$ codirection at all points in the plane $x_1 = a$, i.e., $\big((a,\boldsymbol{x}');\pm\boldsymbol{dx}_1\big) \notin \mathrm{WF}(f)$ for all $\boldsymbol{x}' \in \mathbb{R}^2$), then $\mathcal{L}(f)$ is smooth in the $\pm\boldsymbol{dx}_1$ codirection above all points on the plane $x_1 = a$. To see this we observe that because supp f *is compact and wavefront sets are conical and closed, one can find a function $g \in \mathscr{C}_c^\infty(\mathbb{R})$ is not zero near $x_1 = a$ and a sufficiently small neighborhood U of $\boldsymbol{e}_1$ such that $g(x_1)(1-\mathcal{M})(f)$ is smooth, and so $g\mathcal{L}(f)$ is smooth in the $\pm\boldsymbol{dx}_1$ codirection at all points. Thus, $\mathcal{L}(f)$ is smooth in this direction at all points in the plane $x_1 = a$.*

That is, wavefront is not added if f is smooth in this codirection at all points on the plane $x_1 = a$. However, Example A.5 *demonstrates that wavefront can be spread in the plane $x_1 = a$ if f has wavefront in the $\pm\boldsymbol{dx}_1$ direction at points in this plane.*

We now introduce a new operator, $\mathcal{L}_\triangle$, which is related to an operator of Louis and Maaß for cone beam CT. Louis and Maaß adapted Lambda tomography to cone beam tomography in a very clever way [17] by taking a Laplacian in the detector plane before taking cone-beam backprojection. This adds extra singularities to the reconstruction as proven in general in [8] and for the cone beam transform in $\mathbb{R}^3$ in [7, 14]. The natural generalization of the Louis–Maaß operator to our setting is

$$\mathcal{L}_\triangle(f) := \mathcal{P}^*_{\theta_{\mathrm{cut}}}\big((\mu - \Delta_{\boldsymbol{\omega}^\perp})\mathcal{P}(f)\big), \tag{A.14}$$

where $\Delta_{\boldsymbol{\omega}^\perp}$ is the Laplacian operator in the detector plane and $\mu \geq 0$. In Example A.5 we will show that $\mathcal{L}_\triangle$ adds stronger singularities than $\mathcal{L}$.

Anastasio et al. [1], Katsevich [15], and Ye, Yu, and Wang [31] have developed refinements of Louis and Maaß's operator for cone beam CT. They decrease the added singularities by taking a derivative in only one direction rather than taking the Laplacian in the detector plane. This is analogous to our operator $\mathcal{L}$, in which the derivative is D_σ^2. Although these results are related, they do not apply to parallel beam data, as our methods do.

The arguments in our proof of Theorem A.1 can be used to show that $\mathcal{L}_\triangle$ is a PDO with a singular symbol

$$\left(\frac{\varphi(\arg(\boldsymbol{\xi}') + \pi/2)}{\|\boldsymbol{\xi}'\|}\right)\big(\mu + \|\boldsymbol{\xi}\|^2\big)$$

and (A.7) and (A.8) hold for $\mathcal{L}_\triangle$. For fixed $\boldsymbol{\xi}'$, the symbol of $\mathcal{L}_\triangle$ is of order 2 as $\boldsymbol{\xi}_1 \to \infty$, although it is of order 1 in other directions. The symbol of $\mathcal{L}$ is more mildly singular since, although it is not differentiable when $\boldsymbol{\xi}' = \boldsymbol{0}$ (on the ξ_1-axis), it is of order zero as $\boldsymbol{\xi}_1 \to \infty$ when $\boldsymbol{\xi}'$ is fixed.

Our next example justifies the addition of the set $\mathcal{A}$ in (A.8), a set on which wavefront can be added by $\mathcal{L}$ and $\mathcal{L}_\triangle$. The example also shows how $\mathcal{L}_\triangle$ adds stronger singularities than $\mathcal{L}$.

EXAMPLE A.5. *Let $\alpha \in \mathbb{R}$, and let $\epsilon > 0$ be arbitrary. We construct a function $f \in \mathscr{H}^{\alpha+1}(\mathbb{R}^3)$ supported in $S := [-1/2, 1/2]^3$ with the following properties:*

1. *$\mathcal{L}_\triangle(f) \notin \mathscr{H}^\alpha_{\mathrm{loc}}(\mathbb{R}^3)$, and $\mathcal{L}_\triangle(f)$ has $\mathscr{H}^\alpha$ wavefront set in the $\pm\boldsymbol{dx}_1$ direction even though f is in $\mathscr{H}^{\alpha+1}$ everywhere. This is true even for points outside* $\operatorname{supp} f$, *points at which f is smooth.*
2. *$\mathcal{L}(f)$ is in $\mathscr{H}^{\alpha+1}_{\mathrm{loc}}(\mathbb{R}^3)$ but has $\mathscr{H}^{\alpha+1+\epsilon}$ wavefront set in the $\pm\boldsymbol{dx}_1$ codirection outside of* $\operatorname{supp} f$. *Therefore, $\mathcal{L}$ also spreads singularities, but, for this case, the singularities are weaker. This weakening is suggested by the fact that $\mathcal{L}^0$, which is defined as $\mathcal{L}$ with $\mu = 0$, is continuous of order one in Sobolev scale.*

The actual construction of f goes as follows: Let $\epsilon' = \min\{\epsilon, 1/2\}$, and let $\phi_1 \in \mathscr{H}^{\alpha+1}(\mathbb{R})$ with $\operatorname{supp}\phi_1 = [-1/2, 1/2]$ *such that*

$$\mathrm{WF}^{\alpha+1+\epsilon'}(\phi_1) = \big\{(x_1, t\boldsymbol{dx}_1) \bigm| x_1 \in [-1/2, 1/2],\ t \neq 0\big\}. \tag{A.15}$$

Also, let ϕ_2 be a nonnegative smooth function in $\mathbb{R}^2$ with $\operatorname{supp}\phi_2 = [-1/2, 1/2]^2$.

For $x_1 \in \mathbb{R}$ and $\boldsymbol{x}' \in \mathbb{R}^2$ define $f(x_1, \boldsymbol{x}') = \phi_1(x_1)\phi_2(\boldsymbol{x}')$. For $g \in \mathscr{C}^\infty_c(\mathbb{R}^2)$ define

$$\mathcal{H}(g)(\boldsymbol{x}') := \int_{\boldsymbol{y}' \in \mathbb{R}^2} \frac{\varphi\big(\arg(\boldsymbol{y}')\big)}{\|\boldsymbol{y}'\|} g(\boldsymbol{x}' + \boldsymbol{y}')\, d\boldsymbol{y}'; \tag{A.16}$$

then $\mathcal{H}$ is really $\mathcal{P}^*_{\theta_{\mathrm{cut}}}\mathcal{P}$ restricted to a fixed plane (compare with (A.2)*). Since $\mathcal{H}$ is a classical PDO, $\mathcal{H}$ is continuous from domain $\mathscr{C}^\infty_c(S)$ to $\mathscr{C}^\infty(\mathbb{R}^2)$.*

It is straightforward to show that $\mathcal{L}_\triangle(f) = -\phi_1''\mathcal{H}(\phi_2) + \phi_1\mathcal{H}\big((\mu - \Delta_{\boldsymbol{x}'})\phi_2\big)$, where ϕ_1'' is the second derivative of ϕ_1. Since ϕ_1 is chosen to be in $\mathscr{H}^{\alpha+1}$ and not $\mathscr{H}^{\alpha+1+\epsilon'}$, the first term in the expression for $\mathcal{L}_\triangle(f)$ is not in $\mathscr{H}^\alpha_{\mathrm{loc}}$, although the other terms are. Thus, $\mathcal{L}_\triangle(f)$ is not in $\mathscr{H}^\alpha_{\mathrm{loc}}$. Because of (A.15),

$$\mathrm{WF}^\alpha\big(\mathcal{L}_\triangle(f)\big) = \big\{(x_1, \boldsymbol{x}', t\boldsymbol{dx}_1) \bigm| x_1 \in [-1/2, 1/2],\ \boldsymbol{x}' \in \operatorname{supp}\mathcal{H}(\phi_2),\ t \neq 0\big\}. \tag{A.17}$$

Furthermore, since φ_2 is nonnegative and not the zero function, $\mathcal{H}(\varphi_2)$ has unbounded support. Thus $\mathcal{L}_\triangle$ adds Sobolev wavefront both inside and outside $\operatorname{supp} f$ *even though $f \in \mathscr{H}^{\alpha+1}_c(\mathbb{R}^3)$.*

In a similar way, one shows that $\mathcal{L}(f) = \phi_1\mathcal{H}\big((\mu - \Delta_{\boldsymbol{x}'})\phi_2\big)$, and so $\mathcal{L}f$ is in $\mathscr{H}^{\alpha+1}_{\mathrm{loc}}$ (because ϕ_1 is in $\mathscr{H}^{\alpha+1}$ and the other term is smooth), but

$$\begin{aligned}&\mathrm{WF}^{\alpha+1+\epsilon'}\big(\mathcal{L}(f)\big)\\ &\qquad = \Big\{(x_1, \boldsymbol{x}', t\boldsymbol{dx}_1) \bigm| x_1 \in [-1/2, 1/2],\ \boldsymbol{x}' \in \operatorname{supp}\mathcal{H}\big((\mu - \Delta_{\boldsymbol{x}'})\phi_2\big),\ t \neq 0\Big\}.\end{aligned}$$

Note that $\operatorname{supp}\mathcal{H}\big((\mu - \Delta_{\boldsymbol{x}'})\phi_2\big)$ *must be unbounded,*[10] *so $\mathcal{L}$ spreads singularities of f, but they are weaker than those for $\mathcal{L}_\triangle(f)$.*

To state Theorem 6.3, we need a little more notation. Covectors in $T^*(Y)$ will be denoted by $\big((y_1, y_{\boldsymbol{\sigma}}, \theta); \nu_1\boldsymbol{dy}_1 + \nu_2\boldsymbol{dy}_{\boldsymbol{\sigma}} + \nu_3\boldsymbol{d\theta}\big)$, where $(\nu_1, \nu_{\boldsymbol{\sigma}}, \nu_\theta) \in \mathbb{R}^3$ and $\boldsymbol{dy}_1$ is the covector dual to the tangent vector $\partial/\partial y_1$, $\boldsymbol{dy}_{\boldsymbol{\sigma}}$ is dual to $\partial/\partial y_\sigma$, and $\boldsymbol{d\theta}$ is dual to $\partial/\partial\theta$. Using these conventions we can state the following theorem that gives the basic microlocal analysis of $\mathcal{P}$ with the limited data given in our ET problem.

THEOREM A.6. *Let f be a distribution of compact support on $\mathbb{R}^3$, $\theta_{\max} \in\,]0, \pi/2[$, and assume $\mathcal{P}(f)(\boldsymbol{y}, \theta)$ is given on an open set $U \subset Y$. Moreover, let $(y_1, y_{\boldsymbol{\sigma}}, \theta_0) \in U$,*

[10] The two-dimensional version of the proof of (A.3) shows that the two-dimensional Fourier transform $\mathcal{F}\mathcal{H}(\phi_2)$ is a product including $\varphi\big(\arg(\boldsymbol{\xi}') + \pi/2\big)$ and so is zero on an open set. If $\mathcal{H}(\phi_2)$ had compact support, then $\phi_2 = 0$ since $\mathcal{F}\mathcal{H}(\phi_2)$ would be real-analytic.

let $\boldsymbol{\xi}_0$ be a nonzero vector perpendicular to $\boldsymbol{\omega}(\theta_0)$ written as $\boldsymbol{\xi}_0 = \xi_1\boldsymbol{e}_1 + \xi_{\boldsymbol{\sigma}}\boldsymbol{\sigma}(\theta_0)$, and assume $\xi_{\boldsymbol{\sigma}} \neq 0$ (i.e., $\boldsymbol{\xi}_0$ is not parallel to $\boldsymbol{e}_1$). Finally, let $\boldsymbol{x}_0 \in \ell(y_1, y_{\boldsymbol{\sigma}}, \theta_0)$. Then, $(\boldsymbol{x}_0; \boldsymbol{\xi}_0\boldsymbol{dx}) \in \mathrm{WF}^{\alpha}(f)$ if and only if

$$\text{(A.18)} \qquad \Big((y_1, y_{\boldsymbol{\sigma}}, \theta_0); \xi_1\boldsymbol{dy}_1 + \xi_{\boldsymbol{\sigma}}\boldsymbol{dy}_a + \big(\xi_{\boldsymbol{\sigma}}\boldsymbol{x}\cdot\boldsymbol{\omega}(\theta_0)\big)\boldsymbol{d\theta}\Big) \in \mathrm{WF}^{\alpha+1/2}\big(\mathcal{P}(f)\big).$$

The proof follows from the fundamental results in [9] that show that Radon transforms are FIOs and also from the analysis of the general X-ray transform in [8] (see also [2]). The proof involves first calculating the canonical relation of $\mathcal{P}$, next noting that $\mathcal{P}$ is elliptic, and finally using the calculus of FIOs [12] to tell what $\mathcal{P}f$ does to the wavefront set. A proof of this result is given for more general curves of directions in S^2 given in [22].

Acknowledgments. The authors are indebted to Prof. Ulf Skoglund for invaluable help, support, and knowledge as well as for generous access to data sets. The authors thank Prof. Jan Boman of Stockholm University for extremely valuable comments that helped clarify the microlocal analysis in the article. We thank Dr. Hans Rullgård of Stockholm University for carefully reading the manuscript, and we thank the referees for their thorough reviews and helpful comments. The first author thanks Prof. Dr. Alfred Louis and Universität des Saarlandes and Prof. Dr. Andreas Rieder and Universität Karlsruhe for their hospitality while this article was being researched and written.

REFERENCES

[1] M. A. Anastasio, Y. Zou, E. Y. Sidky, and X. Pan, *Local cone-beam tomography image reconstruction on chords*, J. Opt. Soc. Amer. A, 24 (2007), pp. 1569–1579.

[2] J. Boman and E. T. Quinto, *Support theorems for real analytic Radon transforms on line complexes in* $\mathbb{R}^3$, Trans. Amer. Math. Soc., 335 (1993), pp. 877–890.

[3] D. J. De Rosier and A. Klug, *Reconstruction of three dimensional structures from electron micrographs*, Nature, 217 (1968), pp. 130–134.

[4] D. Fanelli and O. Öktem, *Electron tomography: A short overview with an emphasis on the absorption potential model for the forward problem*, Inverse Problems, 24 (2008), 013001.

[5] A. Faridani, D. V. Finch, E. L. Ritman, and K. T. Smith, *Local tomography* II, SIAM J. Appl. Math., 57 (1997), pp. 1095–1127.

[6] A. Faridani, E. L. Ritman, and K. T. Smith, *Local tomography*, SIAM J. Appl. Math., 52 (1992), pp. 459–484.

[7] D. V. Finch, I.-R. Lan, and G. Uhlmann, *Microlocal analysis of the restricted X-ray transform with sources on a curve*, in Inside Out: Inverse Problems and Applications, Math. Sci. Res. Inst. Publ. 47, G. Uhlmann, ed., Cambridge University Press, Cambridge, UK, 2003, pp. 193–218.

[8] A. Greenleaf and G. Uhlmann, *Non-local inversion formulas for the X-ray transform*, Duke Math. J., 58 (1989), pp. 205–240.

[9] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, AMS, Providence, RI, 1977.

[10] P. W. Hawkes and E. Kasper, *Principles of Electron Optics. Volume* 3. *Wave Optics*, Academic Press, London, 1994.

[11] W. Hoppe, R. Langer, G. Knesch, and C. Poppe, *Protein-kristallstrukturanalyse mit elektronenstrahlen*, Naturwissenschaften, 55 (1968), pp. 333–336.

[12] L. Hörmander, *Fourier integral operators,* I, Acta Math., 127 (1971), pp. 79–183.

[13] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, Vol. I, Springer-Verlag, New York, 1983.

[14] A. I. Katsevich, *Cone beam local tomography*, SIAM J. Appl. Math., 59 (1999), pp. 2224–2246.

[15] A. I. Katsevich, *Improved cone beam local tomography*, Inverse Problems, 22 (2006), pp. 627–643.

[16] P. Kuchment, K. Lancaster, and L. Mogilevskaya, *On local tomography*, Inverse Problems, 11 (1995), pp. 571–589.

[17] A. K. Louis and P. Maaß, *Contour reconstruction in 3-D X-ray CT*, IEEE Trans. Med. Imaging, 12 (1993), pp. 764–769.

[18] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM Monogr. Math. Model. Comput. 5, SIAM, Philadelphia, 2001.

[19] B. Petersen, *Introduction to the Fourier Transform and Pseudo-Differential Operators*, Pitman, Boston, 1983.

[20] E. T. Quinto, *Singularities of the X-ray transform and limited data tomography in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, SIAM J. Math. Anal., 24 (1993), pp. 1215–1225.

[21] E. T. Quinto, *Local algorithms in exterior tomography*, J. Comput. Appl. Math., 199 (2007), pp. 141–148.

[22] E. T. Quinto, T. Bakhos, and S. Chung, *Local tomography in 3-D SPECT*, in Proceedings of the Interdisciplinary Workshop on Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT), Pisa, Italy, 2007, to appear.

[23] L. Reimer, *Transmission Electron Microscopy*, 4th ed., Springer Ser. Opt. Sci. 36, Springer-Verlag, New York, 1997.

[24] S. Sandin, L.-G. Öfverstedt, A. C. Wikström, O. Wrange, and U. Skoglund, *Structure and flexibility of individual immunoglobulin G molecules in solution*, Structure, 12 (2004), pp. 409–415.

[25] V. I. Semyanistyi, *Homogeneous functions and some problems of integral geometry on spaces of constant curvature*, Soviet Math. Dokl., 2 (1961), pp. 59–62.

[26] K. T. Smith, *Inversion of the X-ray transform*, in Inverse Problems, SIAM-AMS Proc. 14, AMS, Providence, RI, 1984, pp. 41–52.

[27] K. Tryggvason, J. Patrakka, and J. Wartiovaara, *Hereditary proteinuria syndromes and mechanisms of proteinuria*, New England J. Medicine, 354 (2006), pp. 1387–1401.

[28] E. I. Vainberg, I. A. Kazak, and V. P. Kurozaev, *Reconstruction of the internal three-dimensional structure of objects based on real-time integral projections*, Soviet J. Nondestructive Testing, 17 (1981), pp. 415–423.

[29] B. K. Vainshtein, V. V. Barynin, and G. V. Gurskaya, *The hexagonal crystalline structure of catalase and its molecular structure*, Dokl. Akad. Nauk SSSR, 182 (1968), pp. 569–572 (in Russian); Soviet Phys. Dokl., 13 (1969), pp. 838–841 (in English).

[30] J. Wartiovaara, L.-G. Öfverstedt, J. Khoshnoodi, J. Zhang, E. Makela, S. Sandin, V. Ruotsalainen, R. H. Cheng, H. Jalanko, U. Skoglund, and K. Tryggvason, *Nephrin strands contribute to a porous slit diaphragm scaffold as revealed by electron tomography*, J. Clin. Invest., 114 (2004), pp. 1475–1483.

[31] Y. Ye, H. Yu, and G. Wang, *Cone beam pseudo-lambda tomography*, Inverse Problems, 23 (2007), pp. 203–215.

[32] J. M. Zuo, *Electron detection characteristics of a slow-scan CCD camera, imaging plates and film, and electron image restoration*, Microsc. Res. Tech., 49 (2000), pp. 245–268.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1304–1315

BOUNDARY CONDITIONS FOR THE MICROSCOPIC FENE MODELS*

CHUN LIU† AND HAILIANG LIU‡

Abstract. We consider the microscopic equation of finite extensible nonlinear elasticity (FENE) models for polymeric fluids under a steady flow field. It is shown that for the underlying Fokker–Planck type of equations, any preassigned distribution on the boundary will become redundant once the nondimensional number $Li := \frac{Hb}{k_B T} \geq 2$, where H is the elasticity constant, $\sqrt{b}$ is the maximum dumbbell extension, T is the temperature, and k_B is the usual Boltzmann constant. Moreover, if the probability density function is regular enough for its trace to be defined on the sphere $|m| = \sqrt{b}$, then the trace is necessarily zero when $Li > 2$. These results are consistent with our numerical simulations as well as some recent well-posedness results by preassuming a zero boundary distribution.

Key words. microscopic finite extensible nonlinear elasticity models, boundary condition, polymer fluids, Fokker–Planck equation, Fichera function

AMS subject classifications. 34F05, 35K65, 35K20, 65C30

DOI. 10.1137/060667700

1. Introduction.

The two-scale macro-micro models have been proven successful in describing the dynamics of many polymeric fluids. The systems usually consist of a macroscopic momentum equation (the force balance equation) and a microscopic evolution of the probability distribution functions (PDFs) through Fokker–Planck equations [4, 5]. The coupling of the micro-macro interaction is through the transport of the PDF in the microscopic equations and the induced elastic stress in the macroscopic equations. It is through this interaction, the competition between the kinetic energy and the (multiscale) elastic energies, that all different hydrodynamical and rheological properties of these materials arise [5, 14]. Let y be the macroscopic Eulerian (observer's) coordinate and m the microscopic molecule configurational variable. The distinguished representation of the variables represents the nature of the scale separation of these models. Let $u = u(y,t)$ be the macroscopic velocity field of the flow and $y(X,t)$ be the induced flow map (trajectory) with macroscopic material coordinate X. $f = f(t,m,y)$, $(m,y) \in \mathbb{R}^{2d}$ is the PDF of the molecule separation. The dependence of the macroscopic coordinate of f is attributed to the macroscopic anisotropy of the materials. Moreover, the models assume that the microscopic deformation is the same as the macroscopic deformation through the macroscopic covariant (or anticovariant) deformation of m. In the case of $m = F\tilde{m}$, with $F = \frac{\partial y}{\partial X}$ the (macroscopic) deformation tensor induced by the flow map $y(X,t)$, and $\tilde{m}$ the undeformed configuration, we have the following microscopic evolution equation [17]:

$$\partial_t f + u \cdot \nabla_y f + \nabla_m \cdot (\kappa m f) = \frac{2}{\gamma}\left[(\nabla_m \cdot (\nabla_m U f) + k_B T \Delta_m f)\right], \tag{1.1}$$

*Received by the editors August 18, 2006; accepted for publication (in revised form) January 10, 2008; published electronically April 11, 2008.
http://www.siam.org/journals/siap/68-5/66770.html

†Department of Mathematics, Pennsylvania State University, 218 McAllister Building, University Park, PA 16802 (liu@math.psu.edu). This author's research was partially supported by NSF grants DMS-0405850 and DMS-0509094.

‡Mathematics Department, Iowa State University, Carver Hall 434, Ames, IA 50011 (hliu@iastate.edu). This author's research was partially supported by NSF grant DMS-0505975 and the Ames Laboratory (DOE facility).

where $\kappa = \nabla_y u$ is the strain rate tensor; U denotes the spring potential; γ is the friction coefficient; T is the absolute temperature; and k_B is the Boltzmann constant. Equation (1.1) models both the convection and stretching of the polymers by the macroscopic flow and the microscopic convection diffusion evolution. The latter mechanism can be interpreted with its corresponding SDE [11, 16]:

$$dm = \left(-\frac{2}{\gamma}\nabla_m U\right) dt + \sqrt{\frac{4k_BT}{\gamma}}dW_t,$$

where W_t is a standard Brownian motion, which in turn gives the Fokker–Planck dynamic to the PDF. Much of the material properties can be attributed to different microscopic energies. The simplest spring potential is given by the Hookean law $U(m) = H|m|^2/2$, where H is the elasticity constant. This potential has the distinguished feature that the system is closed under second moment closure, which yields the usual Oldroyd models [3, 5, 17]. A more commonly used model is with the following finite extensible nonlinear elasticity (FENE) potential:

$$U(m) = -\frac{Hb}{2}\log\left(1 - \frac{|m|^2}{b}\right),$$

which takes into account a finite-extensibility constraint by assigning infinite energy when the molecule length approaches $\sqrt{b}$, the maximum dumbbell extension [3]. In this case, the convective spring force becomes

$$\nabla_m U = \frac{Hm}{1 - |m|^2/b}, \tag{1.2}$$

which also becomes infinity on the boundary of $B_{\sqrt{b}}$. Intuitively, this should mean that the Fokker–Planck equation (1.1) is defined only on the open ball $\Omega = \{m \in \mathbb{R}^d : |m|^2 < b\}$, where $d = 2, 3$, disregarding the boundary condition preimposed on the solution of (1.1). On the other hand, the diffusion due to the thermofluctuation does have infinite propagation speed. Also, the Brownian motion is unbounded in the L^∞ norm. The main complexity with the FENE potential lies mainly with the singularity of the equation at the boundary. In this paper, we will discuss boundary conditions for the Fokker–Planck equation alone, with the fluid velocity being steady and homogeneous. The velocity gradient will be treated as a constant matrix. Observe that since (1.1) experiences singularity on the sphere $|m| = \sqrt{b}$, the data may not necessarily be well defined. The main issues of our interest in this work are the following:

- Are the boundary conditions necessary or redundant?
- If the PDF solution is regular enough to have a trace on the boundary, what is its trace, regardless of whether or not the data is preimposed?

These issues for the underlying FENE models are fundamental and attracted much attention in the study of the well-posedness in certain weighted Sobolev spaces (see, e.g., [15]) as well as in the two-dimensional SDE framework [13]. However, the whole issue remained open. Our key observation in this paper is that the answer to the above questions hinges on whether the nondimensional quantity

$$Li := \frac{Hb}{k_BT} \tag{1.3}$$

crosses a critical value 2. The main goals of this paper are to show two new results: (i) for the underlying Fokker–Planck equation (1.1), any boundary condition will become redundant once the nondimensional number $Li \geq 2$; (ii) if the PDF f is regular enough for its trace to be defined on the sphere $|m| = \sqrt{b}$, then the trace is necessarily zero when $Li > 2$. Physically speaking, no boundary condition is necessary for the case with the given spring at low temperature or the case at fixed temperature with the large product of H and b. To put our work in a proper perspective we recall that well-posedness of the coupled micro-macro system or Fokker–Planck equation alone has attracted much attention recently [2, 6, 7, 12]. In particular, the local existence results of [23] in the weighted Sobolev space will also force the zero boundary condition. Up to now, most works have been with prescribed boundary conditions. On the other hand, numerical simulations seemed to indicate that the solutions are not sensible to such boundary conditions. We also refer the reader to [15] for the study of large-time behavior of the coupled micro-macro system and a rigorous formulation of the no-flux boundary condition. We now proceed to identify the key quantity Li defined in (1.3) by making the following scaling:

$$(y, u, t, m, b) \to \left(\frac{y}{L_0}, \frac{u}{U_0}, \frac{t}{T_c}, \frac{m}{l}, \frac{b}{l^2} \right),$$

where $T_c := L_0/U_0$ is the macroscopic convective time scale and $l := \sqrt{\frac{k_B T}{H}}$ serves as the mesoscopic length scale of the spring. We further introduce the nondimensional parameter $De = \frac{T_r}{T_c}$, where $T_r = \frac{\gamma}{4H}$ characterizes the mesoscopic relaxation time scale of the spring, and De is often called the Deborah number, which is a unique parameter in non-Newtonian fluids. Putting all of the above together and still using (y, u, t, m, b) for the scaled quantities, (1.1) thus reduces to

$$\partial_t f(t, m) + \nabla_m \cdot (\kappa m f) = \frac{1}{2\,De}(\nabla_m \cdot (\nabla_m U f) + \Delta_m f), \tag{1.4}$$

where $\kappa = \nabla_y u$ is the steady homogeneous velocity gradient. $\mathrm{Tr}(\kappa) = \nabla_y \cdot u = 0$ for the incompressible flow. Note that the corresponding square of radius for the nondimensional configuration variable m is exactly the key parameter $Li = b/l^2 = \frac{Hb}{k_B T}$ as given in (1.3) (though still denoted by b in what follows). The potential in nondimensional form thus reads

$$\nabla_m U = \frac{m}{1 - |m|^2/b}. \tag{1.5}$$

In order to simplify the notation, we simply take $De = 1$ and obtain

$$\partial_t f(t, m) + \nabla_m \cdot (\kappa m f) = \frac{1}{2}(\nabla_m \cdot (\nabla_m U f) + \Delta_m f). \tag{1.6}$$

For (1.6) with scaled potential (1.5) our first result is that $b = 2$ is a critical extension parameter for deciding whether boundary conditions are necessary for a well-defined problem (in the sense that the well-posedness results are expected to be obtained in standard Sobolev spaces).

Theorem 1.1. *Consider* (1.6) *in* $\{m \mid |m| < \sqrt{b}\}$ *for* $t > 0$ *subject to certain initial data. The extension parameter* $b = 2$ *is a critical value in the sense that the Dirichlet boundary condition leads to a well-defined problem, provided that* (i) *when* $b < 2$*, the distribution* f *on boundary* $|m|^2 = b$ *must be imposed; and* (ii) *when* $b \geq 2$*, any preassigned distribution on boundary* $|m|^2 = b$ *will become redundant.*

It is known that the parameter range of physical interest is $b > 2$ [3]. Therefore, in this regime for b, the original problem can be formulated as follows: find a distribution function $f(t, m)$ such that (1.6) holds for $t > 0$ and

$$f(0, m) = f_0(m), \quad |m| < \sqrt{b}, \tag{1.7}$$

where $f_0 \geq 0$ is a given bounded measurable function.

Here we would like to make several remarks:

(1) The statement in this theorem is justified based on the use of Fichera's criterion [8], which is sketched in the appendix.

(2) The well-posedness for the case $b < 2$ with imposed distribution on $|m| = \sqrt{b}$ follows from Theorem 4.2 in the appendix, with an extra constraint on the nontrivial shear rate k. The proof of the well-posedness for the case $b \geq 2$ is more delicate. In this work the redundancy of boundary conditions is stated in the sense of trace for weak solutions. We note that boundary conditions can also be discussed in terms of trajectories of the stochastic process; see, e.g., Stroock and Varadhan [20]. The second order linear equations with nonnegative characteristic form attracted much interest in the 1960's and 1970's [18, 21]. The existence results are customarily in Sobolev spaces with certain disjointness assumptions on the relevant boundary parts with the irrelevant parts and the non(negative) characteristic parts. Regularity of solutions with continuity up to the boundary can be further discussed [18]. As for the FENE models considered in this paper, we will present a description of the existence of the full system (coupled with the flow field) in a separate paper.

(3) The corresponding analogue of this statement in the SDE framework is known [13], where for the two-dimensional case the authors show that if $b \geq 2$, then the trajectories of the stochastic process representing the evolution of the end-to-end vector does not touch the boundary of radius $\sqrt{b}$, which means the polymer does not reach its maximal extensibility. Our second result determines the trace of the PDF on the sphere $|m|^2 = b$.

THEOREM 1.2. *Consider the initial value problem* (1.6)–(1.7) *in* $|m| < \sqrt{b}$. *Let* $f_0(m)$ *be a bounded measurable function with* $\mathrm{supp}(f_0(m)) \subset \{m, |m| \leq \sqrt{b^*}, b^* < b\}$. *Then for* $b > 2$ *the solution* $f(t, m)$ *of* (1.6) *remains bounded and satisfies*

$$|f| \leq |f_0| \left(\frac{b - |m|^2}{b - b^*} \right)^{b/2 - \alpha} e^{Kt},$$

where α *and* K *satisfy*

$$0 < \alpha < \frac{b}{2} - 1, \quad K > K^* := \frac{\beta^2}{16b\alpha(b - 2 - 2\alpha)} - \rho(b - 2\alpha)$$

with $\beta = \rho(b - \alpha)r^2 + 2\alpha(d + b - 2 - 2\alpha)$ *and* $\rho = \sqrt{\mathrm{Tr}(\kappa^\top \kappa)}$.

In comparison we mention that the solution to the SDE associated with (1.6) is shown to exist and has trajectorial uniqueness if and only if $b \geq 2$ [12]. We also refer the reader to [2] for an existence result with prescribed zero boundary data. Our results also show that for $b \geq 2$ well-posedness requires no prescribed boundary value on $|m|^2 = b$, and for $b > 2$ the distribution function, if regular enough to have a trace, must have zero trace:

$$f(t, m)|_{|m| = \sqrt{b}} = 0. \tag{1.8}$$

In other words, one is not allowed to prescribe boundary data on the sphere $|m| = \sqrt{b}$ other than (unnecessary) $f = 0$ or a natural no-flux boundary condition. The difficulty of the problem lies in the singularity of the equation occurring at the boundary. The key to our approach is to rewrite the equation into a second order equation having standard nonnegative characteristic form, for which we apply the Fichera function criterion to check when boundary conditions are unnecessary [8, 18, 21]. We further investigate the trace of the PDF on the sphere $|m| = \sqrt{b}$ where no data is preimposed. Our approach is to convert the equation by a delicate transformation in such a way that the resulting equation supports a maximum principle. This paper is organized as follows: in section 2, we use the Fichera function criterion to prove Theorem 1.1. Section 3 is devoted to the trace analysis of the PDF on the sphere $|m| = \sqrt{b}$. The presentation is split into two parts, without and with homogeneous flow involved.

2. Critical parameter $b = 2$ and boundary conditions. In this section we shall show that $b = 2$ is a critical value in the sense that for $b < 2$ a boundary condition is necessary and when $b \geq 2$ the boundary distribution becomes redundant. Note that (1.6) has a singular lower order term on boundary $|m| = \sqrt{b}$; our approach is to first transform this equation into a second equation degenerating near the boundary. We then employ the method of the Fichera function [1, 8, 18, 21] to study the corresponding relevant boundary value points on the boundary [18], as sketched in the appendix. We now introduce the following transformation:

$$f(t, m) = g(t, m) \exp(-U(m)), \tag{2.1}$$

which gives

$$\partial_t g + \nabla_m \cdot (\kappa m g) - \nabla_m U \cdot (\kappa m g) = \frac{1}{2}[\Delta_m g - \nabla_m U \cdot \nabla_m g]. \tag{2.2}$$

The right-hand side of the equation becomes the dual form of the original Fokker–Planck equation [9, 19]. We note a different transformation in [6, 10, 22], $f(t, m) = g(t, m) \exp(-U(m)/2)$, which was used to remove the singularity at the boundary in the resulting equation. Applying further rescaling,

$$x = \sqrt{2}m, \quad r^2 = 2b, \quad v(t, x) = g(t, m), \tag{2.3}$$

we obtain

$$\partial_t v + \nabla_x \cdot (\kappa x v) + a(x) \cdot (\nabla_x v - \kappa x v) = \Delta_x v, \tag{2.4}$$

where

$$a(x) := \frac{bx}{r^2 - |x|^2}.$$

Note that $\nabla_x \cdot (\kappa x) = \mathrm{Tr}(\kappa) = 0$; the above equation reduces to

$$\partial_t v + (a(x) + \kappa x) \cdot \nabla_x v - a(x) \cdot \kappa x v = \Delta_x v. \tag{2.5}$$

Once v is determined, the PDF f can be recovered through

$$f(t, m) = v(t, \sqrt{2}m)(1 - |m|^2/b)^{b/2}. \tag{2.6}$$

Rewrite (2.5) as

$$L(v) = 0, \quad x \in \mathbb{R}^d, \tag{2.7}$$

where

$$
\begin{aligned}
L(v) := (r^2 - |x|^2)\Delta_x v - (r^2 - |x|^2)v_t \\
- (bx + (r^2 - |x|^2)\kappa x) \cdot \nabla_x v + bx^\top \kappa x v
\end{aligned}
$$

has a standard form:

$$L(v) :\equiv a^{kj}(\xi)D_{kj}v + b^k(\xi)D_k v + c(\xi)v = 0, \quad k, j = 0 \cdots d.$$

Here the repeated indices are summed from 1 to d, $\xi = (t, x)$:

$$a^{00} = 0, \quad b^0 = -(r^2 - |x|^2), \quad c(\xi) = bx^\top \kappa x$$

and

$$a^{kk}(\xi) = (r^2 - |x|^2), \quad b^k = -[bx_k + (r^2 - |x|^2)\kappa_{kj}x_j], \quad k = 1 \cdots d.$$

Note that the new equation is degenerate at boundary $|x| = r$. This second order equation has nonnegative characteristic form in domain $\Omega = \{(t, x),\, 0 < t < T^*,\, |x| < r\}$ for any $T^* > 0$, since

$$a^{kj}(\xi)y_k y_j \geq 0$$

for any real vector y and any point $\xi \in \Omega$. Hence there are no negative characteristic points on the boundary.

Next we check the sign of Fichera's function

$$\mathfrak{F} = (b^k - a^{kj}_{\xi_j})n_k$$

at points on $\partial\Omega$. At boundary $|x| = r$, $0 < t < T^*$, one has $n_0 = 0$, $n_k = -x_k/r$, $k = 1 \cdots d$; thus

$$
\begin{aligned}
\mathfrak{F}(t, x) &= \sum_{k=1}^{d} (b^k - a^{kk}_{x_k})n_k \\
&= \sum_{k=1}^{d} (-bx_k + 2x_k) \cdot (-x_k/r) \\
&= (b - 2)r.
\end{aligned}
$$

If $\mathfrak{F} \geq 0$, that is, $b \geq 2$, all boundary points are irrelevant and no boundary condition is needed. Otherwise, in the case $b < 2$, all boundary points are relevant and an appropriate boundary condition has to be imposed. We now examine the boundary $t = T^*$ and $|x| < r$, on which one has $n^0 = -1$, $n^k = 0$; thus

$$
\begin{aligned}
\mathfrak{F}(t, x) &= (b^k - a^{kj}_{\xi_j})n_k = b^0 n_0 \\
&= r^2 - |x|^2 > 0.
\end{aligned}
$$

No condition needs to be imposed at $t = T^*$ either. Similarly at $t = 0$, $|x| < r$, one obtains $\mathfrak{F}(t, x) = |x|^2 - r^2 < 0$; thereby a condition at $t = 0$, the initial condition, has to be imposed. This completes the proof of Theorem 1.1.

Remark. For the transformed equation $L[v] = 0$ with $\kappa = 0$, we have

$$\frac{1}{2}D_i b^i - \frac{1}{2}D_{ij}a^{ij} - c = d\left(1 - \frac{b}{2}\right) > 0$$

for $b < 2$. Hence the existence theorem (Theorem 4.2) in the appendix applies only to the case $b < 2$.

3. The trace of the distribution function on the sphere $|m| = \sqrt{b}$ for $b > 2$. We now restrict ourselves to the case of $b > 2$. The proof in the above section shows that the presence of the fluid velocity does not affect the relevancy of the boundary points with respect to the equation. In this section, we will show, if the solution exists and assumes a trace on the boundary, that the trace of the resulting PDF has to be zero.

3.1. No-flow case. We will start from the case $\kappa = 0$. Equation (2.4) becomes

$$\partial_t v + a(x) \cdot \nabla_x v = \Delta_x v, \quad a(x) := \frac{bx}{r^2 - |x|^2}, \tag{3.1}$$

with the initial condition

$$v(0, x) = v_0(x) = f_0\left(\frac{x}{\sqrt{2}}\right)\left(1 - \frac{|x|^2}{r^2}\right)^{-b/2}, \quad \operatorname{supp}(v_0) \subset [-r, r];$$

we are going to show that there exists an α satisfying $0 < \alpha < b/2$ and a $K > 0$ such that

$$|v(t, x)| \leq Me^{Kt}(r^2 - x^2)^{-\alpha} \quad \forall t > 0.$$

Combining with the original transformation (2.1) and (2.3), which yield

$$f(t, m) = v(t, x)(1 - |x|^2/r^2)^{b/2},$$

we arrive at

$$|f(t, m)| \leq C(r^2 - x^2)^{b/2-\alpha}e^{Kt}. \tag{3.2}$$

This leads to the zero trace for the PDF:

$$f(t, m)|_{|m|^2=b} = 0 \quad \forall t > 0.$$

The main difficulty of Theorem 1.2 lies in the singularity at the boundary. Equation (2.7) for v solves $L(v) = 0$ with

$$L(v) := (r^2 - |x|^2)\Delta_x v - bx \cdot \nabla_x v - (r^2 - |x|^2)\partial_t v.$$

We now introduce the transformation

$$v(t, x) := w(t, x)(r^2 - |x|^2)^{-\alpha}e^{Kt},$$

with α and K to be determined. A simple calculation gives

$$\begin{aligned}
\partial_t v &= (w_t + Kw)(r^2 - |x|^2)^{-\alpha}e^{Kt},\\
\nabla_x v &= [\nabla_x w(r^2 - |x|^2)^{-\alpha} + 2\alpha w x(r^2 - |x|^2)^{-\alpha-1}]e^{Kt},\\
\Delta_x v &= [\Delta_x w(r^2 - |x|^2)^{-\alpha} + 4\alpha x \cdot \nabla_x w(r^2 - |x|^2)^{-\alpha-1}\\
&\qquad + 4\alpha(\alpha + 1)w|x|^2(r^2 - |x|^2)^{-\alpha-2}]e^{Kt}\\
&\qquad + 2\alpha d w(r^2 - |x|^2)^{-(\alpha+1)}e^{Kt}.
\end{aligned}$$

Substitution of these terms into the equation $L(v) = 0$ multiplied by $(r^2-|x|^2)^{\alpha+1}e^{-Kt}$ gives

$$A(w) = 0,$$

where the operator $A(w)$ is defined as

$$\begin{aligned} A(w) := (r^2-|x|^2)^2\Delta_x w + (4\alpha-b)(r^2-|x|^2)x\cdot\nabla_x w \\ -(r^2-|x|^2)^2\partial_t w + c(x)w, \end{aligned}$$

in which the coefficient

$$c(x) = -K(r^2-|x|^2)^2 + 2\alpha[dr^2+(2\alpha+2-d-b)|x|^2].$$

In order to apply a maximum principle to $A(w)=0$, we need to choose α and K such that $c<0$ in $\Omega(T^*)$. Setting $|x|^2=\theta r^2$, we have

$$\begin{aligned} c &= -Kr^4(1-\theta)^2 + 2\alpha dr^2 + 2\alpha(2\alpha+2-d-b)\theta r^2 \\ &= -r^2\left\{Kr^2\theta^2 - 2((2\alpha+2-d-b)\alpha+Kr^2)\theta + Kr^2 - 2d\alpha\right\}. \end{aligned}$$

Thus as a function of θ, c achieves its maximum

$$c = K^{-1}\alpha\left\{2Kr^2(2\alpha+2-b) + \alpha(2\alpha+2-d-b)^2\right\}$$

at

$$\theta^* = 1 + \frac{\alpha}{Kr^2}(2-d-b+2\alpha).$$

The coefficient c can be made negative if its maximum value is negative, which is true provided that

$$\alpha < \frac{b}{2}-1$$

and

$$Kr^2 > \frac{\alpha(d+b-2-2\alpha)^2}{2(b-2-2\alpha)} > 0.$$

With these choices we apply the maximum principle [18] to the equation $A(w)=0$ and find that w achieves a positive maximum only at initial time, i.e., in the region $\{(0,x),\, |x|<r^2\}$. Therefore we have

$$0 \le w(t,x) \le \|w(0,\cdot)\|_{L^\infty}.$$

Note that

$$w_0(x) = v_0(x)(r^2-|x|^2)^\alpha = f_0(m)r^h(r^2-|x|^2)^{\alpha-b/2}.$$

Assume that $f_0(m)\neq 0$ for $|m|^2 \le b^* < b$. Then

$$\|w_0\|_\infty \le \|f_0\|_\infty r^b(r^2-2b^*)^{\alpha-b/2}.$$

Thus from

$$f(t,m) = v(t,x)(1-|x|^2/r^2)^{b/2} = w(t,x)r^{-b}(r^2-|x|^2)^{b/2-\alpha}e^{Kt},$$

it follows that

$$|f(t,m) \le \|f_0\|_\infty \le \left(\frac{r^2-|x|^2}{r^2-2b^*}\right)^{b/2-\alpha} e^{Kt}.$$

Replacing $r^2=2b$ and $|x|^2=2|m|^2$ we have obtained the desired estimate stated in Theorem 1.2. Therefore the trace of f on the sphere $|m|=\sqrt{b}$ must be null.

3.2. Coupled with flow $\kappa \neq 0$. We now show the null trace when flow is involved. In this case (2.7) has the form $L(v) = 0$ with

$$\begin{aligned} L(v) := {} & (r^2 - |x|^2)\Delta_x v - (bx + (r^2 - |x|^2)\kappa x) \cdot \nabla_x v \\ & - (r^2 - |x|^2)\partial_t v + (bx^\top \kappa x)v. \end{aligned}$$

We again apply the transformation

$$v(t, x) := w(t, x)(r^2 - |x|^2)^{-\alpha} e^{Kt},$$

with α and K to be determined. This transformation applied to $L(v)(r^2-|x|^2)^{\alpha+1}e^{-Kt}$ leads to the following equation:

$$B(w) = 0,$$

with the operator $B(w)$ being

$$\begin{aligned} B(w) = {} & (r^2 - |x|^2)^2\Delta_x w + (r^2 - |x|^2)[(4\alpha - b)x \\ & - (r^2 - |x|^2)\kappa x] \cdot \nabla_x w - (r^2 - |x|^2)^2\partial_t w + c(x)w, \end{aligned}$$

and the coefficient of the last term being

$$\begin{aligned} c(x) = {} & -K(r^2 - |x|^2)^2 + 2\alpha[dr^2 + (2\alpha + 2 - d - b)|x|^2] \\ & + (b - 2\alpha)x^\top \kappa x(r^2 - |x|^2). \end{aligned}$$

Using a similar argument as in the no-flow case, we proceed to determine α and K so that c stays negative in $\Omega(T)$. Let ρ be the largest eigenvalues of the deformation tensor

$$S = (\kappa + \kappa^\top)/2;$$

one has

$$x^\top \kappa x \leq \rho|x|^2.$$

We first choose α in such a way that $\alpha < b/2 - 1$, which implies $h - 2\alpha > 0$. Thus we obtain

$$\begin{aligned} c(x) \leq \bar{c} = {} & -K(r^2 - |x|^2)^2 + 2\alpha[dr^2 + (2\alpha + 2 - d - b)|x|^2] \\ & + \rho(b - 2\alpha)|x|^2(r^2 - |x|^2) \\ = {} & -[K + \rho(b - 2\alpha)]|x|^4 + [2Kr^2 + \rho(h - 2\alpha)r^2 \\ & + 2\alpha(2\alpha + 2 - d - b)]|x|^2 + 2\alpha dr^2 - Kr^4 \\ \leq {} & 2\alpha dr^2 - Kr^4 \\ & + \frac{[2Kr^2 + \rho(b - 2\alpha)r^2 + 2\alpha(2\alpha + 2 - d - b)]^2}{4[K + \rho(h - 2\alpha)]} \\ \leq {} & 2\alpha r^2(2\alpha + 2 - b) + \frac{\beta^2}{4[K + \rho(b - 2\alpha)]}, \end{aligned}$$

where

$$\beta := \rho(b - 2\alpha)r^2 - 2\alpha(2\alpha + 2 - d - b).$$

Therefore we can choose K such that

$$K > \frac{\beta^2}{8\alpha r^2(b-2-2\alpha)} - \rho(b-\alpha),$$

and the following is always true:

$$c(x) \le \bar{c} < 0, \quad |x| \le r^2.$$

We thus can apply the maximum principle [18] to the equation

$$B(w) = 0$$

and obtain that u achieves its positive maximum only at initial time, i.e., in the region $\{(0,x),\, |x| < r^2\}$. This will give the result that

$$0 \le w(t,x) \le \|w(0,\cdot)\|_{L^\infty}.$$

Converting back to f we prove the results stated in Theorem 1.2.

4. Appendix.

In this appendix, we recall some basic facts concerning the Fichera function which we used in section 2. Consider the second order equation

$$L[u] := a^{jk}D_{jk}u + b^i D_i u + cu = f \text{ in } \Omega$$

with the condition

$$a^{ij}(x)\xi_i\xi_j \ge 0$$

for any $\xi \in \mathbb{R}^d$ and $x \in \Omega$. This class of equations with noncharacteristic form includes equations of elliptic and parabolic types, first order equations, the equations of Brownian motion, and others. The first boundary value problem in its general form was set up by Fichera [8]. We assume that $a^{ij} \in C^2(\Omega)$, $b^i \in C^1(\Omega)$, and $c \in C^0(\Omega)$. Let n denote the unit outward normal vector to $\Gamma = \partial\Omega$ at $x \in \Gamma$. The Fichera function is defined as

$$\mathfrak{F} = (b^k - D_j a^{kj})n_k(x) : \Gamma \to \mathbb{R}.$$

Thus the boundary is classified into several parts based on the sign of Fichera function $\mathfrak{F}$:

$$\Gamma = \Gamma_e \cup \Gamma_h,$$

where $\Gamma_e := \{x \in \Gamma,\, a^{ij}n_in_j > 0\}$ is the noncharacteristic (positive characteristic) part, and $\Gamma_h := \{x \in \Gamma,\, a^{ij}n_in_j = 0\}$, in which there are two subsets (irrelevant and relevant parts):

$$\Gamma_+ = \{x,\, \mathfrak{F}(x) \ge 0\}, \quad \Gamma_- = \{x,\, \mathfrak{F}(x) < 0\}.$$

The classical theory of Fichera [8] says that the Dirichlet boundary condition leads to a well-posed problem: find a function u in $\Omega \cup \Gamma$ such that

$$L(u) = f \text{ in } \Omega, \quad u = g \text{ on } \Gamma_e \cup \Gamma_-. \tag{4.1}$$

In other words, no Dirichlet boundary data is necessarily imposed on Γ_-, where $\mathfrak{F} \ge 0$.

It was shown that the sign of the Fichera function at points on Γ_h does not change under smooth nondegenerate changes of independent variables in the equation. Consequently, we have the following theorem.

THEOREM 4.1 (see [18]). *The subsets $\Gamma_e, \Gamma_+, \Gamma_-$ of the boundary Γ, defined for the operator L, remain invariant under smooth nonsingular changes of independent variables in the equation.*

To define the weak solution for the underlying problem, we introduce an adjoint operator as

$$L^*[v] = -D_{ij}(a^{ij}v) - D_i(b^i v) + cv.$$

Then for $v \in C^2(\bar{\Omega})$ and $v = 0$ on $\Gamma_e \cup \Gamma_-$ the weak formulation of the equation becomes

$$\int_\Omega uL^*[v]dx = \int_\Omega vL[u]dx.$$

Weak solution in $L^2(\Omega)$. A bounded measurable function $u(x)$ will be called a weak solution of the above problem with $u \in \{L^2(\Omega),\, u = 0,\, x \in \Gamma_e \cup \Gamma_-\}$ if

$$\int_\Omega uL^*[v]dx = (f, v)$$

for all $v \in \{C^2(\bar{\Omega}),\, v = 0,\, x \in \Gamma_e \cup \Gamma_-\}$, where $(\cdot, \cdot)$ denotes the inner product in L^2. Weak solution in other spaces such as the L^p space or general Hilbert space can be defined; and solution smoothness can be further studied once existence of a weak solution is established. Existence results have been proven under various assumptions, e.g., the following theorem.

THEOREM 4.2. *Suppose the inequality*

$$\frac{1}{2}D_i b^i - \frac{1}{2}D_{ij}a^{ij} - c \geq c_0 > 0$$

is satisfied in $\bar{\Omega}$, and let $f \in L^2(\Omega)$. Then there exists a function $u \in L^2(\Omega)$ which is a weak solution of (4.1) *in the sense stated above.*

Acknowledgments. We acknowledge discussion with and suggestions by Gary Lieberman. We also thank the referees for valuable suggestions and pointing out two relevant references [13, 15]. We also want to thank the IMA for the hospitality and support during our stay for the *Thematic Year on Mathematics of Materials and Macromolecules: Multiple Scales, Disorder, and Singularities.*

REFERENCES

[1] M. BARDI AND S. BOTTACIN, *Characteristic and Irrelevant Boundary Points for Viscosity Solutions of Nonlinear Degenerate Elliptic Equations*, Preprint 25, Dip. di Mathematica P. E. A, University di Padova, Padova, Italy, 1998.

[2] J. W. BARRETT, C. SCHWAB, AND E. SULI, *Existence of global weak solutions for some polymeric flow models*, Math. Models Methods Appl. Sci., 15 (2005), pp. 939–983.

[3] R. B. BIRD, C. F. CURTISS, R. C. ARMSTRONG, AND P. HASSAGER, *Dynamics of Polymeric Liquids, Volume* 2: *Kinetic Theory*, Wiley Interscience, New York, 1987.

[4] R. B. BIRD, O. HASSAGER, R. C. ARMSTRONG, AND C. F. CURTISS, *Dynamics of Polymeric Fluids,* 2, *Kinetic Theory*, John Wiley and Sons, New York, 1977.

[5] M. DOI AND S. F. EDWARDS, *The Theory of Polymer Dynamics*, Oxford University Press, Oxford, UK, 1986.

[6] Q. Du, C. Liu, and P. Yu, *FENE dumbbell model and its several linear and nonlinear closure approximations*, Multiscale Model. Simul., 4 (2005), pp. 709–731.

[7] W.-N. E, T. Li, and P. Zhang, *Well-posedness for the dumbell model of polymeric fluids*, Comm. Math. Phys., 248 (2004), pp. 409–427.

[8] G. Fichera, *Sulle equazioni differenziali lineari ellittico-paraboliche del secondo ordine*, Atti Accad. Naz. Lincei. Mem. Cl. Sci. Fis. Mat. Nat. Sez. I (8), 5 (1956), pp. 1–30.

[9] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer, Berlin, 1997.

[10] Y. Guo, *The Vlasov-Maxwell-Boltzmann system near Maxwellians*, Invent. Math., 153 (2003), pp. 593–630.

[11] P. Halin, G. Lielens, R. Keunings, and V. Legat, *The Lagrangian particle method for macroscopic and micro-macro viscoelastic flow computations*, J. Non-Newtonian Fluid Mech., 79 (1998), pp. 387–403.

[12] B. Jourdain, T. Lelièvre, and C. Le Bris, *Existence of solution for a micro-macro model of polymeric fluid: The FENE model*, J. Funct. Anal., 209 (2004), pp. 162–193.

[13] B. Jourdain and T. Lelièvre, *Mathematical analysis of a stochastic differential equation arising in the micro-macro modelling of polymeric fluids*, in Probabilistic Methods in Fluids, I. M. Davies, N. Jacob, A. Truman, O. Hassan, K. Morgan, and N. P. Weatherill, eds., World Scientific, River Edge, NJ, 2002, pp. 205–223.

[14] R. G. Larson, *The Structure and Rheology of Complex Fluids*, Oxford University Press, New York, 1995.

[15] C. Le Bris, B. Jourdain, T. Lelièvre, and F. Otto, *Long-time asymptotic of a multi-scale model for polymeric fluid flows*, Arch. Ration. Mech. Anal., 181 (2006), pp. 97–148.

[16] C. Le Bris and P. L. Lions, *Renormalized solutions of some transport equations with partially $W^{1,1}$ velocities and applications*, Ann. Mat. Pura Appl. (4), 183 (2004), pp. 97–130.

[17] F. H. Lin, C. Liu, and P. Zhang, *On a micro-macro model for polymeric fluids near equilibrium*, Comm. Pure Appl. Math., 60 (2006), pp. 838–866.

[18] O. A. Oleinik and E. V. Radkevič, *Second Order Equations with Nonnegative Characteristic Form*, AMS, Providence, RI, and Plenum Press, New York, 1973.

[19] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.

[20] D. Strook and S. R. S. Varadhan, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.

[21] K. Taira, *Diffusion Processes and Partial Differential Equations*, Academic Press, Boston, 1988.

[22] P. Yu, Q. Du, and C. Liu, *From micro to macro dynamics via a new closure approximation to the FENE model of polymeric fluids*, Multiscale Model. Simul., 3 (2005), pp. 895–917.

[23] H. Zhang and P. Zhang, *Local existence for the FENE-dumbbell model of polymeric fluids*, Arch. Ration. Mech. Anal., 181 (2006), pp. 373–400.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1316–1332

TRAVELING WAVES AND SHOCKS IN A VISCOELASTIC GENERALIZATION OF BURGERS' EQUATION*

VICTOR CAMACHO†, ROBERT D. GUY‡, AND JON JACOBSEN†

Abstract. We consider traveling wave phenomena for a viscoelastic generalization of Burgers' equation. For asymptotically constant velocity profiles we find three classes of solutions corresponding to smooth traveling waves, piecewise smooth waves, and piecewise constant (shock) solutions. Each solution type is possible for a given pair of asymptotic limits, and we characterize the dynamics in terms of the relaxation time and viscosity.

Key words. non-Newtonian fluids, fast-slow dynamics, vanishing viscosity solutions

AMS subject classifications. 35L67, 76A10, 35Q53

DOI. 10.1137/070687840

1. Introduction.

Burgers' equation

$$u_t + uu_x = \epsilon u_{xx} \tag{1.1}$$

is perhaps the simplest model that couples the nonlinear convective behavior of fluids with the dissipative viscous behavior. Introduced by Burgers [5] as a model for turbulence, (1.1) and its inviscid counterpart,

$$u_t + uu_x = 0, \tag{1.2}$$

are essential for their role in modeling a wide array of physical systems such as traffic flow, shallow water waves, and gas dynamics [17, 18, 19, 23]. The equations also provide fundamental pedagogical examples for many important topics in nonlinear PDEs such as traveling waves, shock formation, similarity solutions, singular perturbation, and numerical methods for parabolic and hyperbolic equations (see, e.g., [9, 14, 20, 23]).

The parabolic equation (1.1) has the property that smooth initial data yield smooth solutions for all $t > 0$. In contrast, smooth initial data for the hyperbolic equation (1.2) can develop jump discontinuities in finite time (shock formation). One technique for studying shock wave solutions of (1.2) is to study smooth traveling wave solutions of (1.1) in the limit as $\epsilon \to 0$.

In this paper we consider how the addition of viscoelasticity affects traveling wave solutions of Burgers' equation. The equations we consider are

$$u_t + uu_x = \sigma_x, \tag{1.3}$$

$$\sigma_t + u\sigma_x - \sigma u_x = \alpha u_x - \beta\sigma. \tag{1.4}$$

*Received by the editors April 10, 2007; accepted for publication (in revised form) January 14, 2008; published electronically April 11, 2008.
http://www.siam.org/journals/siap/68-5/68784.html

†Mathematics Department, Harvey Mudd College, 301 Platt Blvd., Claremont, CA 91711 (vcamacho@hmc.edu, jacobsen@math.hmc.edu). The research of the first author was supported by a grant from the Baker Foundation. The research of the third author was supported by a grant from the Beckman Research Program at Harvey Mudd College.

‡Mathematics Department, University of California Davis, Davis, CA 95616 (guy@math.ucdavis.edu). The research of this author was supported in part by NSF grants DMS-0139926 and DMS-0540779.

The constitutive law (1.4) resembles a one-dimensional version of the upper convected Maxwell model [11]. The relaxation time is $\lambda = \beta^{-1}$, and $\alpha = \mu\lambda^{-1}$ could be interpreted as the elastic modulus of the material if there were no relaxation of stress ($\beta = 0$). In the other limit of instantaneous relaxation of stress ($\lambda \to 0$), (1.4) reduces to $\sigma = \mu u_x$, and the system (1.3)–(1.4) is equivalent to Burgers' equation (1.1) with fluid viscosity $\mu = \epsilon$.

The remainder of the paper is organized as follows. In section 2 we give a brief introduction to viscoelastic fluids and explain the reduction and constitutive law for our model. We show in section 3 that traveling wave solutions to (1.3)–(1.4) exist only when the viscosity (or elastic modulus) is above a certain threshold. As the viscosity approaches this threshold, singularities in the derivative appear, and numerical experiments suggest that shocks develop when the viscosity is below threshold. The system (1.3)–(1.4) is nonconservative, and therefore the classical theory for systems of conservation laws (cf. [9, 23]) cannot be used to analyze singular solutions. A generalized theory of weak solutions to nonconservative hyperbolic equations has been developed for such problems [2, 7, 8].

We take a different approach and analyze the shock solutions by introducing an additional viscosity to regularize the problem. Using singular perturbation theory, we show in section 4 that traveling waves exist for all parameters in the regularized problem, and the waves limit to shock solutions as the additional viscosity goes to zero. This method of *vanishing viscosity* is a well-known technique for analyzing weak solutions of nonconservative hyperbolic equations such as the Hamilton–Jacobi equations [9]. Finally, in section 5 we discuss the effect of different parameters on the solution structure, how the results depend on the choice of one-dimensional reduction, and a possible application of the results to numerical methods for viscoelastic flows.

2. Viscoelastic fluids. In this section we discuss how the constitutive law in (1.4) is related to a standard constitutive law for viscoelastic fluids. The discussion here is not meant to be extensive. For more comprehensive treatments of viscoelastic fluids, see [3, 4, 11, 12].

The incompressible Navier–Stokes equations are

$$\rho\left(\boldsymbol{u}_t + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = -\nabla p + \mu\Delta\boldsymbol{u}, \tag{2.1}$$

$$\nabla\cdot\boldsymbol{u} = 0. \tag{2.2}$$

The momentum equation (2.1) can be expressed as

$$\rho\left(\boldsymbol{u}_t + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = -\nabla p + \nabla\cdot\boldsymbol{\sigma}_{\mathbf{v}}, \tag{2.3}$$

where the (Newtonian) viscous stress $\boldsymbol{\sigma}_{\mathbf{v}}$ is defined by

$$\boldsymbol{\sigma}_{\mathbf{v}} = 2\mu\boldsymbol{D} = \mu\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\mathrm{T}}\right). \tag{2.4}$$

This Newtonian constitutive law means that the fluid stress is proportional to the deformation rate tensor. In contrast, the stress in viscoelastic fluids includes some time history of the deformation.

One of the simplest constitutive laws for viscoelastic materials is the Maxwell model. Consider a linear spring and dashpot in series, with spring constant k and damping coefficient μ. The stress, σ, in the element is

$$\lambda\dot{\sigma} + \sigma = \mu\dot{\epsilon}, \tag{2.5}$$

where ϵ is the strain in the element, and $\lambda = k/\mu$ is the relaxation time. The linear Maxwell model for a continuum is

$$\lambda \boldsymbol{\sigma}_t + \boldsymbol{\sigma} = 2\mu \boldsymbol{D}. \tag{2.6}$$

However, this is not a valid constitutive law because it is not frame invariant [11]. That is, the stress depends on the reference frame. Frame invariance is achieved by choosing an appropriate time derivative, akin to the material derivative for the velocity field. One frame invariant time derivative is the upper convected derivative, defined by

$$\overset{\triangledown}{\boldsymbol{S}} = \boldsymbol{S}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{S} - \nabla \boldsymbol{u}\, \boldsymbol{S} - \boldsymbol{S}\, \nabla \boldsymbol{u}^{\mathrm{T}}. \tag{2.7}$$

Replacing the partial time derivative in (2.6) with the upper convected derivative gives the upper convected Maxwell (UCM) equation

$$\lambda \overset{\triangledown}{\boldsymbol{\sigma}} + \boldsymbol{\sigma} = 2\mu \boldsymbol{D}. \tag{2.8}$$

The ij component in (2.8) satisfies

$$\lambda \left(\frac{\partial \sigma_{ij}}{\partial t} + u_k \frac{\partial \sigma_{ij}}{\partial x_k} - \frac{\partial u_i}{\partial x_k} \sigma_{kj} - \sigma_{ik} \frac{\partial u_j}{\partial x_k} \right) + \sigma_{ij} = \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right), \tag{2.9}$$

where summation is over the repeated index k. Although there are many other frame invariant derivatives, in this paper we consider a one-dimensional reduction, in which case they yield identical reductions.

A one-dimensional version of the UCM equation is

$$\lambda \left(\sigma_t + u\sigma_x - \sigma u_x \right) + \sigma = \mu u_x. \tag{2.10}$$

However, there are other reasonable choices for a one-dimensional UCM equation. For example, the equation for σ_{11} when $\boldsymbol{u} = (u_1, 0, 0)$ is

$$\lambda \left(\sigma_t + u\sigma_x - 2\sigma u_x \right) + \sigma = 2\mu u_x, \tag{2.11}$$

where we have dropped the subscripts on the stress and velocity. The upper convected derivative must be used in (2.8) because this is the time derivative of a tensor in a moving continuum. In one dimension, the stress is a scalar, so it would also be reasonable to simply use the material derivative for the time derivative. In this case the constitutive law is

$$\lambda \left(\sigma_t + u\sigma_x \right) + \sigma = \mu u_x. \tag{2.12}$$

In this paper we analyze the first UCM equation (2.10). While all three models have similar results, (2.10) is more robust, in that all of the phenomena that occur in (2.11) and (2.12) also occur in (2.10). In section 5 we discuss how the results change if (2.11) or (2.12) is used instead.

Equation (2.10) is equivalent to (1.4). This is seen by dividing through by the relaxation time λ to get

$$\sigma_t + u\sigma_x - \sigma u_x = \alpha u_x - \beta\sigma, \tag{2.13}$$

where

$$\alpha = \mu\lambda^{-1}, \tag{2.14}$$

$$\beta = \lambda^{-1}. \tag{2.15}$$

The parameter α could be interpreted as the elastic modulus of the material if there were no relaxation of stress ($\beta = 0$). It is somewhat arbitrary whether the constitutive law is expressed in terms of the relaxation time (λ) and viscosity (μ) or elastic modulus (α) and decay rate (β). In this paper we primarily use the latter, but sometimes we express results using both sets of parameters for additional insight.

In section 4 we consider a modification to the Maxwell constitutive law (1.4). We include a second viscous term, one without memory, so that the system becomes

$$u_t + uu_x = \sigma_x + \epsilon u_{xx}, \tag{2.16}$$

$$\sigma_t + u\sigma_x - \sigma u_x = \alpha u_x - \beta\sigma. \tag{2.17}$$

The addition of the second viscous term can be considered as a one-dimensional version of the Oldroyd-B constitutive law [12].

We note that the one-dimensional constitutive law studied in this paper is not a *physical* reduction from the three-dimensional UCM. It is a reduction in the same sense that Burgers' equation is a reduction. One may wonder what, if any, physical significance there is to the problem that we analyze in this paper. Using high-resolution Godunov schemes for the advection terms in the Navier–Stokes equations requires solving Burgers' equation [1]. Analogously, systems of the form (2.10) and (2.11) arise in the application of wave propagation schemes to viscoelastic fluids [10, 22]. This was the original inspiration for this study but not the sole motivation. It is interesting to explore what happens to traveling waves in Burgers' equation (1.1) if the viscous term is replaced by a viscoelastic term, and the most natural starting point is the Maxwell model. Thus the one-dimensional constitutive laws considered were chosen to resemble the UCM equation.

3. Traveling waves.

To find traveling wave solutions to (1.3)–(1.4) we consider solutions of the form $u(x,t) = U(\xi)$ and $\sigma(x,t) = S(\xi)$, where $\xi = x - ct$ for some constant c. In traveling wave coordinates, the system is

$$-cU' + UU' = S', \tag{3.1}$$

$$-cS' + US' - SU' = \alpha U' - \beta S. \tag{3.2}$$

We consider traveling waves that correspond to heteroclinic connections between two equilibrium points with given velocity values at infinity. The equilibrium points of the system correspond to all states with $S = 0$, and thus we assume the following asymptotic boundary conditions:

$$U(-\infty) = u_\ell, \quad S(-\infty) = 0, \tag{3.3}$$

$$U(\infty) = u_r, \quad S(\infty) = 0. \tag{3.4}$$

In the next section we examine for which values of u_ℓ, u_r, α, and β do solutions of this problem exist.

3.1. Existence.

Integrating (3.1) gives the stress in terms of the velocity as

$$S = \frac{U^2}{2} - cU + A, \tag{3.5}$$

where A is the integration constant. Applying the boundary conditions, the wave speed and integration constant are

$$A = \frac{u_\ell u_r}{2}, \tag{3.6}$$

$$c = \frac{u_\ell + u_r}{2}. \tag{3.7}$$

Note that if a traveling wave exists, then it moves with the same speed as traveling waves in Burgers' equation (1.1) and shock waves in the inviscid Burgers equation (1.2).

We obtain the equation for the velocity profile, U, by using (3.1) and (3.5) to eliminate S and S' in (3.2) to get

$$U' = \frac{-\beta\big(U\,(U/2 - c) + A\big)}{U\,(U/2 - c) + c^2 - A - \alpha}. \tag{3.8}$$

Using (3.6) and (3.7), this simplifies to

$$U' = \frac{-\beta(U - u_\ell)(U - u_r)}{(U - u_\ell)(U - u_r) + 2\left(\left(\frac{u_\ell - u_r}{2}\right)^2 - \alpha\right)}. \tag{3.9}$$

From the dynamics of this equation we extract conditions for the existence of traveling waves. The two equilibrium points are clearly $U = u_\ell$ and $U = u_r$, and a traveling wave corresponds to a one-dimensional flow from one equilibrium point to the other. There are two cases to consider: $u_\ell > u_r$ and $u_\ell < u_r$.

First we suppose that $u_\ell > u_r$. For a traveling wave to exist, we need that $U' < 0$ for $U \in (u_r, u_\ell)$. The numerator of (3.9) is positive in this interval. The maximum value of $(U - u_\ell)(U - u_r)$ is 0, and so the denominator is always negative, provided $((u_\ell - u_r)/2)^2 - \alpha < 0$, in which case $U' < 0$ for $U \in (u_r, u_\ell)$.

Next, consider the case $u_\ell < u_r$. A traveling wave exists if $U' > 0$ for $U \in (u_\ell, u_r)$. As before, the numerator of (3.9) is positive for $U \in (u_\ell, u_r)$, and thus we examine the sign of the denominator. The minimum value of $(U - u_\ell)(U - u_r)$ is $-\left((u_\ell - u_r)/2\right)^2$, in which case it follows that $U' > 0$, provided $((u_\ell - u_r)/2)^2 - 2\alpha > 0$.

Combining these two cases, we have the following result: a traveling wave solution to (1.3)–(1.4) with boundary conditions (3.3)–(3.4) exists if and only if

$$u_\ell > u_r \quad \text{and} \quad \alpha > \left(\frac{u_\ell - u_r}{2}\right)^2 \tag{3.10}$$

or

$$u_\ell < u_r \quad \text{and} \quad 2\alpha < \left(\frac{u_\ell - u_r}{2}\right)^2. \tag{3.11}$$

Equivalently, no traveling wave solutions exist if

$$\frac{(u_\ell - u_r)^2}{8} \le \alpha \le \frac{(u_\ell - u_r)^2}{4}. \tag{3.12}$$

Using (2.14) to express this condition in terms of the relaxation time and viscosity, we see that no traveling wave solution exists if

$$\frac{(u_\ell - u_r)^2}{8} \le \frac{\mu}{\lambda} \le \frac{(u_\ell - u_r)^2}{4}. \tag{3.13}$$

In comparison, for the viscous Burgers equation (1.1), traveling waves with $u_\ell > u_r$ exist for any positive viscosity. By adding elasticity we see that, for a fixed relaxation time λ, there is now a minimal viscosity required for such waves to exist. In the following sections we explore what happens to these wave solutions when the viscosity is reduced beyond this minimal viscosity.

3.2. Wave profile. The shape of the wave is found by integrating (3.9). The solution is

$$\beta\,(\xi - \xi_0) = \frac{2\Big(\left(\frac{u_\ell - u_r}{2}\right)^2 - \alpha\Big)}{u_\ell - u_r} \log\left|\frac{U(\xi) - u_r}{U(\xi) - u_\ell}\right| - U(\xi). \tag{3.14}$$

When a traveling wave exists, the profile is defined implicitly by (3.14). However, when a traveling wave fails to exist, we can still plot the implicit solutions of (3.14). In Figure 3.1 we plot the curve defined by (3.14) for four different values of α, while keeping the other parameter values fixed at $u_\ell = 2$, $u_r = 0$, and $\beta = 1$. For these parameter values, a traveling wave exists when $\alpha > 1$. In Figure 3.1(a) the wave profile is shown for $\alpha = 1.2$. As α approaches 1, the wave profile approaches the piecewise linear function shown in Figure 3.1(b). As α is decreased further, the curve becomes multivalued and the asymptotic values are no longer satisfied. Figure 3.1(c) shows the solution for $\alpha = 0.9$. As α decreases even further, the solution of (3.14) returns to being single-valued but no longer yields a traveling wave solution with the given asymptotic limits. This transition occurs at $\alpha = \frac{1}{2}\left(\frac{u_\ell - u_r}{2}\right)^2$, which corresponds to when U' returns to being one-signed (now positive), corresponding to the lower limit of (3.13). Figure 3.1(d) shows the solution for $\alpha = 0.25$.

3.3. Numerical simulations. In this section we consider numerical simulations of the full PDE system (1.3)–(1.4). According to (3.10), when $u_\ell > u_r$ there is a minimal viscosity in order for traveling waves to exist. In numerical simulations of this case, these traveling wave solutions appear to be stable and travel with the speed $c = (u_\ell + u_r)/2$, as in (3.7). We found that for any initial data, as long as the asymptotic limits were maintained, the solution approached the traveling wave profile given by (3.14). On the other hand, according to (3.11), when $u_\ell < u_r$, traveling waves exist as long as the viscosity is below a certain threshold. In simulations of the PDE system for this case, these waves did not appear to be stable; rather the solutions always rarefy. Accordingly, from this point on we consider only the stable case of $u_\ell > u_r$.

We next consider what happens when the viscosity is below the minimal value, corresponding to the implicit plots shown in Figure 3.1(c)–(d). We solve the full system (1.3)–(1.4) numerically by splitting the update at each time step into three substeps. First we take a step including only the advection terms

$$u_t + uu_x = 0, \tag{3.15}$$

$$\sigma_t + u\sigma_x = 0 \tag{3.16}$$

and use an upwinding method. Next we take a step including the elastic terms

$$u_t = \sigma_x, \tag{3.17}$$

$$\sigma_t - \sigma u_x = \alpha u_x. \tag{3.18}$$

We linearize the σu_x term in each grid cell by treating this term as $\sigma_j^n u_x$ through the time step, where σ_j^n is the value of the stress at time step n at grid cell j. This

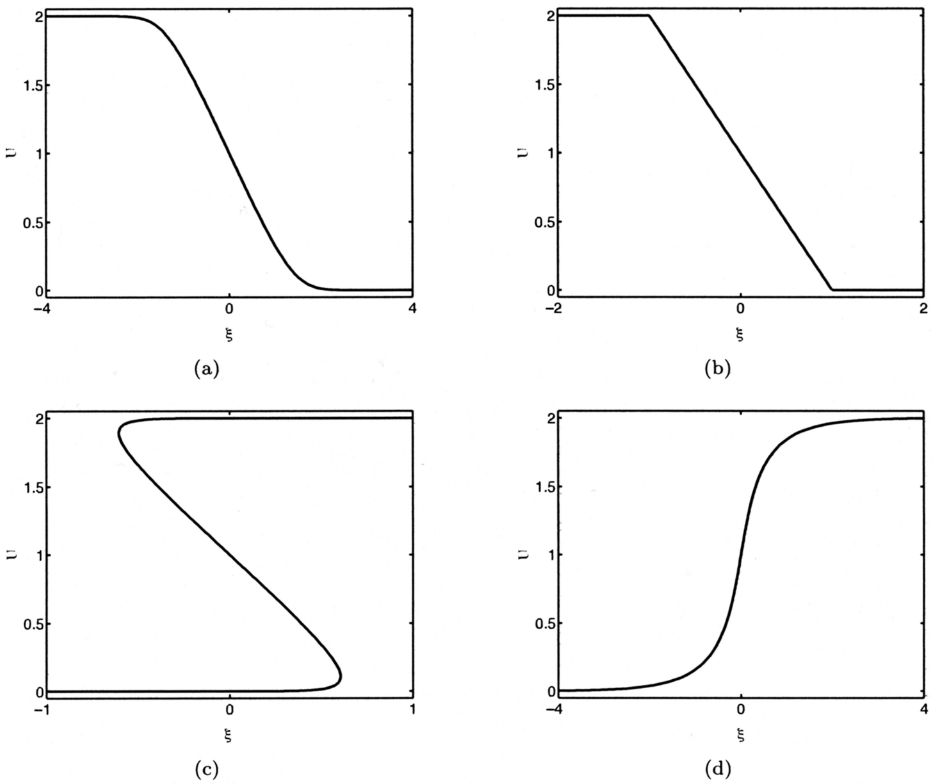


FIG. 3.1. *Plots of the solution curves to* (3.14)*. The parameters are* $u_\ell = 2$*,* $u_r = 0$*,* $\beta = 1$*. Four different values of* α *are plotted:* (a) $\alpha = 1.2$*,* (b) $\alpha = 1$*,* (c) $\alpha = 0.9$*,* (d) $\alpha = 0.25$*. For these values of* u_ℓ *and* u_r*, no wave exists for* $\alpha < 1$*.*

linearized system is a variable coefficient wave equation, which we update by a wave propagation method as described in [16]. Finally, we update the stress by taking a step of

$$\sigma_t = -\beta\sigma. \tag{3.19}$$

For the initial condition we set the velocity equal to the traveling wave profile corresponding to the viscous Burgers equation with a given viscosity and set the stress to zero.

As suggested by Figure 3.1(c)–(d), we find two distinct cases, corresponding to whether

$$\frac{1}{2}\left(\frac{u_\ell - u_r}{2}\right)^2 < \alpha < \left(\frac{u_\ell - u_r}{2}\right)^2 \tag{3.20}$$

or

$$0 < \alpha < \frac{1}{2}\left(\frac{u_\ell - u_r}{2}\right)^2. \tag{3.21}$$

In both cases we find that the solutions develop into traveling waves, now, however, with jump discontinuities in the wave profile. These numerical solutions propagate with the wave speed $c = (u_\ell + u_r)/2$, the same wave speed as smooth traveling wave solutions. When α satisfies (3.20) the profile is piecewise smooth, with two shocks, as indicated in Figure 3.2(a). We refer to this solution as the double-shock solution. As α ranges between the limiting values of (3.20) the height of each jump discontinuity ranges from 0 when $\alpha = \left(\frac{u_\ell - u_r}{2}\right)^2$ to 1 when $\alpha = \frac{1}{2}\left(\frac{u_\ell - u_r}{2}\right)^2$, which yields a piecewise constant solution. This piecewise constant solution persists when α satisfies (3.21), as indicated in Figure 3.2(b). This resembles a classic shock solution of the Riemann problem for the inviscid Burgers equation.

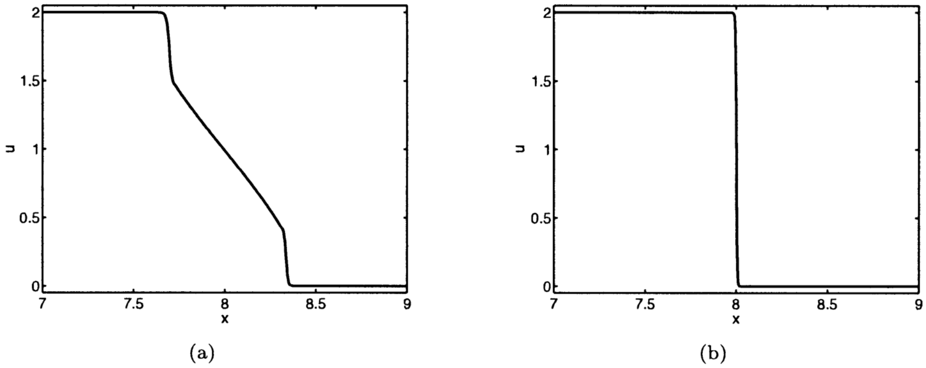


FIG. 3.2. *Plots of the wave profile found by solving* (1.3)–(1.4) *with smooth traveling wave initial data. The simulations were run until the profile stabilized. The smooth waves develop apparent jump discontinuities, whose type depends on whether α satisfies* (3.20) *or* (3.21)*, and travel with fixed speed. The parameter values are $u_\ell = 2, u_r = 0, \beta = 1$, and* (a) $\alpha = 0.8$*;* (b) $\alpha = 0.25$.

When solving equations with discontinuities care must be taken in order to capture the correct solution. These numerical solutions may not be the correct solutions, but they raise several questions that warrant further investigation. For example, as the PDE is not given by a system of conservation laws, what is the "correct" weak solution? In the case of the double-shock solution, what determines the shock height? What determines the shape of the solution between the two shocks? Why is it that we see a double-shock solution? In the next section we answer these questions by introducing a second viscous term to regularize the equations and analyzing the system in the limit of small viscosity.

4. Vanishing viscosity solution.

In this section we add a viscous regularization term on the velocity:

$$u_t + uu_x = \sigma_x + \epsilon u_{xx}, \tag{4.1}$$

$$\sigma_t + u\sigma_x - \sigma u_x = \alpha u_x - \beta\sigma \tag{4.2}$$

for $\epsilon > 0$. With the extra viscous term, this system can be viewed as a one-dimensional version of the Oldroyd-B constitutive law [12]. To study the double-shock and shock solutions of (1.3)–(1.4) we consider traveling wave solutions of this extended system in the limit $\epsilon \to 0$.

In traveling wave coordinates, the system becomes

$$-cU' + UU' = S' + \epsilon U'', \tag{4.3}$$

$$-cS' + US' - SU' = \alpha U' - \beta S. \tag{4.4}$$

Integrating (4.3), applying the asymptotic boundary conditions, and eliminating U' in (4.4) yields the system

$$\epsilon U' = \frac{1}{2}(U-u_\ell)(U-u_r) - S, \tag{4.5}$$

$$\epsilon(U-c)S' = (S+\alpha)\left(\frac{1}{2}(U-u_\ell)(U-u_r) - S\right) - \epsilon\beta S. \tag{4.6}$$

This system has precisely two equilibrium points: $(u_\ell, 0)$ and $(u_r, 0)$. A traveling wave solution of the PDE system (4.1)–(4.2) corresponds to a heteroclinic orbit connecting these two equilibrium points, as in Figure 4.1 (recall that we are assuming $u_\ell > u_r$).

Note that if a traveling wave of the original system (1.3)–(1.4) exists (i.e., when $\alpha > (u_\ell - u_r)^2/4$), then the wave corresponds to the trajectory in the phase plane defined by (3.5), or, equivalently,

$$S = \frac{1}{2}(U-u_\ell)(U-u_r). \tag{4.7}$$

This is the U-nullcline from (4.5) (for all ϵ).

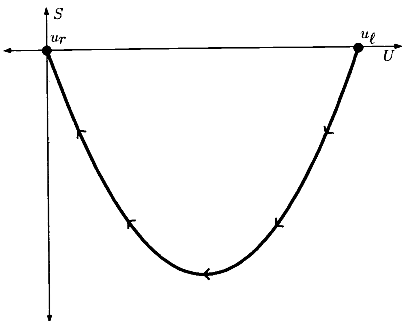


FIG. 4.1. *Heteroclinic orbit corresponding to traveling wave solution of system* (4.1)–(4.2).

The system (4.5)–(4.6) exhibits symmetric behavior about the line $U = c$, where $c = (u_\ell + u_r)/2$ is the wave speed for the inviscid case ($\epsilon = 0$). In particular, if $(\widehat{U}(\xi), \widehat{S}(\xi))$ solves (4.5)–(4.6) with $\widehat{U} > c$ for $\xi \in (-b, \xi_0)$ and $\widehat{U}(\xi_0) = c$, then $(U(\xi), S(\xi)) = (2c - \widehat{U}(-\xi + 2\xi_0), \widehat{S}(-\xi + 2\xi_0))$ solves (4.5)–(4.6) for $\xi \in (\xi_0, b + 2\xi_0)$, with $U < c$ and $U(\xi_0) = c$. This corresponds to the reflection of the trajectory through the line $U = c$.

The Jacobian of the system at the equilibrium point $(u_\ell, 0)$ is

$$J = J(u_\ell, 0) = \begin{bmatrix} \frac{d}{2\epsilon} & -\frac{1}{\epsilon} \\ \frac{\alpha}{\epsilon} & -\frac{2(\alpha+\beta\epsilon)}{\epsilon d} \end{bmatrix}, \tag{4.8}$$

where $d = u_\ell - u_r$. Since $\det(J) = -\frac{\beta}{\epsilon} < 0$, it follows that $(u_\ell, 0)$ is a saddle point for all $\epsilon > 0$. Thus the reflection through $U = c$ maps the unstable manifold of $(u_\ell, 0)$ to the stable manifold of $(u_r, 0)$. For this reason, to establish the existence of a heteroclinic orbit connecting the two, it suffices to establish that the unstable manifold of $(u_\ell, 0)$ crosses the line $U = c$.

The positive eigenvalue of $J(u_\ell, 0)$ is

$$\lambda_\ell = \frac{1}{4\epsilon d}\left(d^2 - 4(\alpha+\beta\epsilon) + \sqrt{(d^2 - 4(\alpha+\beta\epsilon))^2 + 16d^2\beta\epsilon}\right), \tag{4.9}$$

with an associated eigenvector

$$\boldsymbol{v}_\ell = \left[1, \frac{d^2+4\alpha+4\beta\epsilon-\sqrt{(d^2+4\alpha+4\beta\epsilon)^2-16\alpha d^2}}{4d}\right]. \tag{4.10}$$

The expansion of λ_ℓ for small ϵ is

$$\lambda_\ell = \frac{1}{4\epsilon d}\left((d^2-4\alpha)+|d^2-4\alpha|\right)+O(1). \tag{4.11}$$

Thus,

$$\alpha > \frac{d^2}{4} \quad \text{implies} \quad \lambda_\ell = O(1) \quad \text{as } \epsilon \to 0, \tag{4.12}$$

and

$$\alpha < \frac{d^2}{4} \quad \text{implies} \quad \lambda_\ell = \frac{1}{\epsilon}\left(\frac{d^2-4\alpha}{2d}\right)+O(1) \quad \text{as } \epsilon \to 0. \tag{4.13}$$

This transition occurs precisely at the critical α value in (3.10), which determines the existence of traveling waves of the original system ($\epsilon = 0$). Thus the onset of the solutions containing shocks corresponds to the introduction of a fast dynamic along the unstable manifold of $(u_\ell, 0)$ as $\epsilon \to 0$. Our motivation for introducing the viscous regularization was to understand the behavior of the wave solutions in the limit of $\epsilon \to 0$. Accordingly, we now focus on the case $0 < \alpha < d^2/4$, the range for which classical traveling waves of the original system ($\epsilon = 0$) fail to exist. There are two cases, depending on whether $0 < \alpha < d^2/8$ or $d^2/8 < \alpha < d^2/4$.

4.1. Case 1: $d^2/8 < \alpha < d^2/4$. The U-nullcline is the parabola given by (4.7). There are two distinct nullclines for S which correspond to the solutions of

$$(S+\alpha)\left(\frac{1}{2}(U-u_\ell)(U-u_r)-S\right)-\epsilon\beta S = 0. \tag{4.14}$$

To plot the S-nullclines, we arrange (4.14) to

$$(U-c)^2 = 2S+\frac{d^2}{4}+\frac{2\beta S}{\alpha+S}\epsilon. \tag{4.15}$$

When $\epsilon = 0$, the curve

$$(U-c)^2 = 2S+\frac{d^2}{4} \tag{4.16}$$

is identical to the U-nullcline given by (4.7).

One S-nullcline is located above (in the U-S plane) the horizontal line $S = -\alpha$ and the other below this line. For $-\alpha < S < 0$, the last term in (4.15), $2\beta S/(\alpha+S)$, is always negative. This decreases U^2, meaning that there is an S-nullcline just above the U-nullcline (just below for $S > 0$). As $\epsilon \to 0$, this S-nullcline converges to the U-nullcline.

On the second S-nullcline, $S < -\alpha$. In this region, the last term in (4.15) is always positive, and for S close to $-\alpha$ it dominates the linear term. The minimum value of S on the U-nullcline is $-d^2/8$. Since $\alpha > d^2/8$, this second S-nullcline is

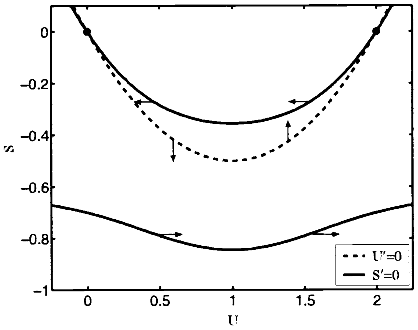


FIG. 4.2. *Typical nullclines for system* (4.5)–(4.6) *with* $d^2/8 < \alpha < d^2/4$. *Here the parameter values are* $u_\ell = 2$, $u_r = 0, \beta = 1, \alpha = 0.6, \epsilon = 0.1$.

below the U-nullcline and bounded away from it as $\epsilon \to 0$. A sample plot of all three nullclines is shown in Figure 4.2.

To find a traveling wave solution, we show that the unstable manifold of $(u_\ell, 0)$ flows to the line $U = c$. The eigenvector $\boldsymbol{v}_\ell$ from (4.10) is tangent to the unstable manifold at $(u_\ell, 0)$. Expanding this eigenvector for small ϵ yields

$$\boldsymbol{v}_\ell = \left[1, \frac{2\alpha}{d}\right] + \epsilon\left[0, \frac{-8\alpha\beta}{d(d^2 - 4\alpha)}\right] + O(\epsilon^2). \tag{4.17}$$

Thus, in the limit as $\epsilon \to 0$ the eigenpair $(\lambda_\ell, \boldsymbol{v}_\ell) \to (\infty, [1, 2\alpha/d])$. The slope of the U-nullcline at $(u_\ell, 0)$ is $d/2$ (independent of ϵ), and the slope of the S-nullcline at $(u_\ell, 0)$ is $d/2(1 + \beta\epsilon/\alpha)^{-1} = d/2(1 - \beta\epsilon/\alpha) + O(\epsilon^2)$. Thus, for $\epsilon < \alpha/\beta$, the unstable manifold enters the region above both the S- and U-nullclines whenever $\alpha < d^2/4$. Moreover, as $\epsilon \to 0$ the speed with which it enters the region approaches infinity.

The trajectories of the system (4.5)–(4.6) satisfy

$$\frac{dS}{dU} = \frac{(S + \alpha)F(U, S) - \epsilon\beta S}{(U - c)F(U, S)}, \tag{4.18}$$

where $F(U, S) = \frac{1}{2}(U - u_\ell)(U - u_r) - S$. Note that $F(U, S) = 0$ defines the U-nullcline and is the leading order approximation of the S-nullcline above it. The unstable manifold quickly flows away from these nullclines into the region where $F(U, S) = O(1)$. In this case, the curves defined by (4.18) are approximated by

$$\frac{dS}{dU} = \frac{(S + \alpha)}{(U - c)}. \tag{4.19}$$

The solutions of (4.19) are lines of the form $|S + \alpha| = A(U - c)$. The solution passing through the equilibrium $(u_\ell, 0)$ has slope $A = 2\alpha/d$, which is precisely the slope of the unstable manifold as $\epsilon \to 0$. Therefore the leading order approximation to the unstable manifold is

$$S = \frac{2\alpha}{d}(U - c) - \alpha, \tag{4.20}$$

which is a valid approximation as long as this trajectory remains away from the nullclines. The line (4.20) eventually intersects the S-nullcline. To leading order, this intersection occurs at

$$U^* = \frac{4\alpha}{d} + u_r, \tag{4.21}$$

$$S^* = \frac{2\alpha}{d^2}\left(4\alpha - d^2\right). \tag{4.22}$$

Since $\alpha > d^2/8$, it follows that $U^* > c$ at the point of intersection. Near the nullclines, the solution to the system (4.5)–(4.6) can be approximated by the quasi-steady solution

$$S = \frac{1}{2}(U - u_\ell)(U - u_r) + O(\epsilon). \tag{4.23}$$

This trajectory intersects the line $U = c$. Thus, by the symmetry of the system, this solution is part of a heteroclinic orbit connecting the points $(u_\ell, 0)$ and $(u_r, 0)$ and corresponds to a traveling wave solution of (4.1)–(4.2).

The above analysis explains the double-shock solution. When $\alpha < d^2/4$, the dynamics near the point $(u_\ell, 0)$ on the unstable manifold are very fast $(O\left(\epsilon^{-1}\right))$. Leaving the equilibrium point, the unstable manifold moves away from the nullclines, but eventually this trajectory approaches the nullclines near the point (U^*, S^*) away from the equilibrium point. This path in phase space (in the limit $\epsilon \to 0$) corresponds to the shock. Once near the nullclines, the solution flows along the nullclines to the line $U = c$. The flow between the point (U^*, S^*) and its reflected point $(2c - U^*, S^*)$ corresponds to the smooth portion of the double-shock solution between the two shocks. Figure 4.3(a) shows the path of the heteroclinic orbit in phase space corresponding to a double-shock solution. The path shown was generated by integrating (4.5)–(4.6) for $\epsilon = 10^{-3}$. The trajectory is very close to our asymptotic solution, which is not shown because it is indistinguishable from the numerical solution on this scale. In Figure 4.3(b) we show the wave profile for decreasing values of ϵ. The solutions were generated by integrating (4.5)–(4.6) for $U > c$ and using the symmetry condition for $U < c$. For finite ϵ the wave is smooth, but, as the figure indicates, the profile approaches the double-shock solution as $\epsilon \to 0$.

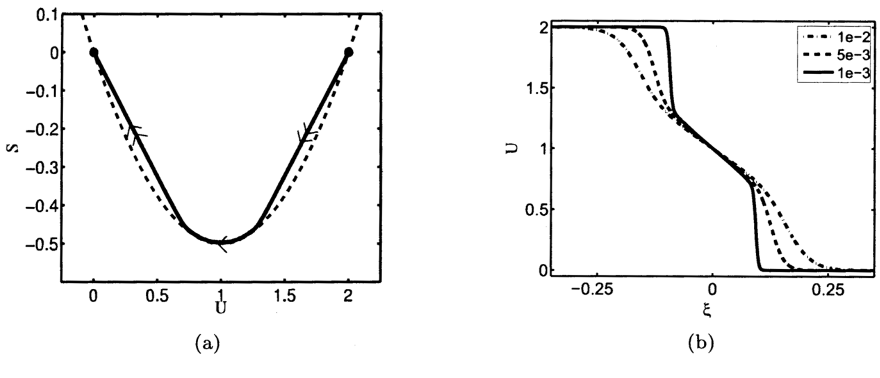


FIG. 4.3. (a) *Path of the heteroclinic orbit for the double-shock traveling wave. The double arrows indicate that the dynamics are much faster along these paths, which correspond to the shocks in the limit $\epsilon \to 0$. The trajectory shown is for $\epsilon = 10^{-3}$, generated by integrating* (4.5)–(4.6). *The solution is indistinguishable from the asymptotic solution on the scale shown.* (b) *For finite ϵ, the wave profile is smooth, but as $\epsilon \to 0$ the solution approaches the double-shock wave. The parameter values are $u_\ell = 2$, $u_r = 0, \beta = 1, \alpha = 0.65$.*

The height of each of the shocks in the double-shock solution is given by

$$[u] = u_\ell - U^* = \frac{d^2 - 4\alpha}{d}. \tag{4.24}$$

Below $\alpha = d^2/4$ smooth traveling waves no longer exist, and at this value of α the shock height is zero. As α decreases from this value, the height of the shocks increases. When $\alpha = d^2/8$, the height of each shock is $d/2$ so that the two shocks come together, and the double-shock solution as analyzed in this section no longer exists. What happens below this value of α is considered in the next section.

4.2. Case 2: $0 < \alpha < d^2/8$. Much of the analysis from the previous section applies to this case. However, one exception is that the S-nullcline above the U-nullcline no longer converges to the U-nullcline as $\epsilon \to 0$. As before, one of the S-nullclines is located above the line $S = -\alpha$ and the other below. Recall that the U-nullcline is the parabola (4.7), and the minimum value of S on this nullcline is $-d^2/8$. When $\alpha < d^2/8$, the line $S = -\alpha$ intersects the U-nullcline, so that as $\epsilon \to 0$ the S-nullcline above the U-nullcline remains bounded away from the U-nullcline for a range of U values. This S-nullcline limits to

$$S = \begin{cases} \frac{1}{2}(U - u_\ell)(U - u_r) & (U - c)^2 > \frac{d^2 - 4\alpha}{8}, \\ -\alpha & (U - c)^2 \le \frac{d^2 - 4\alpha}{8}. \end{cases} \tag{4.25}$$

A sample plot of the nullclines is shown in Figure 4.4 for small ϵ.

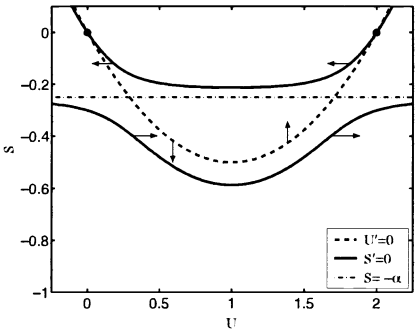


FIG. 4.4. *Typical nullclines for system* (4.5)–(4.6) *with* $0 < \alpha < d^2/8$. *Here the parameter values are* $u_\ell = 2$, $u_r = 0, \beta = 1, \alpha = 0.25, \epsilon = 0.05$.

As before, the unstable manifold of $(u_\ell, 0)$ flows into the region above the S-nullcline, and once the trajectory is $O(\epsilon)$ away from the equilibrium point the dynamics are fast $(O(\epsilon^{-1}))$. This unstable manifold is again approximated by the line (4.20). The unstable manifold eventually brings the flow back to the S-nullcline (4.25). These two curves intersect at the point $(U, S) = (c, -\alpha)$, and by symmetry the stable manifold of $(u_r, 0)$ also flows from this point. Thus the solution does not travel along the S-nullcline at all because the region of fast dynamics leaving $(u_\ell, 0)$ connects with the region of fast dynamics entering $(u_r, 0)$. Figure 4.5(a) shows the path of the heteroclinic orbit connecting $(u_\ell, 0)$ and $(u_r, 0)$ corresponding to the single-shock traveling wave. This solution was generated by integrating (4.5)–(4.6) for $\epsilon = 0.02$. The asymptotic solution is indistinguishable from the numerical solution on this scale. Figure 4.5(b) shows the wave profile for decreasing values of ϵ. For finite ϵ the wave profile is smooth, but it approaches a single shock as $\epsilon \to 0$.

The numerical simulations from section 3.3 suggested that for $\alpha < d^2/8$ the traveling wave solution was the shock solution from the inviscid Burgers equation. This analysis confirms this but provides more information on the structure of this shock for small viscosity. This shock is really a degenerate double-shock solution, in that the two shocks meet in the middle of the wave profile.

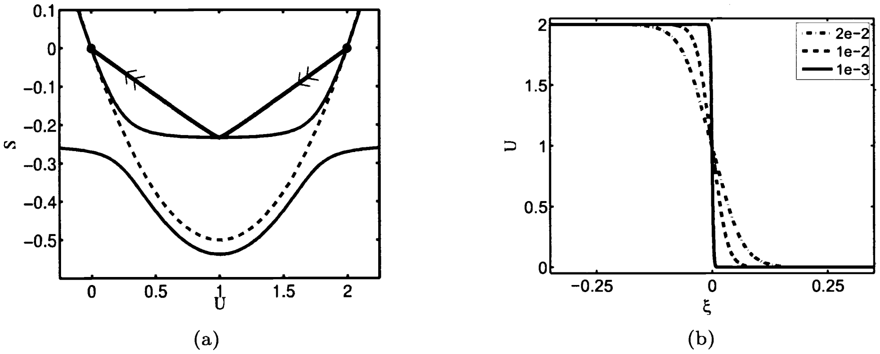


FIG. 4.5. (a) *Path of the heteroclinic orbit for the single-shock traveling wave that occurs when* $\alpha < d^2/8$. *The double arrows indicate the fast dynamics along these paths which correspond to the shocks in the limit* $\epsilon \to 0$. *This solution was generated by integrating* (4.5)–(4.6) *for* $\epsilon = 0.02$. *On this scale the asymptotic solution is indistinguishable from the numerical solution.* (b) *For finite* ϵ *the wave profile is smooth, but as* $\epsilon \to 0$ *the solution approaches the single-shock wave. The parameter values in both plots are* $u_\ell = 2$, $u_r = 0, \beta = 1, \alpha = 0.25$.

5. Discussion.

For given asymptotic values of the velocity, u_ℓ and u_r with $u_\ell > u_r$, the viscoelastic Burgers model (1.3)–(1.4) has three different types of traveling wave solutions, depending on the value of the elastic modulus α. For $\alpha > d^2/4$, smooth traveling waves exist, where $d = u_\ell - u_r$. When $d^2/8 < \alpha < d^2/4$, the profile of the traveling wave is piecewise smooth with two jump discontinuities, and when $\alpha < d^2/8$ the wave solution is a single shock. In all three cases the wave travels with unique speed $c = (u_\ell + u_r)/2$.

We address the physical significance of the threshold in the elastic modulus α for traveling waves to exist. For simplicity, consider the case in which $u_\ell = -u_r$, so that the speed of the traveling wave is 0. The condition $\alpha > d^2/4$ for a wave to exist reduces to $\alpha > u_\ell^2$, or $\sqrt{\alpha} > u_\ell$. The system linearized about $u = u_\ell$, $\sigma = 0$ is

$$u_t + u_\ell u_x = \sigma_x, \tag{5.1}$$

$$\sigma_t + u_\ell \sigma_x = \alpha u_x - \beta\sigma, \tag{5.2}$$

which can be written in the form

$$\boldsymbol{q}_t + A\boldsymbol{q}_x = B\boldsymbol{q}, \tag{5.3}$$

where $\boldsymbol{q} = (u, \sigma)^{\mathrm{T}}$. The wave speeds of this linearized system are $u_\ell \pm \sqrt{\alpha}$. The wave speeds are the sum of the advective speed u_ℓ and the elastic wave speeds $\pm\sqrt{\alpha}$. The advection terms tend to steepen the wave, which generates elastic forces that oppose this steepening. As long as the elastic wave speed is faster than the advective wave speed, smooth traveling waves exist. In the viscous Burgers equation ($\sigma = \epsilon u_x$), the viscous stresses propagate instantaneously, but in the viscoelastic model the elastic stresses propagate at a finite speed. Thus, the smooth traveling wave breaks down when the advective speed surpasses the elastic speed.

Recall that $\alpha = \mu/\lambda$, where μ and λ are the viscosity and relaxation time, respectively. For a fixed relaxation time, each of the three types of wave solutions is possible, depending on the size of the viscosity. For large enough viscosity, the smooth traveling wave results, and as the viscosity is decreased the solution transitions to the double-shock wave and then to the single-shock wave. Equivalently, for a fixed viscosity, the type of wave depends on the size of the relaxation time. The progression from the smooth wave to the double-shock wave to the single-shock wave occurs as the relaxation time increases. The regions of parameter space where the different wave solutions occur are illustrated in Figure 5.1.

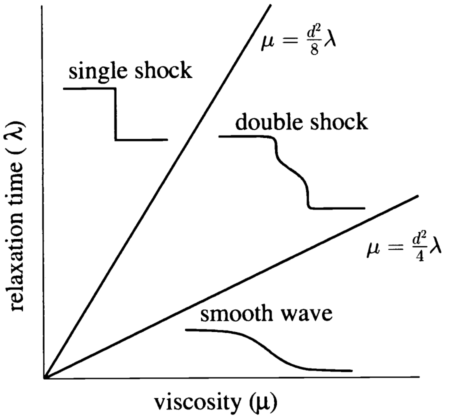


FIG. 5.1. *The values of the relaxation time λ and the viscosity μ determine the type of traveling wave solution. In parameter space the line $\mu = d^2\lambda/4$ is the boundary between smooth waves and double-shock solutions, and the line $\mu = d^2\lambda/8$ is the boundary between double shocks and single shocks.*

For $\lambda = 0$, only the smooth traveling wave is possible. In the limit that $\lambda \to 0$ for fixed μ, the constitutive law reduces $\sigma = \mu u_x$, and the model becomes Burgers' equation (1.1). This limit corresponds to the constitutive law for a viscous fluid. Taking the limit $\beta \to 0$ for a fixed value of α, the constitutive law limits to that of an elastic solid. The transitions between the different wave types are independent of the value of β. If we nondimensionalize the problem, the value of $\beta^{-1} = \lambda$ determines the time scale of the problem, which is related to the steepness of the wave profiles. As β gets smaller, the wave profiles steepen, meaning that as $\beta \to 0$ all wave solutions tend to shocks.

In section 2, we presented several different one-dimensional reductions of the UCM equation and in the remainder of the paper presented an analysis based on (2.10). However, the techniques employed apply to all three constitutive laws. Repeating the analysis for (2.11), we find that again there are smooth traveling waves for $\alpha > d^2/4$, but, for $\alpha < d^2/4$, only the single-shock solutions occur. For (2.12) there is a transition from a smooth traveling wave to a double-shock solution at $\alpha = c^2$, and the single-shock solution is approached as $\alpha \to 0$. Because (2.10) exhibits all three behaviors, we chose to present this case.

There are many different constitutive laws for viscoelastic fluids. In this paper we used the UCM model (Oldroyd-B when $\epsilon \neq 0$) because it is perhaps the simplest differential constitutive law and has been extensively studied in the past. Others have studied viscoelastic generalization of Burgers' equation [13, 21], and it would be interesting to explore how the behavior of the wave solutions analyzed in this paper is affected by different constitutive laws.

The problem in this paper is interesting in part because of its classical nature, but the analysis of one-dimensional waves in viscoelastic generalizations of Burgers' equations could also be used to develop numerical schemes for viscoelastic fluids. High-resolution finite volume methods have been used successfully in simulating high Reynolds number flows [1]. The algorithm for discretizing the convection terms in [1] is based on numerical methods for conservation laws [6]. These methods require solving one-dimensional Riemann problems, and it is not clear how to adapt this approach to nonconservative systems. Wave propagation algorithms [15] are more easily adapted to nonconservative problems, but these methods also require being able to solve one-dimensional Riemann problems. Recently finite volume methods for viscoelastic flows have been proposed [10, 22]. The techniques from this paper could be adapted to solve the Riemann problems that arise in these methods.

Acknowledgments. The authors would like to thank Andy Bernoff and Jim Keener for helpful discussions while writing this paper. We also thank the referees for their insightful comments.

REFERENCES

[1] J. B. Bell, P. Colella, and H. M. Glaz, *A second-order projection method for the incompressible Navier–Stokes equations*, J. Comput. Phys., 85 (1989), pp. 257–283.

[2] H. A. Biagioni, *A Nonlinear Theory of Generalized Functions*, 2nd ed., Lecture Notes in Math. 1421, Springer-Verlag, Berlin, 1990.

[3] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, 2nd ed., Vol. 1, Wiley, New York, 1987.

[4] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, 2nd ed., Vol. 2, Wiley, New York, 1987.

[5] J. M. Burgers, *A mathematical model illustrating the theory of turbulence*, in Advances in Applied Mechanics, Academic Press, New York, 1948, pp. 171–199.

[6] P. Colella, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200.

[7] J.-F. Colombeau, *New Generalized Functions and Multiplication of Distributions*, North–Holland Math. Stud. 84, North–Holland, Amsterdam, 1984.

[8] J.-F. Colombeau, *Multiplication of Distributions*, Lecture Notes in Math. 1532, Springer-Verlag, Berlin, 1992.

[9] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[10] R. D. Guy and A. L. Fogelson, *A wave propagation algorithm for viscoelastic fluids with spatially and temporally varying properties*, Comput. Methods Appl. Mech. Engrg., to appear.

[11] D. D. Joseph, *Fluid Dynamics of Viscoelastic Liquids*, Springer-Verlag, New York, 1990.

[12] R. G. Larson, *Constitutive Equations for Polymer Melts and Solutions*, Butterworth, Stoneham, MA, 1988.

[13] D. G. Lasseigne and W. E. Olmstead, *Stability of a viscoelastic Burgers flow*, SIAM J. Appl. Math., 50 (1990), pp. 352–360.

[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 1992.

[15] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[16] R. J. LeVeque, *Finite-volume methods for nonlinear elasticity in heterogeneous media*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 93–104.

[17] M. J. Lighthill and G. B. Whitham, *On kinematic waves.* II. *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[18] J. D. Logan, *An Introduction to Nonlinear Partial Differential Equations*, Pure Appl. Math. (N.Y.), John Wiley and Sons, New York, 1994.

[19] P. I. Richards, *Shock waves on the highway*, Operations Res., 4 (1956), pp. 42–51.

[20] J. J. Stoker, *Water Waves. The Mathematical Theory with Applications*, John Wiley and Sons, New York, 1992.

[21] N. Sugimoto and T. Kakutani, *"Generalized Burgers equation" for nonlinear viscoelastic waves*, Wave Motion, 7 (1985), pp. 447–458.

[22] D. Trebotich, P. Colella, G. H. Miller, A. Nonaka, T. Marshall, S. Gulati, and D. Liepmann, *A numerical algorithm for complex biological flow in irregular microdevice geometries*, in Technical Proceedings of the 2004 Nanotechnology Conference and Trade Show, Vol. 2, Nano Science and Technology Institute, Cambridge, MA, 2004, pp. 470–473.

[23] G. B. Whitham, *Linear and Nonlinear Waves*, Pure Appl. Math. (N.Y.), John Wiley and Sons, New York, 1999.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1333–1359

THE CHAPMAN–JOUGUET CLOSURE FOR THE RIEMANN PROBLEM WITH VAPORIZATION*

VINCENT PERRIER[†]

Abstract. This work is devoted to the modeling of phase transition. The thermodynamic model for phase transition chosen is a model with two equations of state, each of them modeling one phase of a given fluid. The mixture equation of state is obtained by an entropy optimization criterion. Both equations of state are supposed to be convex, and a necessary condition is found to ensure the convexity of the mixture equation of state. Then we investigate the Riemann problem for the Euler system with these equations of state. More precisely, we propose to take into account metastable states, which may occur as noted in [J. R. Simões-Moreira and J. E. Shepherd, *J. Fluid Mech.*, 382 (1999), pp. 63–86]. We check whether the Chapman–Jouguet theory can be applied in our context, and that it is consistent with the entropy growth criterion. As the characteristic Lax criterion does not hold for this solution, an additional relation, the kinetic closure, is necessary. The common closure, i.e., the Chapman–Jouguet closure, is proved to be incorrect in general in that context.

Key words. fluid mechanics, phase transition, nonconvex equation of state, Riemann problem, Chapman–Jouguet theory

AMS subject classifications. 76N99, 80A22, 80A20, 76T10, 82B26, 35L65

DOI. 10.1137/060650374

Introduction. We are interested in the study of some problems arising in the modeling of phase transition in compressible fluids. A widely used model for phase transition [2, 1, 14, 5] is the van der Waals model. Nevertheless, the very physical meaning of this model is questionable, because the resulting system of partial differential equations is not hyperbolic. Moreover, according to [16], the shock structure found in [14] does not seem to match with what experiments show. Another approach consists in modeling each phase by an equation of state, and in coupling them by optimizing the entropy [7, 9] to get a mixture equation of state. As explained in [12], the convexity of internal energy is necessary to ensure the local thermodynamic equilibrium. Whether the mixture equation of state is convex or not will be discussed in section 1.

We will then concentrate on the Euler system of partial differential equations, which models flow dynamics with neither viscosity nor thermal conduction. A fundamental step to approximate the solutions of the Euler system with a Godunov method is to solve the Riemann problem, i.e., the Cauchy problem where the initial condition is composed of two different constant states. If the solution of the Riemann problem for the Euler system is easy to solve for a convex equation of state, it becomes much harder when the equation of state suffers from loss of derivative and from local nonconvexity [19, 12, 20, 21, 16, 17, 18, 11], because the common entropy growth criterion fails to ensure the existence and uniqueness of the solution. Based on the experiments of [13], we propose to take into account out-of-thermodynamic-equilibrium states, or *metastable* states. The Chapman–Jouguet (CJ) theory will be used [4, 6], and the compatibility between the model of equation of state and this theory will be discussed.

*Received by the editors January 20, 2006; accepted for publication (in revised form) July 25, 2007; published electronically April 23, 2008. This work was partially supported by the CEA through contract 123-C-BEFI.

http://www.siam.org/journals/siap/68-5/65037.html

[†]Mathématiques Appliquées de Bordeaux and Centre des Lasers Intenses et Applications, Université Bordeaux 1, 33405 Talence Cedex, France (perrier@math.u-bordeaux1.fr).

Table 1
Thermodynamic notation.

T	temperature
τ	specific volume
ρ	density
s	specific entropy
μ	chemical potential
h	specific enthalpy

f	specific free energy
P	pressure
ε	specific internal energy
y	mass fraction
α	volume fraction

The entropy growth criterion will be shown to hold. Finally, we will show that the closure usually used to close the problem [10], i.e., the CJ closure, leads to a solution that does not in general continuously depend on its initial data, and this solution will therefore be rejected.

1. Thermodynamic preliminaries.

1.1. Thermodynamic with phase transition. We suppose that we have two phases of the same fluid. For the sake of simplicity, we will consider that one phase is liquid (subscript l) and the other one is the vapor (subscript v). For each phase, we will take the notation of Table 1. If we consider that we have a mixture of the two phases, then the total specific quantities are defined by

$$\tau_{tot} = y_l \tau_l + y_v \tau_v, \tag{1.1a}$$

$$\varepsilon_{tot} = y_l \varepsilon_l + y_v \varepsilon_v, \tag{1.1b}$$

$$s_{tot} = y_l s_l + y_v s_v. \tag{1.1c}$$

To find the thermodynamic equilibrium, the total entropy (1.1c) must be optimized. Of course, the optimization must be consistent with the following constraints:

$$\text{conservation of total energy} \qquad y_l \varepsilon_l + y_v \varepsilon_v = cste, \tag{1.2a}$$

$$\text{conservation of mass} \qquad y_l + y_v = 1. \tag{1.2b}$$

Moreover, we suppose that the two phases are locally nonmiscible, which means that

$$y_l \tau_l + y_v \tau_v = cste. \tag{1.2c}$$

We choose to optimize (1.1c) with the variables τ, ε, y for each phase. The first and second principles of thermodynamics impose that for each phase

$$\mathrm{d}s = \frac{\mathrm{d}\varepsilon}{T} + \frac{P}{T}\,\mathrm{d}\tau.$$

Then the differential of s_{tot} must belong to the set spanned by the gradients of the constraints; if we denote by $\lambda_1, \lambda_2, \lambda_3$ the Lagrange multipliers associated to the constraints (1.2a), (1.2b), (1.2c), we find (with the notation of Table 1)

$$s_l = \lambda_1 \varepsilon_l + \lambda_2 + \lambda_3 \tau_l, \tag{1.3a}$$

$$\frac{y_l}{T_l} = \lambda_1 y_l, \tag{1.3b}$$

$$\frac{y_l P_l}{T_l} = \lambda_3 y_l, \tag{1.3c}$$

$$s_v = \lambda_1 \varepsilon_v + \lambda_2 + \lambda_3 \tau_v, \tag{1.3d}$$

$$\frac{y_v}{T_v} = \lambda_1 y_v, \tag{1.3e}$$

$$\frac{y_v P_v}{T_v} = \lambda_3 y_v. \tag{1.3f}$$

If we suppose that both phases coexist, then (1.3b) and (1.3e) give

$$T_l = T_v =: T,$$

(1.3c) and (1.3f) give

$$P_l = P_v =: P,$$

and, finally, (1.3a) and (1.3d) lead to

$$\mu_l(P,T) = \mu_v(P,T). \tag{1.4}$$

It is well known that there exist a temperature T_c and a pressure P_c such that there is no longer any difference between the liquid and the gas phases above this temperature and pressure (the fluid is said to be *supercritical*). If we suppose that under these values we have

$$\tau_l(P,T) \neq \tau_v(P,T),$$

then we can apply the implicit function theorem to (1.4) to prove that P is locally a function of T:

$$P = P_{sat}(T).$$

The limit of thermodynamic stability of the mixture is given, in the (τ, P) plane, by the functions $\tau_v(P_{sat}(T),T)$ and $\tau_l(P_{sat}(T),T)$. The set of all the physical states lying between these curves is called the saturation dome. When there is no ambiguity, we will also denote by τ_v the function $T \mapsto \tau_v(P_{sat}(T),T)$ and by τ_l the function $T \mapsto \tau_l(P_{sat}(T),T)$. Thus, the thermodynamic plane can be divided into four parts (see Figure 1): two parts in which one of the two pure phases is stable, one part where a mixture of saturated liquid and vapor is stable and where the specific thermodynamic variables are defined as in (1.1), and one part with the supercritical fluid.

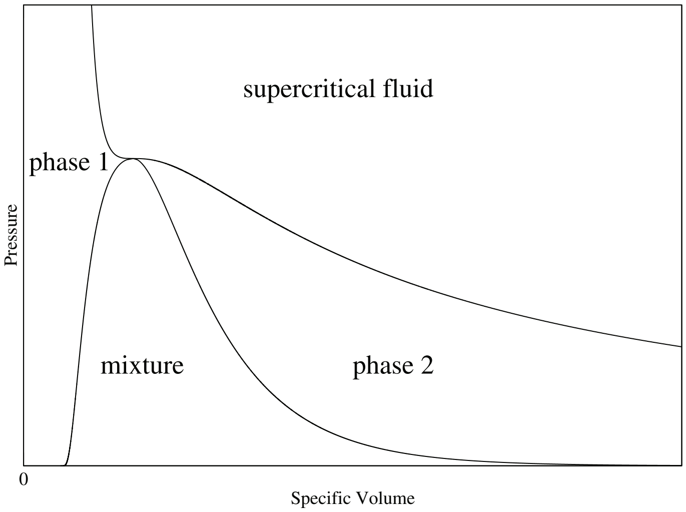


FIG. 1. *The thermodynamic plane (P, τ) is divided into four parts: two parts where the single phases are stable, one part where the mixture is stable, and one part where the fluid is supercritical.*

Moreover, if we differentiate the equality $\mu_v(P,T) = \mu_l(P,T)$ with respect to (P,T), we find the Clausius–Clapeyron relation

$$\frac{\mathrm{d}P_{sat}}{\mathrm{d}T}(T) = \frac{s_v(T) - s_l(T)}{\tau_v(T) - \tau_l(T)}. \tag{1.5}$$

For most of the solid-liquid phase transition (except for special cases such as bismuth or water) and for all of the liquid-gas phase transition, we have for all $T < T_c$

$$\tau_v(T) - \tau_l(T) > 0.$$

This means that in Figure 1, phase 1 is the liquid, and phase 2 is a gas.

For any phase transition, the entropy of the most compact constituent is lower than the entropy of the other constituent. Thus, in our case we have for all $T < T_c$

$$s_v(T) - s_l(T) > 0,$$

which induces $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0$.

1.2. Adimensioned thermodynamic coefficients. We adopted the notation of Table 1. As in [12], we define three adimensioned parameters as

$$\Gamma = -\frac{\tau}{T}\left(\frac{\partial T}{\partial \tau}\right)_s, \qquad \gamma = -\frac{\tau}{P}\left(\frac{\partial P}{\partial \tau}\right)_s, \qquad g = \frac{P\tau}{T^2}\left(\frac{\partial T}{\partial s}\right)_\tau. \tag{1.6}$$

γ is the adiabatic coefficient, and Γ is the Grüneisen coefficient. As in [12], we propose using these three thermodynamic coefficients to express all the thermodynamic quantities. One can then show that the following identities hold:

$$\mathrm{d}s = \frac{P\tau}{T^2}\,\frac{1}{g}\,\mathrm{d}T + \frac{P}{T}\,\frac{\Gamma}{g}\,\mathrm{d}\tau, \tag{1.7a}$$

$$\mathrm{d}s = -\frac{\tau}{T}\,\frac{\Gamma}{\gamma g - \Gamma^2}\,\mathrm{d}P + \frac{P\tau}{T^2}\,\frac{\gamma}{\gamma g - \Gamma^2}\,\mathrm{d}T, \tag{1.7b}$$

$$\mathrm{d}s = \frac{\tau}{T}\,\frac{1}{\Gamma}\,\mathrm{d}P + \frac{P}{T}\,\frac{\gamma}{\Gamma}\,\mathrm{d}\tau, \tag{1.7c}$$

$$\mathrm{d}\tau = -\frac{\tau}{P}\,\frac{g}{\gamma g - \Gamma^2}\,\mathrm{d}P + \frac{\tau}{T}\,\frac{\Gamma}{\gamma g - \Gamma^2}\,\mathrm{d}T, \tag{1.7d}$$

$$\mathrm{d}h = \tau\,\frac{\Gamma + 1}{\Gamma}\,\mathrm{d}P + P\,\frac{\gamma}{\Gamma}\mathrm{d}\tau, \tag{1.7e}$$

$$\mathrm{d}\varepsilon = \tau\,\frac{1}{\Gamma}\,\mathrm{d}P + P\,\frac{(\gamma - \Gamma)}{\Gamma}\,\mathrm{d}\tau. \tag{1.7f}$$

Thermodynamic stability requires that ε be convex as a function of τ and s, which leads to

$$g \geq 0, \qquad \gamma \geq 0, \qquad \gamma g - \Gamma^2 \geq 0. \tag{1.8}$$

1.3. The fundamental derivative. In [15], the fundamental derivative was defined as

$$\mathscr{G} = -\frac{\tau}{2}\,\frac{\left(\frac{\partial^3 \varepsilon}{\partial \tau^3}\right)_s}{\left(\frac{\partial^2 \varepsilon}{\partial \tau^2}\right)_s}.$$

The sign of $\mathscr{G}$ determines whether the Hugoniot curve and the isentropes are convex or not in the (τ, P) plane. We will suppose in the following that $\mathscr{G}$ is positive, so that no undercompressive discontinuity or expansion fans can exist (see [15, 20, 21]).

1.4. Mixture equation of state.

1.4.1. Parameterization of the mixture equation of state. In the following, we will denote with a subscript m all the variables relative to the mixture equation of state. The mixture equation of state is naturally parameterized by y, the mass fraction of the gas, and T, the temperature. Nevertheless, in the next sections, the parameters that will intervene are mostly τ and s. They are linked by the transformation

$$\Phi : \begin{pmatrix} y \\ T \end{pmatrix} \mapsto \begin{pmatrix} y\tau_v(T) + (1-y)\tau_l(T) \\ ys_v(T) + (1-y)s_l(T) \end{pmatrix} = \begin{pmatrix} \tau \\ s \end{pmatrix}. \tag{1.9}$$

Theorem 1. *For all points in the saturation dome, Φ is a local diffeomorphism provided that the equations of state of the liquid and of the gas are both convex.*

Proof. To show that Φ is a local diffeomorphism, it is sufficient to show that its Jacobian does not vanish. Differentiation of (1.9) and using the Clausius–Clapeyron relation lead to

$$\det(\mathrm{D}\Phi) = (\tau_v - \tau_l)\left(y\left(\frac{\mathrm{d}s_v}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_v}{\mathrm{d}T}\right) + (1-y)\left(\frac{\mathrm{d}s_l}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_l}{\mathrm{d}T}\right)\right). \tag{1.10}$$

We supposed that $\tau_v - \tau_l > 0$ (except at the critical point), so it remains to show that

$$y\left(\frac{\mathrm{d}s_v}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_v}{\mathrm{d}T}\right) + (1-y)\left(\frac{\mathrm{d}s_l}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_l}{\mathrm{d}T}\right) \tag{1.11}$$

never vanishes. The term (1.11) is a convex combination of

$$\frac{\mathrm{d}s_v}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_v}{\mathrm{d}T} \quad \text{and} \quad \frac{\mathrm{d}s_l}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_l}{\mathrm{d}T}. \tag{1.12}$$

Using (1.3a), (1.3b) for $P = P_{sat}(T)$ leads to

$$\frac{\mathrm{d}s_b}{\mathrm{d}T} - \frac{\mathrm{d}P_{sat}}{\mathrm{d}T}\frac{\mathrm{d}\tau_b}{\mathrm{d}T} = \frac{\gamma g - \Gamma^2}{g}\frac{P}{\tau}\left(\frac{\mathrm{d}\tau_b}{\mathrm{d}T}\right)^2 + \frac{P\tau}{T^2}\frac{1}{g} > 0 \tag{1.13}$$

for $b = g$ or l, which is positive provided that each pure phase equation of state is convex.

Therefore (1.11) is positive, because it is a convex combination of two terms as is (1.13). As a consequence, $\det(\mathrm{D}\Phi) > 0$. □

Thanks to the parameterization (1.9), we can calculate the adimensioned coefficients defined by (1.6), to show that the following theorem holds.

Theorem 2. *If both equations of state are convex, and if $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0$, then the mixture equation of state is convex, too; i.e., inequalities* (1.8) *hold.*

Proof. We denote by a subscript m the thermodynamic parameters relative to the mixture equation of state.

- *Calculation of* Γ_m. To calculate Γ_m, we first use the chain rule

$$\left(\frac{\partial\tau}{\partial T}\right)_s = \left(\frac{\partial y}{\partial T}\right)_s\left(\frac{\partial\tau}{\partial y}\right)_T + \left(\frac{\partial\tau}{\partial T}\right)_y.$$

Then the differentiation of the definition of mixture entropy shows that

$$\left(\frac{\partial y}{\partial T}\right)_s = \frac{y\frac{\mathrm{d}s_v}{\mathrm{d}T} + (1-y)\frac{\mathrm{d}s_l}{\mathrm{d}T}}{s_l - s_v},$$

which leads to

$$\left(\frac{\partial\tau}{\partial T}\right)_s = -\frac{y\frac{\mathrm{d}s_v}{\mathrm{d}T}+(1-y)\frac{\mathrm{d}s_l}{\mathrm{d}T}}{s_l-s_v}(\tau_l-\tau_v)+y\frac{\mathrm{d}\tau_v}{\mathrm{d}T}+(1-y)\frac{\mathrm{d}\tau_l}{\mathrm{d}T}.$$

Thanks to the Clausius–Clapeyron relation, we find

$$\left(\frac{\partial\tau}{\partial T}\right)_s = y\left(\frac{\mathrm{d}\tau_v}{\mathrm{d}T}-\frac{\mathrm{d}T}{\mathrm{d}P}\frac{\mathrm{d}s_v}{\mathrm{d}T}\right)+(1-y)\left(\frac{\mathrm{d}\tau_v}{\mathrm{d}T}-\frac{\mathrm{d}T}{\mathrm{d}P}\frac{\mathrm{d}s_l}{\mathrm{d}T}\right),$$

which is negative according to what we did for the Jacobian of Φ. Therefore

$$\Gamma_m = -\frac{T}{\tau}\left(\frac{\partial\tau}{\partial T}\right)_s \geq 0.$$

- *Calculation of* γ_m. As $P = P_{sat}(T)$ in the saturation area, we have

$$\left(\frac{\partial\tau}{\partial P}\right)_s = \frac{\mathrm{d}T_{sat}}{\mathrm{d}P}\left(\frac{\partial\tau}{\partial T}\right)_s.$$

Thus $\gamma_m = \Gamma_m\,\frac{T}{P}\,(\frac{\mathrm{d}P}{\mathrm{d}T})_{sat} \geq 0$.

- *Calculation of* g_m. By using the identity

$$\left(\frac{\partial T}{\partial s}\right)_\tau\left(\frac{\partial s}{\partial\tau}\right)_T\left(\frac{\partial\tau}{\partial T}\right)_s = -1,$$

we have

$$\left(\frac{\partial T}{\partial s}\right)_\tau = -\frac{1}{(\frac{\partial s}{\partial\tau})_T\,(\frac{\partial\tau}{\partial T})_s}.$$

Along an isotherm we have $\mathrm{d}s = (s_v - s_l)\mathrm{d}y$ and $\mathrm{d}\tau = (\tau_v - \tau_l)\mathrm{d}y$, so that

$$\left(\frac{\partial s}{\partial\tau}\right)_T = \left(\frac{\mathrm{d}P}{\mathrm{d}T}\right)_{sat}.$$

Therefore, we find

$$\left(\frac{\partial T}{\partial s}\right)_\tau = \left(\frac{\mathrm{d}T}{\mathrm{d}P}\right)_{sat}\frac{\Gamma_m T}{\tau},$$

which induces $\gamma_m g_m = \Gamma_m^2$. As $\gamma_m \geq 0$, this means that $g_m \geq 0$.
Therefore, we proved that $g_m \geq 0$, $\gamma \geq 0$, and that $\gamma_m g_m - \Gamma_m^2 = 0$, so that the convexity of energy is ensured. □

1.5. Adimensioned coefficients near a phase transition boundary. In this section, we continue denoting by the subscript m the thermodynamic coefficients of the mixture equation of state, while the coefficients with no subscript are of the pure phase.

In [12, p. 121], the following identity is proved:

$$\frac{\gamma-\gamma_m}{\gamma_m} = (\gamma g - \Gamma^2)\left(\frac{T}{\tau}\left(\frac{\mathrm{d}s_b}{\mathrm{d}P}\right)_{sat}\right)^2 > 0,$$

with $b = g$ or l. This identity proves that isentropes are stiffer in the pure phases than in the mixture. In the same manner it is proved that

$$\frac{\Gamma_m}{\Gamma} = \frac{\gamma_m - \xi}{\gamma - \xi}, \tag{1.14}$$

with $\xi = -\frac{\tau}{P}(\frac{\mathrm{d}P}{\mathrm{d}\tau})_{sat}$. As in [12], we suppose that the isentropes can be parameterized by τ, so that

$$\frac{\gamma_m - \xi}{\gamma - \xi} > 0, \tag{1.15}$$

and so that Γ is positive too, because $\Gamma_m > 0$.

1.6. Retrograde and regular behavior. In [16], the retrogradicity r was introduced to study the behavior of isentropes near a phase transition boundary

$$r = \left(\frac{\partial T}{\partial \tau}\right)_P \left(\frac{\mathrm{d}s_b}{\mathrm{d}P}\right).$$

Thanks to (1.7d) and as Γ is positive near a phase transition boundary, $(\frac{\partial T}{\partial \tau})_P$ is positive, so that the sign of r is the same as the sign of $\frac{\mathrm{d}s_b}{\mathrm{d}P}$.

We suppose now that a fluid undergoes a rarefaction isentrope: this is the only regular transformation that a fluid can undergo. In the (S, T) plane, this transformation is drawn as a vertical line. As the transformation is undercompressive, the temperature decreases (at least near the phase transition boundary, because $\Gamma > 0$ and $\Gamma_m > 0$). If $r > 0$, then the isentrope crosses the saturation curve from the pure phase to the mixture phase (as on both sides of the left-hand part of Figure 2 and the liquid side of the right-hand part of Figure 2). In this case, the fluid is said to be *regular*. If r is negative, then the isentrope crosses the saturation curve from the mixture to the pure phase, as on the gas side of the right-hand part of Figure 2.

In [12, p. 121], other expressions of r are given:

$$r = \frac{\Gamma_m}{\Gamma}\frac{\gamma g - \Gamma^2}{\gamma_m}\frac{\mathrm{d}s_b}{\mathrm{d}T} = \frac{\gamma - \gamma_m}{\gamma_m}\frac{\xi}{\xi - \gamma}.$$

Experiments show that the liquid saturation curve is always regular. The gas saturation curve can be either regular or retrograde.

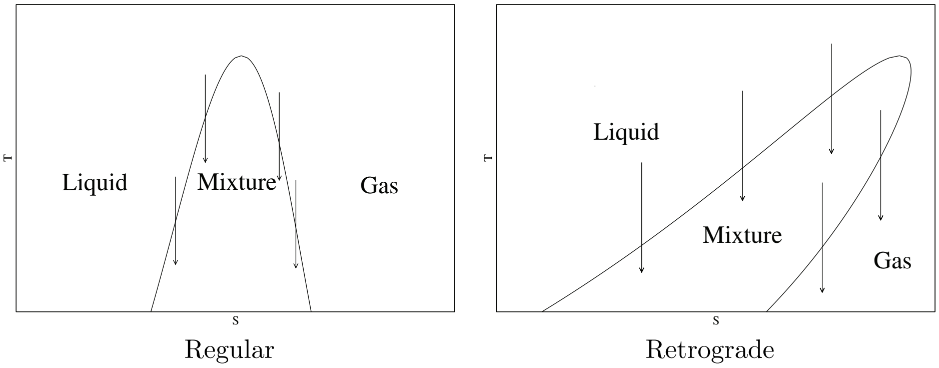


FIG. 2. *The saturation dome in the (S, T) plane. On the left, the fluid is regular: all the isentropes (drawn as arrows) cross the saturation dome from the pure phase to the mixture. On the right, the fluid is retrograde: the isentropes are crossing the liquid saturation curve from the pure phase to the mixture, whereas it is the contrary on the gas side.*

1.7. Validity domain of a model. To quickly compute a solution of the Riemann problem for fluid flows, simplified equations of state (perfect gas or stiffened gas, for example) are often preferred to tabulated ones. Nevertheless such equations of state often have only a narrow range of validity, outside of which they do not have a physical behavior (negative energy, nonconvexity).

If we want to use a simplified equation of state for both liquid and gas, we have to care not only about the physical behavior of the two equations of state but also about the mixture equation of state computed. If we look at the properties needed in section 1, we see that the property $\frac{\mathrm{d}P_{\mathrm{sat}}}{\mathrm{d}T} > 0$ is fundamental to ensuring the local convexity of energy. Nevertheless, it is not always true, as we show now with examples.

1.7.1. The two perfect gas model. This model was proposed by [8, 7]. The two phases are modeled with a perfect gas equation of state. To complete the equation of state, we suppose, moreover, that $C_v = 1$ for each fluid. We denote by $\overline{\Gamma}_i$ the Grüneisen coefficient of the phase i. Then we have

$$\varepsilon_i(P,\tau) = \frac{P\tau}{\overline{\Gamma}_i}, \tag{1.16a}$$

$$s_i(P,T) = \log\left(T\left(\frac{\overline{\Gamma}_i T}{P}\right)^{\overline{\Gamma}_i}\right), \tag{1.16b}$$

$$\mu_i(P,T) = (\overline{\Gamma}_i + 1)T - T\log\left(T\left(\frac{\overline{\Gamma}_i T}{P}\right)^{\overline{\Gamma}_i}\right). \tag{1.16c}$$

The equation $\mu_1(P,T) = \mu_2(P,T)$ can be explicitly solved to get $P = \beta T$, with $\beta = \exp(1)(\frac{\Gamma_2^{\Gamma_2}}{\Gamma_1^{\Gamma_1}})^{\frac{1}{\Gamma_1-\Gamma_2}}$. We see here that the condition $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0$ always holds. The limits of the saturation dome are given by the equations

$$T = \varepsilon = \frac{P_{sat}(T)\tau_i(T)}{\Gamma_i},$$

which gives $\tau_i(T) = \frac{\Gamma_i}{\beta}$. Thus, $T \mapsto \tau_i(T)$ is a constant function. In particular, the critical point does not exist. If we decide, for example, that $\Gamma_1 < \Gamma_2$, then we get the projections of the phase diagram in the (P,τ) plane and in the (S,T) plane that are drawn in Figure 3. The mixture equation of state can be explicitly calculated:

$$\begin{cases} P(\tau,\varepsilon) = \Gamma_2\dfrac{\varepsilon}{\tau} & \text{if} \quad \tau \leq \tau_2, \\ P(\tau,\varepsilon) = \Gamma_2\dfrac{\varepsilon}{\tau_2} = \Gamma_1\dfrac{\varepsilon}{\tau_1} & \text{if} \quad \tau_2 \leq \tau \leq \tau_1, \\ P(\tau,\varepsilon) = \Gamma_1\dfrac{\varepsilon}{\tau} & \text{if} \quad \tau_1 \leq \tau. \end{cases}$$

Nevertheless, we remark that the heaviest phase is described by the lowest adiabatic coefficient, which is in contradiction to what is described, for example, in [22, Chapter XI]. Thus, the two perfect gas model is a good mathematical model because the mixture equation of state can be explicitly calculated, but it cannot give a good account for the physical behavior.

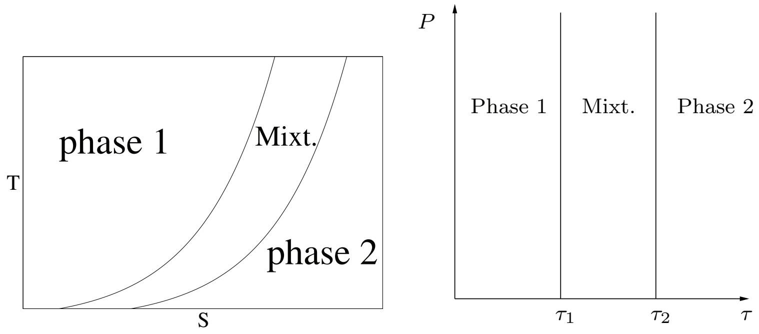


FIG. 3. *Shape of the saturation dome for the two perfect gas model. We note that the fluid is always retrograde.*

1.7.2. The two stiffened gas model. We model the two phases of a fluid with the stiffened gas equation of state, for which we have (see [9])

$$\varepsilon(P,\tau) = \frac{P+\overline{\gamma}P^{\infty}}{\overline{\gamma}-1}\tau + q, \tag{1.17a}$$

$$s(P,T) = C_v \log\left(\frac{T^{\overline{\gamma}}}{(P+P^{\infty})^{\overline{\gamma}-1}}\right) + q', \tag{1.17b}$$

$$G(P,T) = (\overline{\gamma}C_v - q')T - C_v T \log\left(\frac{T^{\overline{\gamma}}}{(P+P^{\infty})^{\overline{\gamma}-1}}\right) + q. \tag{1.17c}$$

For this equation of state, the adimensioned coefficients are given by

$$\gamma = \overline{\gamma}\left(1+\frac{P^{\infty}}{P}\right), \qquad \Gamma = \overline{\gamma}-1, \qquad g = \frac{(\overline{\gamma}-1)P}{P+P^{\infty}}.$$

The conditions $\gamma > 0$ and $g > 0$ are ensured if $\overline{\gamma} > 1$. In [9], the coefficients $q, q', C_v, \overline{\gamma}, P^{\infty}$ were calculated for the gaseous and liquid phases to fit with the saturation curves near $T = 298$K. These coefficients are given in Table 2. The resulting function $P_{\text{sat}}(T)$ was drawn in Figure 4. For the two stiffened gas model, we cannot be sure that the functions $T \mapsto s_v(T) - s_l(T)$ and $T \mapsto \tau_v(T) - \tau_l(T)$ simultaneously vanish. Therefore, the critical point does not really exist. As we saw in section 1.4, we need $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0$ to ensure the convexity of the mixture equation of state. Thus, the model is valid only when $\tau_v(T) - \tau_l(T)$ and $s_v(T) - s_l(T)$ are both positive. In our example, with the coefficients of Table 2, the function $T \mapsto P_{sat}(T)$ is drawn in Figure 4. We can see that the limit is near $T = 970\,\mathrm{K}$, for which $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T}$ vanishes.

1.8. Positivity of the fundamental derivative. In section 1.4, we found some criteria to ensure the convexity of the mixture equation of state. Nevertheless, we did not manage to find a simple criterion that can also ensure the positivity of the fundamental derivative of the mixture equation of state. Actually, as shown

TABLE 2
Thermodynamic coefficients for the liquid and gaseous phase of dodecane.

Phase	$\overline{\gamma}$	P^{∞}	C_v	q	q'
Gas	1.025	0	1956.45	-237547	-24485
Liquid	2.35	4.10^8 Pa	1077.7	-755269	0

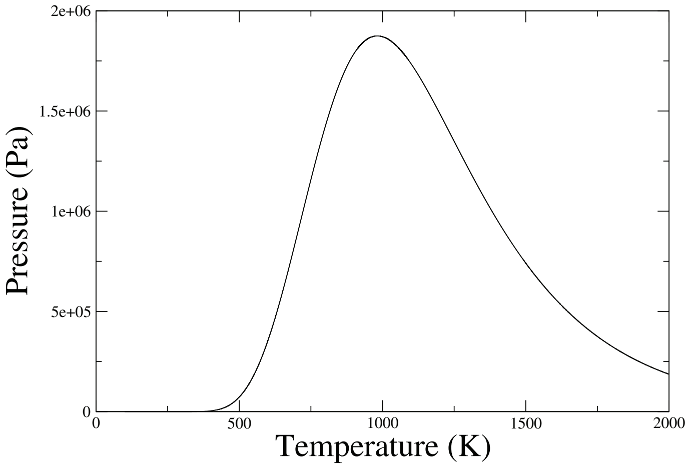


FIG. 4. *Numerical computation of the behavior of $P_{\text{sat}}(T)$ for two phases of the stiffened gas model with the coefficients of Table* 2. *For temperatures below* $970K$, $T \mapsto P_{\text{sat}}(T)$ *increases. For* $T \approx 970K$, *the function* $T \mapsto s_v(T) - s_l(T)$ *vanishes and its sign changes, whereas the function* $T \mapsto \tau_v(T) - \tau_l(T)$ *does not vanish. As a consequence,* $T \mapsto P_{\text{sat}}(T)$ *does not increase anymore, and the equation of state is no longer valid.*

numerically in Figure 5 for the thermodynamic coefficients of Table 2, the fundamental derivative of the mixture equation of state can be either positive or negative, even if both equations of state have a positive fundamental derivative. In that example, the equation of state can be considered as valid if $T \leq 715$K, where $\mathscr{G} > 0$.

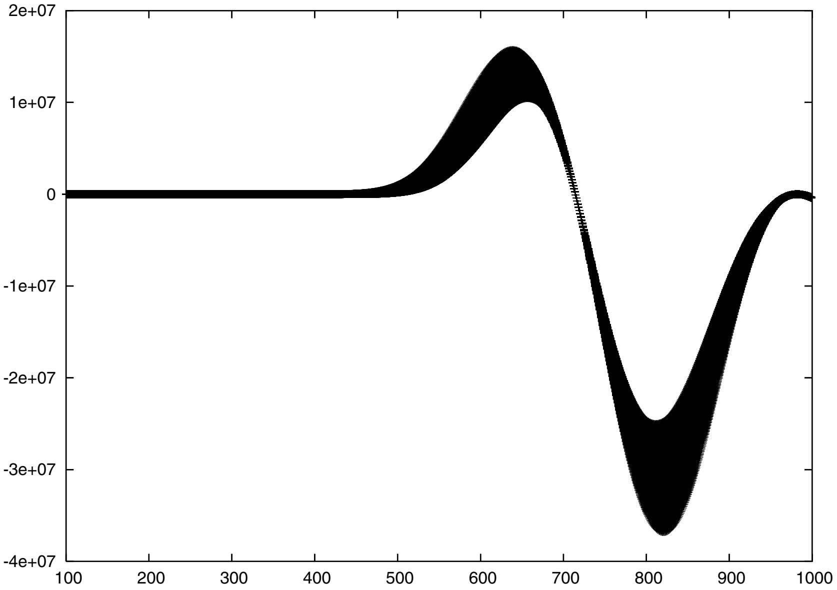


FIG. 5. *In this figure, we draw all the* $T \mapsto \mathscr{G}$ *for* $y \in [0;1]$. *The equation of state is the one of Table* 2. *For low temperatures* ($T \leq 715K$), *the fundamental derivative is positive, whereas for* $715 \leq T \leq 970K$ *the fundamental derivative is negative.*

2. Recollections of the Chapman–Jouguet theory (see [6, pp. 142–160]). Based on the experiments of [13, 16, 17], we propose to take into account out-of-thermodynamic-equilibrium states for solving the Riemann problem. By "out-of-equilibrium states" we mean *metastable states*, or *overheated* states, i.e., pure fluids that have a pressure P and a specific volume τ that lie in the saturation dome. Existence of such states is due to some phenomena such as capillarity, for example. In [13, 16, 17] it was observed that phase transition waves were self-similar waves, so that Rankine–Hugoniot relations hold across them: $\left[F(U) - \sigma U\right] = 0$, where σ is the velocity of the discontinuity. These relations can be put in the following form (see [6], for example):

$$\dot{M} = \frac{u_1 - u_0}{\tau_1 - \tau_0}, \tag{2.1a}$$

$$\dot{M}^2 = -\frac{P_1 - P_0}{\tau_1 - \tau_0}, \tag{2.1b}$$

$$\varepsilon_1 - \varepsilon_0 + \frac{1}{2}(P_1 + P_0)(\tau_1 - \tau_0) = 0, \tag{2.1c}$$

where $\dot{M}$ is the flow rate across the wave: $\dot{M} = \rho(u - \sigma)$. The interest of writing the Rankine–Hugoniot relations as in (2.1) is that the last two equations are purely thermodynamic. Equation (2.1b) describes the Rayleigh line in the (τ, P) plane. Equation (2.1c) describes the Crussard curve. The very difference with classical shock relations is that the set of the downstream states is not described with the same equation of state as that of the upstream one. For that sort of wave, we can use the CJ theory. Let us recall the main points of that theory (see [6] or [4] for the details and the proofs).

PROPERTY 1 (position of the initial state and the Crussard curve). *Suppose that the equation of state $(\tau, s) \mapsto P(\tau, s)$ has the properties*

$$\left(\frac{\partial P}{\partial \tau}\right)_s < 0 \qquad \textit{and} \qquad \left(\frac{\partial P}{\partial s}\right)_\tau > 0, \tag{2.2a}$$

and the reaction is exothermic:

$$\varepsilon_1(\tau_0, p_0) < \varepsilon_0(\tau_0, p_0). \tag{2.2b}$$

Then the point A_0 corresponding to the upstream state is under the Crussard curve.

In the first property, note that (2.2a) is always true provided that γ and Γ are both positive (thanks to (1.7c)). We will find in section 3.1 a condition to ensure the exothermic property (2.2b).

PROPERTY 2. *Suppose, moreover, that*

$$\left(\frac{\partial^2 P}{\partial \tau^2}\right)_s > 0;$$

then the Crussard curve is convex. Hence, the Rayleigh line (2.1b) *and the Crussard curve are crossing in zero or two points.*

Note that the Property 2 supposes that the pressure can be differentiated twice, which is not the case in our application, because of the local loss of derivative due to phase transition. Nevertheless, if Properties 1 and 2 hold for the equation of state

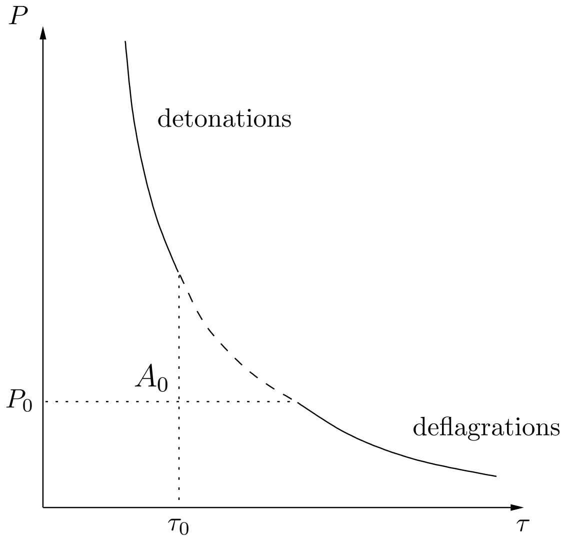


FIG. 6. *The Crussard curve related to an initial point* (τ_0, P_0). *The curve is cut into three parts: the upper one is the detonation branch, the lower one is the deflagration branch, and the middle part does not match the negative slope of the Rayleigh line.*

of the downstream states, then the Crussard curve can be schematically drawn as in Figure 6. The Crussard curve is cut into two parts: the upper part is called the detonation branch, and the lower one is the deflagration branch. In the middle part of the curve, $\frac{P_1-P_0}{\tau_1-\tau_0} > 0$. This does not match the negative slope of the Rayleigh line (2.1b).

Each part of the Crussard curve is itself cut into two parts, separated by the tangential point of the Rayleigh line with the Crussard curve (the existence of such a tangential point can be shown under some assumption on the asymptotic behavior of equation of state). Both branches are schematically drawn in Figure 7.

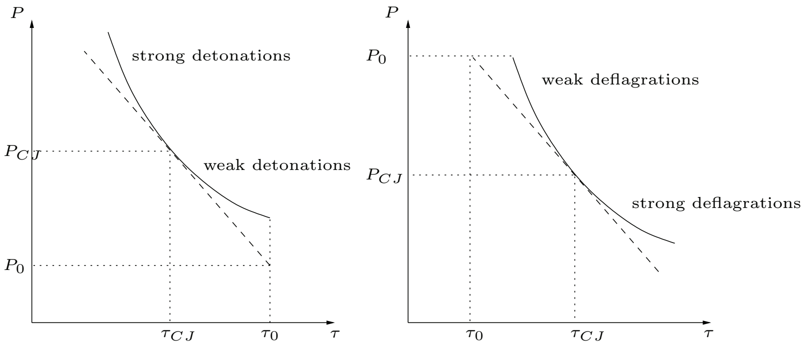


FIG. 7. *Zoom on the Crussard curve; on the left side, the detonation branch* $(P \geq P_0)$ *is cut into two parts by the CJ point. The upper part is the part of the strong detonations, and the lower part is called the part of the weak detonations. On the right side, the deflagration branch* $(P \leq P_0)$ *is cut into two parts by the CJ point. The upper part is the part of the weak deflagrations, and the lower part is called the part of the strong deflagrations.*

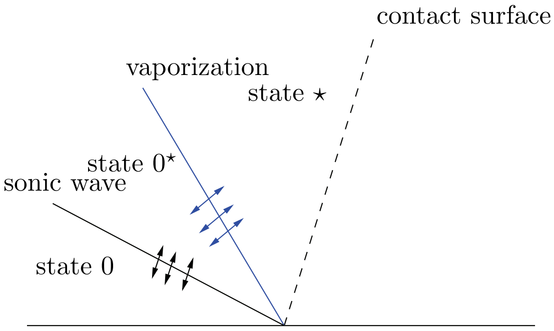


FIG. 8. *Structure of the half-Riemann problem: the state* 0 *is linked with the state* $0^\star$ *by a forerunner sonic wave (rarefaction wave or shock). Then the state* $0^\star$ *and the state* $\star$ *are linked with a deflagration wave. Eventually, there is a contact discontinuity.*

PROPERTY 3. *Along the Crussard curve, the velocity* $|v| = |u - \sigma|$ *has a local minimum on the CJ-detonation point and a local maximum on the CJ-deflagration point. More precisely, we have the following properties:*

$$\begin{array}{llll}
\textit{for a strong detonation} & : & |v_0| > c_0, & |v_1| < c_1, \\
\textit{for a weak detonation} & : & |v_0| > c_0, & |v_1| > c_1, \\
\textit{for a weak deflagration} & : & |v_0| < c_0, & |v_1| < c_1, \\
\textit{for a strong deflagration} & : & |v_0| < c_0, & |v_1| > c_1.
\end{array}$$

This last property is very important because it can allow us to know the structure of the half-Riemann problem with a combustion wave provided that we know which "family" the combustion wave belongs to. In our case, we are interested in waves which transform a heavy phase into a lighter one. Therefore, we expect that τ will increase, so that we will concentrate only on the deflagration branch of the Crussard curve.

Thanks to Property 3, we can give the structure of the Riemann problem in the case of strong and weak deflagrations. In both cases, from Property 2, the deflagration wave is always subsonic relative to the liquid; for example, if the liquid is on the left, we have $\dot{M} > 0$, and $u_0 - c_0 < \sigma < u_0$. For the position of the wave relative to the fields of 1, we have that

- if the wave is a strong deflagration, then $\sigma < u_1 - c_1$;
- if the wave is a weak deflagration, then $u_1 - c_1 < \sigma < u_1$.

In [4, p. 230], it is shown that under the assumption that a wave is a deflagration and that across that wave, the mass fraction of gas always increases, the reaction is a weak deflagration. We will suppose that we are always in that case in the following.

The structure of the Riemann problem with weak deflagrations is drawn in Figure 8. The problem for deflagrations is that the Lax characteristic criterion is not satisfied (see [6, p. 154]), and the Riemann problem cannot be solved with only the classical relations across the sonic wave and the Rankine–Hugoniot relations across the subsonic wave. There remains one indeterminate. The supplementary relation needed is often called "the kinetic closure."

3. Application to the solution of the Riemann problem with vaporization.

3.1. Useful verifications for the use of CJ theory.

In this section, we check whether the inequalities needed for applying the CJ theory hold.

THEOREM 3.

1. *If both equations of state are convex and if* $\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0$, *then inequalities* (2.2a) *hold.*
2. *If* (P, τ) *lie in the saturation dome and under the condition*

$$\frac{\gamma}{\Gamma} - \frac{T}{P}\frac{\mathrm{d}P_{sat}}{\mathrm{d}T} > 0, \tag{3.1}$$

then the inequality (2.2b) *holds.*

Proof. According to the identity (1.7c), it is sufficient to have γ and $\Gamma > 0$. This is supposed for pure fluids, and this is ensured for the mixture equation of state if $\frac{\mathrm{d}P_{\mathrm{sat}}}{\mathrm{d}T} > 0$, and thus (2.2a) holds.

Let us now check if the inequality (2.2b) is ensured. We suppose that (P, τ) lie in the saturation dome, so that the corresponding equilibrium downstream state is a mixture,

$$\varepsilon_1(\tau, P) = \varepsilon_m(\tau, P) = y_l\varepsilon_l(P, \tau_l(P)) + (1 - y_l)\varepsilon_v(P, \tau_v(P)),$$

and we want to know if $\varepsilon_1(P, \tau) - \varepsilon_0(P, \tau) < 0$, the state 0 being, of course, described by the liquid equation of state. For this, we denote

$$\delta\varepsilon(y_l) = y_l\varepsilon_l(P, \tau_l(P)) + (1 - y_l)\varepsilon_v(P, \tau_v(P)) - \varepsilon_l(P, y_l\tau_l(P) + (1 - y_l)\tau_v(P)),$$

and we immediately see that $\delta\varepsilon(1) = 0$. It remains to show that $\delta\varepsilon$ is an increasing function:

$$\begin{aligned}\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}y_l}(y_l) =& \varepsilon_l(P, \tau_l(P)) - \varepsilon_v(P, \tau_v(P))\\ &-(\tau_l(P) - \tau_v(P))\left(\frac{\partial\varepsilon_l}{\partial\tau}\right)_P(P, y_l\tau_l(P) + (1 - y_l)\tau_v(P)).\end{aligned}$$

Integration of the identity $\mathrm{d}\varepsilon + P\mathrm{d}\tau = T\mathrm{d}s$ across the saturation dome leads to

$$\varepsilon_l(P, \tau_l(P)) - \varepsilon_v(P, \tau_v(P)) + P(\tau_l(P) - \tau_v(P)) = T(s_l(P) - s_v(P)),$$

so that

$$\begin{aligned}\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}y_l}(y_l) =& -P(\tau_l(P) - \tau_v(P)) + T(s_l(P) - s_v(P))\\ &-(\tau_l(P) - \tau_v(P))\left(\frac{\partial\varepsilon_l}{\partial\tau}\right)_P(P, y_l\tau_l(P) + (1 - y_l)\tau_v(P)),\end{aligned}$$

which can be cast into the following form, thanks to (1.5) and (1.7f):

$$\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}y_l}(y_l) = P(\tau_v - \tau_l)\left(\frac{\gamma}{\Gamma} - \frac{T}{P}\frac{\mathrm{d}P_{\mathrm{sat}}}{\mathrm{d}T}\right).$$

As $P > 0$, $\tau_v - \tau_l > 0$, and as (3.1), $\delta\varepsilon$ increases, so that $\delta\varepsilon \leq \delta\varepsilon(1) = 0$. Thus, (2.2b) holds. ∎

Remark 1. Supposing that (τ, P) is always in the saturation dome is not a strong assumption. Indeed, as the upstream state is slightly compressible, its specific volume cannot increase a lot across a sonic wave, and it is likely that a metastable liquid with a specific volume equal to that of a gas at equilibrium cannot exist, except just near the critical point.

Remark 2. The condition (3.1) holds at least in the following two frameworks:

1. The terms $\frac{\gamma}{\Gamma}$ and $\frac{T}{P}\frac{\mathrm{d}P_{\mathrm{sat}}}{\mathrm{d}T}$ can easily be compared near the saturation curve. Indeed, if we use (1.7b) with saturated variables, we find

$$\frac{T^2(\gamma g-\Gamma^2)}{P\tau}\frac{\mathrm{d}s_l}{\mathrm{d}T}=\frac{\gamma}{\Gamma}-\frac{T}{P}\frac{\mathrm{d}P_{\mathrm{sat}}}{\mathrm{d}T}, \tag{3.2}$$

so that we have $\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}T}(1)=(\tau_v-\tau_l)\,\frac{T^2}{\tau}\,\frac{\gamma g-\Gamma^2}{\Gamma}\,\frac{\mathrm{d}s_l}{\mathrm{d}T}$.
Thus, if the liquid saturation curve is regular (which is always the case), then $\frac{\mathrm{d}s_l}{\mathrm{d}T}>0$, so that condition (3.1) is ensured.

2. For a simple model, such as perfect gas or stiffened gas, we have

$$\frac{\gamma}{\Gamma}=\frac{\bar\gamma\left(1+\frac{P^\infty}{P}\right)}{\bar\gamma-1},$$

so that $\frac{\gamma}{\Gamma}$ does not depend on the specific volume. Thus, equality (3.2) holds for any τ, so that condition (3.1) always holds.

Remark 3. The same calculations can be made for liquefaction. Then we find that, near the vapor saturation curve,

$$\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}T}(1)=(\tau_l-\tau_v)\,\frac{T^2}{\tau}\,\frac{\gamma g-\Gamma^2}{\Gamma}\,\frac{\mathrm{d}s_v}{\mathrm{d}T}.$$

Thus, if the fluid is regular, then $\frac{\mathrm{d}s_v}{\mathrm{d}T}<0$, and as $\tau_l-\tau_v<0$, then $\frac{\mathrm{d}\delta\varepsilon}{\mathrm{d}T}(1)>0$, so that locally we have $\varepsilon_m-\varepsilon_v<0$ and the CJ theory can be used. If the fluid is retrograde, then we find that locally $\varepsilon_m-\varepsilon_v>0$ and the CJ theory may be used, but by exchanging the upstream and the downstream states. Note that the Hugoniot curves can enter the saturation dome only in the retrograde case.

The CJ theory also relies heavily on the convexity properties of the Crussard curve (see Property 2 of section 2), which are ensured if the fundamental derivative $\mathscr{G}$ is positive. Nevertheless, even if we suppose that the liquid and the gas equations of state have a positive fundamental derivative, the mixture equation of state can have a negative fundamental derivative, as was shown numerically in section 1.8. This nonpositivity of the fundamental derivative can lead to a wrong behavior of the Crussard curve as shown in Figure 9: the CJ points do not exist anymore, and all the undercompressive downstream states are strong deflagrations. If the sign of the fundamental derivative changed many times along the Crussard curve, we could expect to observe several CJ points. From now, we suppose that $\mathscr{G}>0$.

3.2. Entropy growth criterion. As the particles are crossing the front from the liquid area to a mixture or pure phase area, we have to check whether the entropy growth criterion is ensured, i.e., if the entropy of the downstream state (gas or mixture) is greater than the entropy of the upstream state (liquid). We first prove the following theorem.

THEOREM 4. *Let $s=s_0$ be a liquid isentrope that crosses the liquid saturation curve. To any metastable point (τ_0,P_0) on that isentrope, we map the point (τ_P,P_0), point of constant pressure deflagration (see Figure* 10*). If we suppose that*

$$\forall(\tau_0,P_0)\qquad \gamma(\tau_0,P_0)>\gamma(\tau_P,P_0), \tag{3.3}$$

then $s(\tau_P,P_0)>s_0$.

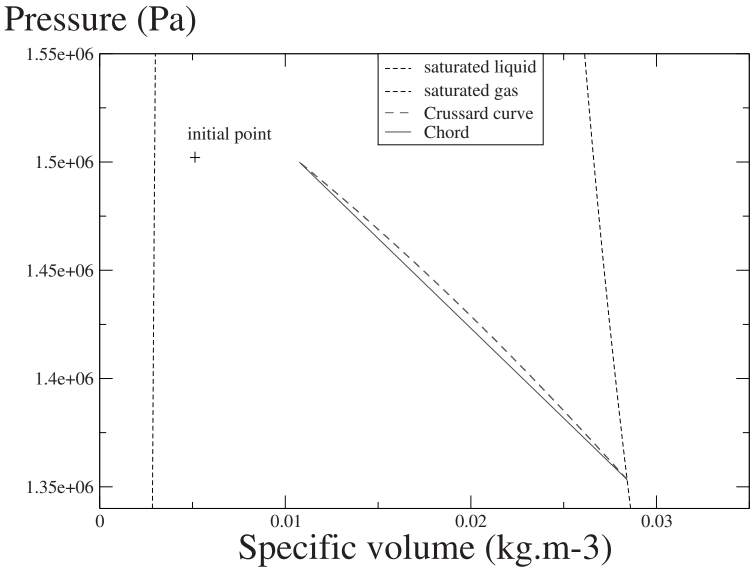


FIG. 9. *Wrong behavior of the mixture Crussard curve if the condition $\mathscr{G} > 0$ is violated. We notice that the Crussard curve is concave, which induces no existence of any CJ point. For a more complicated pair of equations of state, we could expect to observe two or three tangential points if the sign of $\mathscr{G}$ changed two or three times along the Crussard curve.*

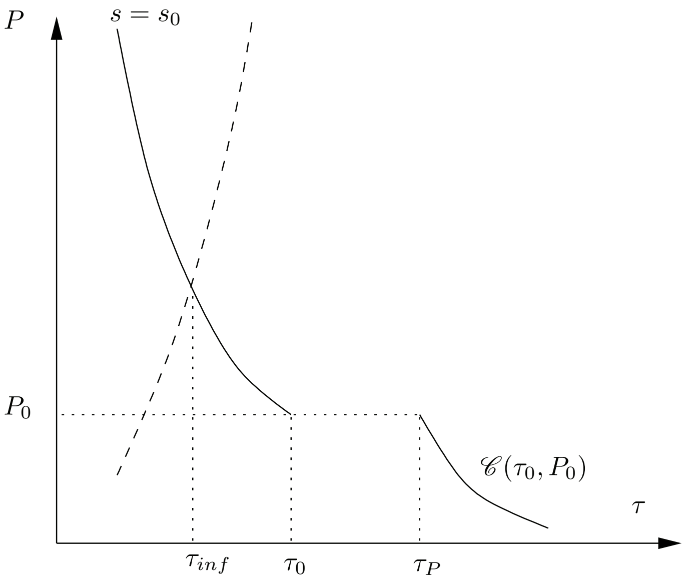


FIG. 10. *Entropy growth criterion. To any point (τ_0, P_0) on a given isentrope $s = s_0$, we associate the point on the Crussard curve (τ_P, P_0). The liquid saturation curve is drawn with dashed lines. When τ_0 is on the liquid saturation curve, we have $\tau_P = \tau_0$ so that $s_P = s_0$. Thus, to show that $s_P \geq s_0$, we only have to prove that entropy of the point P grows when τ_0 increases.*

Proof. To any point (τ_0, P_0) on this isentrope, we associate the point (τ_P, P_0), the point of constant pressure deflagration (see Figure 10). For greater convenience in notation, we suppose that (τ_P, P_0) is a mixture state (i.e., all the linked quantities have m). τ_P is defined by the implicit equation

$$\varepsilon_m(\tau_P, P_0) - \varepsilon_l(\tau_0, P_0) + P_0(\tau_P - \tau_0) = 0. \tag{3.4}$$

Differentiation of (3.4) with respect to τ_P is equal to $\frac{\gamma_m}{\Gamma_m}$, which never vanishes, so that according to the implicit function theorem, τ_P is a $\mathscr{C}^1$ function of τ_0 and P_0. Moreover, we can calculate its derivative with respect to τ_0 and P_0:

$$\begin{cases} \left(\dfrac{\partial \tau_P}{\partial \tau_0}\right)_{P_0} = \dfrac{\gamma_l}{\gamma_m}\dfrac{\Gamma_m}{\Gamma_l}, \\ \left(\dfrac{\partial \tau_P}{\partial P_0}\right)_{\tau_0} = \dfrac{\tau_0}{\gamma_m P_0}\left(\dfrac{\Gamma_m(\Gamma_l+1)}{\Gamma_l} - \dfrac{\tau_P}{\tau_0}(\Gamma_m+1)\right). \end{cases}$$

Besides, as we supposed that the points (P_0, τ_0) belong to the same isentrope, P_0 is actually a function of τ_0 with $\frac{\mathrm{d}P_0}{\mathrm{d}\tau_0} = -\frac{\gamma_l P_0}{\tau_0}$, so that τ_P is a function of the only variable τ_0 and

$$\begin{aligned} \frac{\mathrm{d}\tau_P}{\mathrm{d}\tau_0} &= \left(\frac{\partial \tau_P}{\partial \tau_0}\right)_{P_0} + \frac{\mathrm{d}P_0}{\mathrm{d}\tau_0}\left(\frac{\partial \tau_P}{\partial P_0}\right)_{\tau_0} \\ &= \frac{\gamma_l}{\gamma_m}\left(-\Gamma_m + \frac{\tau_P}{\tau_0}(\Gamma_m+1)\right). \end{aligned}$$

Now, we calculate the entropy variation of the point τ_P when the point (P_0, τ_0) follows the isentrope $s = s_0$:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}\tau_0} &= \frac{\mathrm{d}\tau_P}{\mathrm{d}\tau_0}\left(\frac{\partial s}{\partial \tau}\right)_P + \frac{\mathrm{d}P_0}{\mathrm{d}\tau_0}\left(\frac{\partial s}{\partial P}\right)_\tau \\ &= \frac{\gamma_m P_0}{T\Gamma_m}\left(\frac{\gamma_l \Gamma_m}{\gamma_m}\left(\frac{\tau_P}{\tau_0} - 1\right) + \frac{\tau_P}{\tau_0}\left(\frac{\gamma_l}{\gamma_m} - 1\right)\right). \end{aligned}$$

According to the hypothesis (3.3), $\gamma_l > \gamma_m$. Moreover, as we have $\tau_P - \tau_0 > 0$, s is an increasing function of τ_0. Furthermore, in the limit of no overheating, we have

$$\lim_{\tau_0 \to \tau_{\mathrm{inf}}} \tau_P(\tau_0, P_0) = \tau_{inf},$$

where τ_{inf} is the crossing point of the isentrope $s = s_0$ with the saturation curve. Thus $\lim\limits_{\tau_0 \to \tau_{\mathrm{inf}}} s(\tau_P, P_0) = s_0$. As a conclusion,

$$\forall \tau_0 \geq \tau_{inf} \qquad s(\tau_P, P_0) \geq s_0,$$

which ends the proof. □

Remark (about the hypothesis (3.3)*).*

1. We know that near the saturation curve, we have $\gamma_l > \gamma_m$. For actual data, we have $\gamma_l \gg \gamma_m$. Thus we can suppose that any γ is greater than any γ_m.
2. $\gamma_l > \gamma_v$ just means that the liquid phase is very much less compressible than the gas phase (see, e.g., [22, Chapter XI]).

Given an initial point, we know that the entropy grows from the constant pressure point to the CJ point, so that if the entropy growth criterion holds for the constant pressure deflagration point, it holds for all the downstream states between the constant pressure deflagration point and the CJ point; i.e., we have the following corollary.

Corollary 1. *Under the same hypothesis as in Theorem* 4*, the entropy growth criterion holds for all the weak deflagration points.*

3.3. Behavior of the Crussard curve near the gas saturation curve. In section 1.5, the behavior of the isentropes near the saturation curves was studied. The difference of the differential behavior of the pure phase and the mixture equation of state induced kinks in isentropes. Now we want to study the behavior of the Crussard curve when it crosses the vapor saturation curve. It is more difficult than the study of the isentrope, because the Crussard curve depends not only on the local variables, but also on the starting point (τ_0, P_0).

3.3.1. General study. We denote by C the point at which the Crussard curve crosses the saturation curve, and by

$$\zeta = -\frac{\tau}{P}\frac{\mathrm{d}P}{\mathrm{d}\tau}_{|\mathscr{C}}$$

the adimensioned slope of the Crussard curve.

The first thing we will prove for the behavior of the Crussard curve near the saturation curve is that it can be parameterized by τ under some conditions.

Theorem 5. *If all the equations of state are convex and if* $\Gamma > 0$*, then* $\zeta > 0$*. With the same hypothesis, the Crussard curve can be parameterized by* τ*, even near the saturation curve.*

Proof. As proved in [12, p. 101], we have

$$\zeta = \frac{\frac{\gamma}{\Gamma} - \frac{\Delta P}{2P}}{\frac{1}{\Gamma} + \frac{\Delta\tau}{2\tau}}. \tag{3.5}$$

Across a deflagration wave, we have $\Delta\tau > 0$ and $\Delta P < 0$. Moreover, we proved that $\Gamma_m > 0$, and we suppose that $\Gamma > 0$. The conditions $\gamma > 0$ and $\gamma_m > 0$ were already supposed to ensure the convexity of the specific energy. Then $\zeta > 0$. If we combine (3.5) with the identities near the saturation boundary of section 1.5, we get

$$\frac{\xi - \zeta}{\Gamma}\left(1 + \Gamma\frac{\Delta\tau}{2\tau}\right) = \frac{\xi - \zeta_m}{\Gamma_m}\left(1 + \Gamma_m\frac{\Delta\tau}{2\tau}\right). \tag{3.6}$$

Across a deflagration, we have $\Delta\tau > 0$. Then

$$\frac{\xi - \zeta}{\xi - \zeta_m} > 0,$$

which means that the Crussard curve, near a boundary, can be parameterized by τ. As $\zeta > 0$, the Crussard curve is a diffeomorphism of τ in each side of the saturation curve. As the Crussard curve can locally parameterize the Crussard curve near a boundary, we conclude that the Crussard curve is a homeomorphism of τ. □

Remark. As the Crussard curve is a decreasing homeomorphism in τ, the point of constant pressure deflagration is uniquely defined.

To be more precise on the relative behavior of the isentropes, the Crussard curve, and the saturation curve, we will prove that the following theorem holds.

Theorem 6. *The relative behavior of the isentropes and the Crussard curves, which gives the nature of the deflagration on each side of the saturation curve, follows the alternative*

- *if $\gamma > \xi$, then*
  - *either $\xi \geq \zeta$ (Case 1), and the deflagration is weak on both sides of the saturation curve,*
  - *or $\xi < \zeta$ (Case 2), and point C cannot be a weak deflagration simultaneously on both sides of the saturation curve;*
- *if $\gamma < \xi$, then*
  - *either $\xi \leq \zeta$ (Case 3), and the deflagration is strong on both of the sides of the saturation curve,*
  - *or $\xi > \zeta$ (Case 4), and the deflagration cannot be simultaneously strong on the mixture side and weak on the pure phase side.*

Proof. Equation (3.6) can be rewritten as

$$\frac{\xi - \zeta}{\xi - \zeta_m} = \frac{\frac{1}{\Gamma_m} + \frac{\Delta\tau}{2\tau}}{\frac{1}{\Gamma} + \frac{\Delta\tau}{2\tau}}$$

so that the discontinuity in the slope of the Crussard curve is directly linked with the sign of $\Gamma_m - \Gamma$ (we recall that $\Gamma_m > 0$ and that Γ has the same sign as Γ_m near the saturation curves). Equation (1.14) induces a separation into the following cases:

- $\gamma > \xi$.
  If $\gamma > \xi$ then we also have $\gamma_m > \xi$. As $\gamma_m < \gamma$, we have $\frac{\gamma_m - \xi}{\gamma - \xi} \leq 1$, so that $\Gamma_m \leq \Gamma$. Therefore

  $$\frac{\xi - \zeta}{\xi - \zeta_m} \geq 1.$$

  Suppose first that $\xi - \zeta \geq 0$ (Case 1). Then $\xi - \zeta \geq \xi - \zeta_m$ so that $\zeta \leq \zeta_m \leq \xi$. In that case, as shown in Figure 11, the relative behavior of the isentrope and the Crussard curve shows that on both sides of the saturation curve, the downstream state is a weak deflagration (see Figure 11). In that case, we have $\zeta \leq \zeta_m \leq \xi \leq \gamma_m \leq \gamma$.
  Suppose now that $\xi - \zeta \leq 0$ (Case 2). Then we have $\xi \leq \zeta_m \leq \zeta$. The nature of the deflagration is given by the relative position of the slope of the Crussard curve and the Rayleigh line, so that on the point saturation curve, there are three subcases (see Figure 12):
  - If the Rayleigh line has a lower slope than both of the slopes of the Crussard curve, then the two parts match with strong deflagrations. Thus, we have $\gamma_m \leq \zeta_m$ and $\gamma \leq \zeta$ (see Figure 12(a)).
  - If the slope of the Rayleigh line is between the slopes of the Crussard curve, then the mixture Crussard curve matches with strong deflagrations, whereas the pure phase Crussard curve matches with weak deflagrations. In that case, we have $\gamma_m \leq \zeta_m$ and $\gamma \geq \zeta$ (see Figure 12(b)). In that case we have $\xi \leq \zeta_m \leq \gamma_m \leq \gamma \leq \zeta$.
  - If the Rayleigh line has a greater slope than both of the slopes of the Crussard curve, then point C is a weak deflagration with respect to the pure and the mixture Crussard curves. Therefore, we have $\gamma_m \geq \zeta_m$ and $\gamma \geq \zeta$ (see Figure 12(c)).

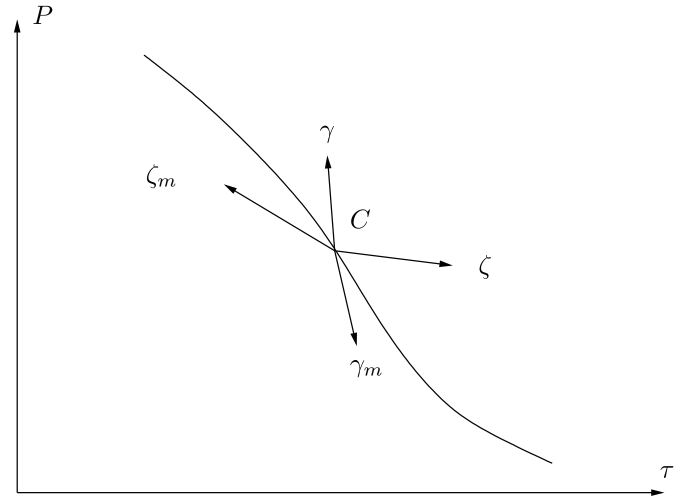


FIG. 11. *Qualitative relative behavior of the isentrope and of the Crussard curve when they cross the vapor saturation curve in the Case* 1. *Arrows represent the half-tangent of the Crussard curve (ζ) and of the isentrope (γ).*

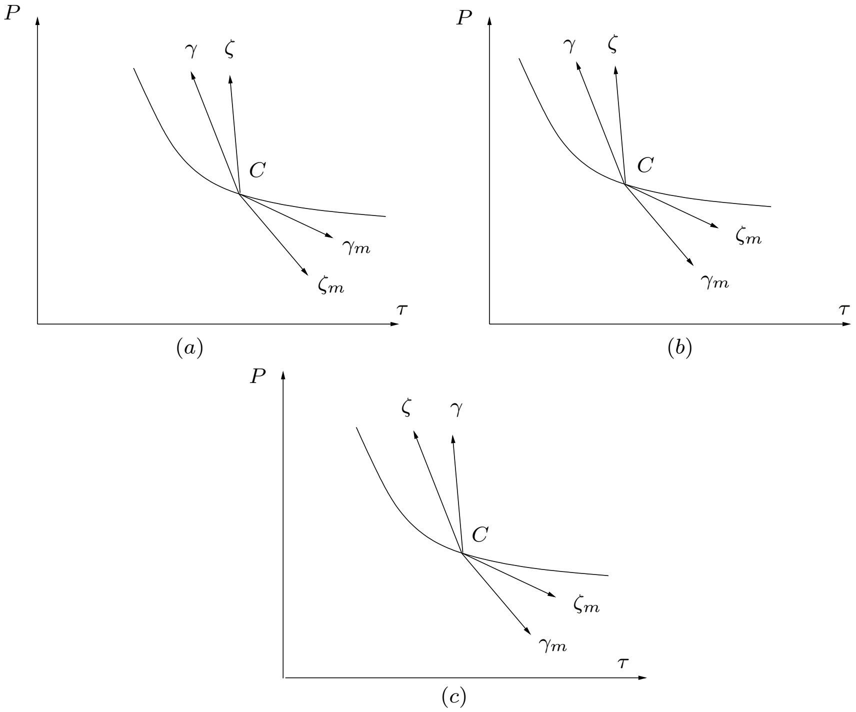


FIG. 12. *Qualitative relative behavior of the isentrope and of the Crussard curve when they cross the vapor saturation curve, Case* 2. *Arrows represent the half-tangent of the Crussard curve (ζ) and of the isentrope (γ).*

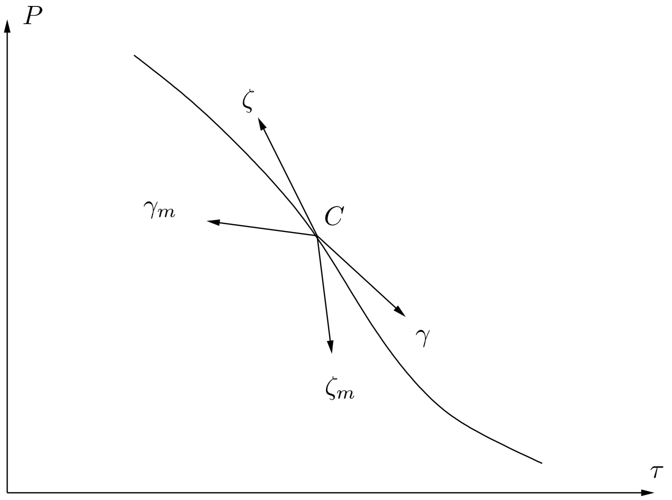


FIG. 13. *Qualitative relative behavior of the isentrope and of the Crussard curve when they cross the vapor saturation curve, Case* 3. *Arrows represent the half-tangent of the Crussard curve* (ζ) *and of the isentrope* (γ).

- $\gamma < \xi$.

  If $\gamma < \xi$, then we also have $\gamma_m < \xi$. As we know that $\gamma_m \leq \gamma$, we have $\frac{\gamma_m - \xi}{\gamma - \xi} \geq 1$, so that $\Gamma_m \geq \Gamma$ (thanks to (1.14)). Therefore

  $$\frac{\xi - \zeta}{\xi - \zeta_m} \leq 1.$$

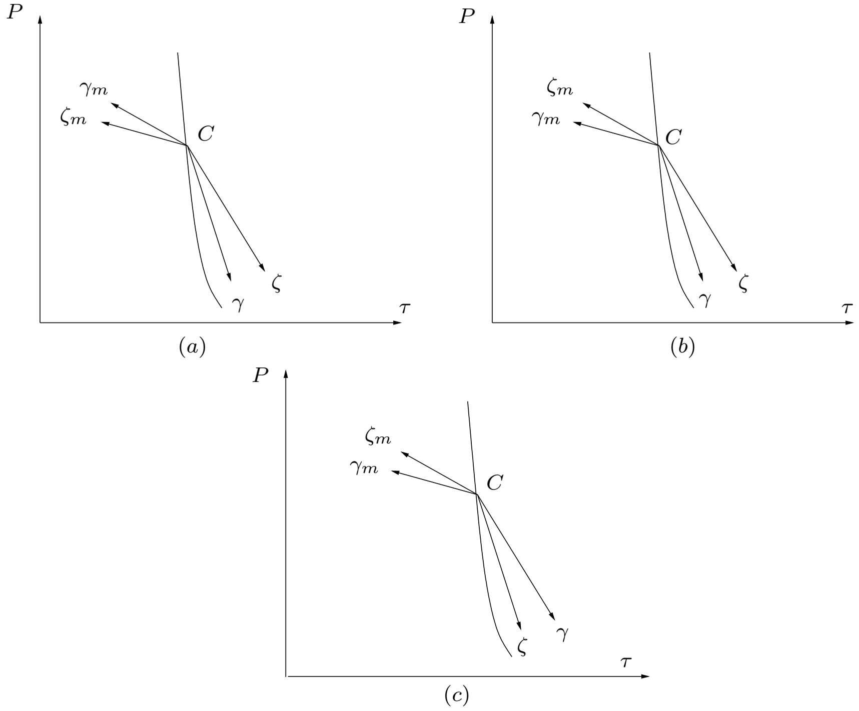
  We suppose first that $\xi - \zeta \leq 0$ (Case 3). Then we immediately have $\zeta \leq \zeta_m$. Thus, we have $\gamma_m \leq \zeta_m$ and $\gamma \leq \zeta$, so that point C matches on both sides of the Crussard curve with strong deflagrations (see Figure 13). In that case we have $\zeta_m \leq \zeta \leq \xi \leq \gamma \leq \gamma_m$.

  We suppose now that $\xi - \zeta \geq 0$ (Case 4). Then we have $\zeta_m \leq \zeta \leq \xi$. The nature of the deflagration is given by the relative position of the slope of the Crussard curve and the Rayleigh line, so that three subcases may happen (see Figure 14):

  - If the Rayleigh line has a lower slope than the slopes on both sides of the Crussard curve in C, then point C is a weak deflagration with respect to the mixture and the pure phase Crussard curve. We then have $\gamma \leq \zeta$ and $\gamma_m \leq \zeta_m$ (see Figure 14(a)).
  - If the slope of the Rayleigh line is between the slopes on each side of the Crussard curve, then point C is a strong deflagration for the mixture Crussard curve and a weak deflagration for the pure phase Crussard curve (see Figure 14(b)). In that case we have $\gamma_m \leq \zeta_m \leq \zeta \leq \gamma \leq \xi$.
  - If the Rayleigh line has a lower slope than both of the slopes of the Crussard curve, then point C is a strong deflagration with respect to the pure and the mixture Crussard curve. Therefore, we have $\gamma_m \geq \zeta_m$ and $\gamma \geq \zeta$ (see Figure 14(c)).

This ends the proof. □

3.3.2. Ill-posedness of the CJ closure. The following result comes immediately from Theorem 6.



(*a*) (*b*) (*c*)

FIG. 14. *Qualitative relative behavior of the isentrope and of the Crussard curve when they cross the vapor saturation curve, Case 4. Arrows represent the half-tangent of the Crussard curve (ζ) and of the isentrope (γ).*

COROLLARY 2. *Let (P_0, τ_0, u_0) be an initial state of liquid at thermodynamic equilibrium, such that the isentrope $\mathscr{C}_s$ coming from this point enters the saturation dome. If $(P_0^\star, \tau_0^\star)$ is a point in $\mathscr{C}_s$, we build $(P^\star, \tau^\star)$ in the following way:*

- *if $(P_0^\star, \tau_0^\star)$ is not in the saturation dome, then $P^\star = P_0^\star$ and $\tau^\star = \tau_0^\star$;*
- *if $(P_0^\star, \tau_0^\star)$ is in the saturation dome, then it is linked with $(P^\star, \tau^\star)$ with a CJ deflagration.*

If $(P^\star, \tau^\star)$ can reach the pure gas phase, then the curve $(P^\star, \tau^\star)$ is discontinuous.

Proof. We suppose that the set described is continuous. As $(P^\star, \tau^\star)$ can reach the saturation dome, it crosses the gas saturation curve at a point (P_c, τ_c). As it is a CJ point, we have $\gamma_m = \zeta_m$, so that we are in Case 4 in Figure 14. The third subcase of Case 4 is excluded because the Rayleigh line is tangential with the Crussard curve, so that the slope of the Rayleigh line is greater than the slope of the Crussard curve in the pure phase side. Thus, the point $\star$ matches with a weak deflagration with respect to the pure phase Crussard curve. As a consequence, the Crussard curve has another CJ point that lies in the pure phase area (see Figure 15), so that the curve $\mathscr{C}_{CJ}$ even has a branch in the pure gas area. $\square$

Eventually, we can state the following theorem.

THEOREM 7. *With the same hypothesis of Corollary 2, if we model the vaporization wave by a CJ deflagration, then the resulting solution of the Riemann problem is ill-posed in the L^1 sense: the solution does not depend continuously on the initial state.*

Proof. We fix a point for $x < 0$ in the liquid area for which the conditions of Corollary 2 hold, and we suppose that on the right, there is some gas. The Riemann

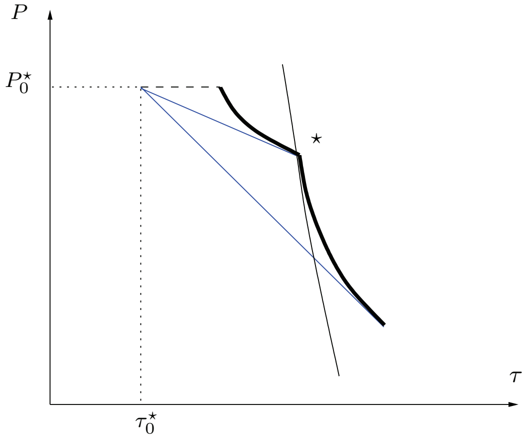


FIG. 15. *Qualitative behavior of the Rayleigh line and the Crussard curve when the Crussard curve crosses the saturation curve on a mixture CJ point. As $\zeta_m \le \zeta$, and as the Rayleigh line is tangential to the mixture Crussard curve, we are in the third subcase of Case* 4 *of the proof of Theorem* 6*. As a consequence, the pure phase side matches with a weak deflagration, too, so that there exists another CJ point.*

problem is composed (from left to right) of a sonic wave, a vaporization wave (if the intermediate state is metastable), a contact discontinuity (across which P and u are constant), and a sonic wave on the gas side. As is usually done [6], to solve the Riemann problem, we intersect the wave curve of the downstream state (sonic wave and maybe followed by a CJ vaporization) of the left side with the wave curve of the sonic wave of the right side, in the plane (P, u). Corollary 2 says that the wave curve of the left side is composed of (at least) two branches (see Figure 16). So that the gas wave curve intersects the liquid wave curve either in one mixture point (case (I)), or in two points (case (II)), or in one pure gas point (case (III)). Existence of case (I) and case (III) implies that we must jump from the mixture to the gas branch of the

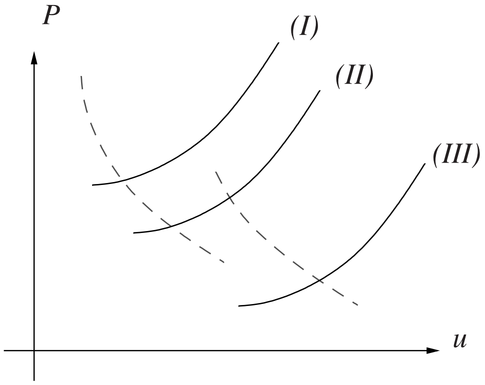


FIG. 16. *Dashed lines: the wave curve of the liquid side. Solid lines: different wave curves for the gas side, depending on the initial state. In case* (I)*, the gas wave curve intersects the liquid one in one mixture point. In case* (II)*, the gas wave curve intersects the liquid curve in the two branches: one mixture and one pure gas point. In case* (III)*, the gas wave curve intersects the liquid curve only on the pure phase branch.*

liquid wave curve. However, jumping from one branch to the other means that we significantly change the vaporized state (thus the L^1_{loc} norm, too) but hardly change the initial state. □

We finish this paper by drawing the curve $\mathscr{C}_{CJ}$ described in Corollary 2 for models we dealt with in section 1.

Example 1: The two perfect gas model.

As an example, we take the two perfect gas model. As we said before, this model enables us to make all the calculations, because the mixture equation of state is explicit.

- *Mixture CJ point.*
  In the case when the downstream state is a mixture, the equation of the Crussard curve is the following:

  $$\frac{P\tau_2}{\Gamma_2} - \frac{P_0\tau_0}{\Gamma_2} + \frac{1}{2}(P+P_0)(\tau-\tau_0) = 0,$$

  which gives an expression of τ as a function of P: $\tau = \tau_0 - \frac{2(P\tau_2 - P_0\tau_0)}{\Gamma_2(P+P_0)}$. The CJ point is such that $\frac{\mathrm{d}\tau}{\mathrm{d}P}(P_{CJ}) = \frac{\tau-\tau_0}{P-P_0}$, so that we find the following equation for P_{CJ}:

  $$\left(\frac{P}{P_0}\right)^2 - 2\frac{\tau}{\tau_0}\frac{P}{P_0} + 1 = 0,$$

  whose undercompressive solution is

  $$P_{CJ} = P_0\left(\frac{\tau_0}{\tau_2} - \sqrt{\left(\frac{\tau_0}{\tau_2}\right)^2 - 1}\right).$$

  τ_{CJ} is then given by

  $$\tau_{CJ} = \tau_2\left(\frac{\tau_0}{\tau_2} + \frac{2\sqrt{\frac{\tau_0}{\tau_2} - 1}}{\Gamma_2\left(\frac{\tau_0}{\tau_2} + 1 - \sqrt{\left(\frac{\tau_0}{\tau_2}\right)^2 - 1}\right)}\right).$$

  Of course, this point can be chosen only when the mixture is stable, that means, when $\tau_{CJ} \leq \tau_1$.
- *Vapor CJ point.*
  The equation of the Crussard curve is then

  $$\frac{\tau P}{\Gamma_1} - \frac{\tau_0 P_0}{\Gamma_2} + \frac{1}{2}(P+P_0)(\tau-\tau_0) = 0.$$

  As in [10], we first calculate the point of constant specific volume detonation, i.e., the downstream state such as $\tau = \tau_0$: $P_\tau = \frac{\Gamma_1}{\Gamma_2}P_0$. If we take the calculations of [10], we get

  $$P_{CJ} = \frac{\Gamma_1 P_0}{\Gamma_2}\left(1 - \sqrt{\left(1 - \frac{\Gamma_2}{\Gamma_1}\right)\left(1 + \frac{\Gamma_2}{\Gamma_1} + \frac{2\Gamma_2}{\Gamma_1(\gamma_1+1)}\right)}\right).$$

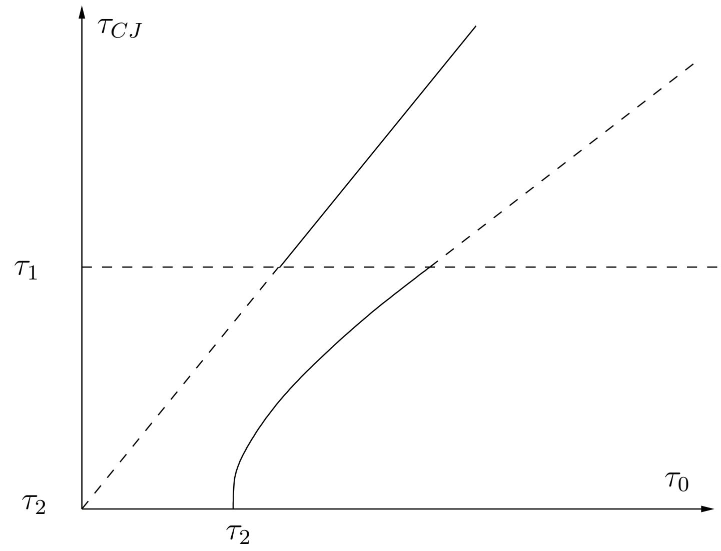


FIG. 17. *Qualitative behavior of the mixture and the vapor CJ points for the two perfect gas model. The horizontal dashed line represents the vapor saturation curve. The increasing line (dashed line, then solid line) is the set of the CJ points for the vapor equation of state. The other function (solid line, then dashed line) is the set of the mixture CJ points. The solid lines of the curves correspond to the part in which they match with the equation of state used.*

We remark that P_{CJ} is a linear function of P_0. If we use the equation of the Crussard curve, we get the following expression for τ_{CJ}:

$$\tau_{CJ} = \tau_0 \frac{\frac{\gamma_2+1}{\gamma_2-1} P_0 + P_{CJ}}{\frac{\gamma_1+1}{\gamma_1-1} P_{CJ} + P_0}.$$

As P_{CJ} is a linear function of P_0, we see that τ_{CJ} is a linear function of τ_0.

The CJ point of pure vapor can be chosen only when $\tau_{CJ} \geq \tau_1$.

The two functions τ_{CJ}, for the vapor and mixture equations of state, are drawn in Figure 17, highlighting the fact that they cannot be linked continuously.

Example 2: The two stiffened gas model.

As already mentioned, the mixture equation of state cannot be computed when we deal with the two stiffened gas model. Therefore we can show only a numerical computation as an illustration. We chose the model of dodecane for which coefficients lie in Table 2. We begin on the point $P_0 = 900000\,\text{Pa}$ with a specific volume of $\tau_0 = 0.0025\,\text{kg.m}^{-3}$. We compute all the states $0^\star$ that can be linked with that initial point via an isentrope. If the state $0^\star$ is overheated (i.e., lies in the saturation dome), then we compute the CJ point(s) corresponding to a mixture downstream state and/or to a pure vapor downstream state. Numerical results are in Figure 18.

4. Conclusion. In section 1 we studied the model with two convex equations of state. In particular, we gave a necessary and sufficient condition for the convexity of the mixture equation of state resulting from an entropy optimization:

$$\frac{\mathrm{d}P_{\text{sat}}}{\mathrm{d}T} > 0.$$

Then we proposed to take into account metastable states in the solution of the Riemann problem. For that, we used the CJ theory. We first proved that this theory

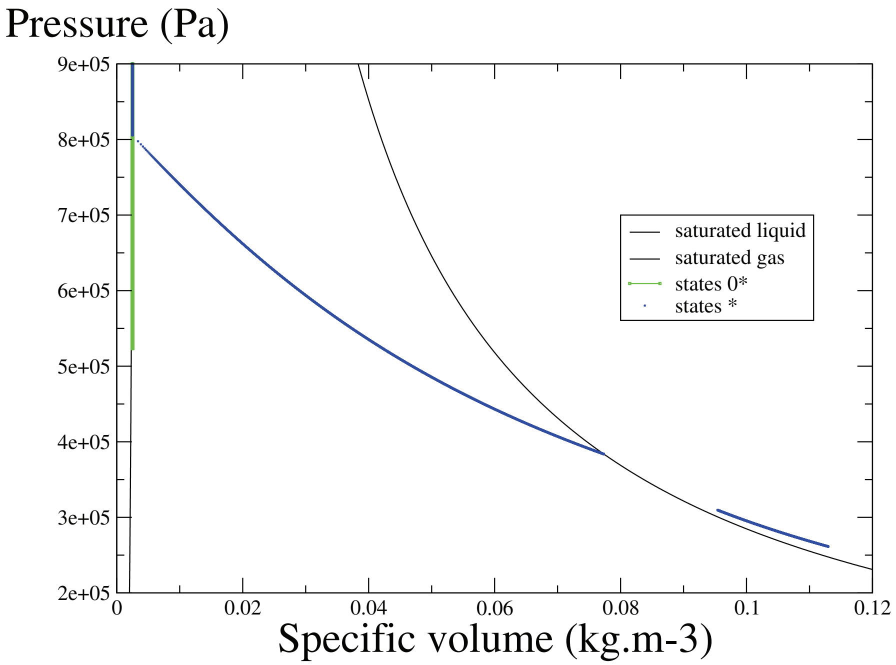


FIG. 18. *Set of all the CJ points that can be reached from a given point. The isentrope is drawn in green and is nearly vertical. In blue, the set of all the CJ points is drawn, showing a jump between the mixture CJ points and the pure vapor CJ point.*

can be applied. We emphasized the link between whether the metastable states are overheated or overcooled and the retrograde or regular behavior of the fluid. In one particular case, when γ/Γ does not depend on τ, the condition of regular behavior of the fluid is necessary and sufficient to ensure that the energy of a metastable liquid is lower than the energy of a mixture at thermodynamic equilibrium with the same pressure and specific volume.

For the entropy growth condition, we proved that it is ensured provided that $\gamma_l > \gamma_m$ and $\gamma_l > \gamma_v$.

The problem with the deflagration waves is that the Lax characteristic criterion is not ensured, so that the problem is underdeterminated. The only thing that we can state with no further hypothesis is that the set of all the downstream states lies in an area limited on the top by the set of all the constant pressure deflagrations, which is continuous, and limited below by the set of all the CJ points, which was proved to be discontinuous thanks to a detailed study of the behavior of the Crussard curve near the saturation curve. As the set of all the CJ points is discontinuous, the use of the CJ closure as in [10] for solving the Riemann problem leads to a solution that does not depend continuously on its initial data in general. A first step in finding a right kinetic closure would be, for example, to study traveling waves for a relaxation model as given in [3]. As we know that liquid-vapor phase transition is governed by a competition between relaxation phenomena and thermal conduction, it would be more relevant (but much harder) to study traveling waves with a relaxation model and thermal conductivity.

Acknowledgment. The author would like to thank Siegfried Müller from the Institut für Geometrie und Praktische Mathematik (Aachen, Germany) for profitable discussions on the Riemann problem with phase transition at thermodynamic equilibrium.

REFERENCES

[1] N. Bedjaoui and P. G. LeFloch, *Diffusive-dispersive traveling waves and kinetic relations.* I. *Nonconvex hyperbolic conservation laws*, J. Differential Equations, 178 (2002), pp. 574–607.

[2] S. Benzoni-Gavage, *Stability of subsonic planar phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 150 (1999), pp. 23–55.

[3] F. Caro, *Modélisation et simulation numérique des transitions de phase liquide-vapeur*, Ph.D. thesis, École Polytechnique, CEA Saclay, France, 2004.

[4] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience Publishers, New York, 1948.

[5] H. T. Fan, *A vanishing viscosity approach on the dynamics of phase transitions in van der Waals fluids*, J. Differential Equations, 103 (1993), pp. 179–204.

[6] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996.

[7] P. Helluy and T. Barberon, *Finite volume simulation of cavitating flows*, Comput. & Fluids, 34 (2005), pp. 832–858.

[8] S. Jaouen, *Étude mathématique et numérique de la stabilité pour des modèles hydrodynamiques avec transition de phase*, Ph.D. thesis, Université Pierre et Marie Curie, CEA Bruyères-le-Châtel, France, 2001.

[9] O. Le Métayer, J. Massoni, and R. Saurel, *Élaboration de lois d'état d'un liquide et de sa vapeur pour les modèles d'écoulements diphasiques*, Int. J. Thermal Sci., 43 (2003), pp. 265–276.

[10] O. Le Métayer, J. Massoni, and R. Saurel, *Modelling evaporation fronts with reactive Riemann solvers*, J. Comput. Phys., 205 (2005), pp. 567–610.

[11] T. P. Liu, *The Riemann problem for general systems of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.

[12] R. Menikoff and B. J. Plohr, *The Riemann problem for fluid flow of real materials*, Rev. Modern Phys., 61 (1989), pp. 75–130.

[13] J. R. Simões-Moreira and J. E. Shepherd, *Evaporation waves in superheated dodecane*, J. Fluid Mech., 382 (1999), pp. 63–86.

[14] M. Slemrod, *Dynamic phase transitions in a van der Waals fluid*, J. Differential Equations, 52 (1984), pp. 1–23.

[15] P. A. Thompson, *A fundamental derivative in gasdynamics*, Phys. Fluids, 14 (1971), pp. 1843–1849.

[16] P. A. Thompson, G. C. Carofano, and Y.-G. Kim, *Shock waves and phase changes in a large heat capacity fluid emerging from a tube*, J. Fluid. Mech., 166 (1986), pp. 57–92.

[17] P. A. Thompson, H. Chaves, G. E. A. Meier, Y.-G. Kim, and H. D. Speckman, *Wave splitting in a fluid of large heat capacity*, J. Fluid. Mech., 185 (1987), pp. 385–414.

[18] P. A. Thompson and K. Lambrakis, *Negative shock waves*, J. Fluid Mech., 60 (1973), pp. 187–208.

[19] A. Voss, *Exact Riemann Solution for the Euler Equations with Nonconvex and Nonsmooth Equation of State*, Ph.D. thesis, Rheinisch-Westfälischen Technischen Hochschule Aachen, Aachen, Germany, 2005; available online at http://www.it-voss.com/papers/thesis-voss-030205-128-final.pdf.

[20] B. Wendroff, *The Riemann problem for materials with nonconvex equations of state.* I. *Isentropic flow*, J. Math. Anal. Appl., 38 (1972), pp. 454–466.

[21] B. Wendroff, *The Riemann problem for materials with nonconvex equations of state.* II. *General flow*, J. Math. Anal. Appl., 38 (1972), pp. 640–658.

[22] Ya. B. Zel'dovich and Yu. P. Raizer, *Physics of Shock Waves and High Temperature Hydrodynamic Phenomena*, Vol. II, Academic Press, New York, London, 1967.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1360–1377

SELF-SIMILAR SOLUTIONS FOR THE TRIPLE POINT PARADOX IN GASDYNAMICS*

ALLEN M. TESDALL[†], RICHARD SANDERS[‡], AND BARBARA L. KEYFITZ[†]

Abstract. We present numerical solutions of a two-dimensional Riemann problem for the compressible Euler equations that describes the Mach reflection of weak shock waves. High resolution finite volume schemes are used to solve the equations formulated in self-similar variables. We use extreme local grid refinement to resolve the solution in the neighborhood of an apparent but mathematically inadmissible shock triple point. The solutions contain a complex structure: instead of three shocks meeting in a single standard triple point, there is a sequence of triple points and tiny supersonic patches behind the leading triple point, formed by the reflection of weak shocks and expansion waves between the sonic line and the Mach shock. An expansion fan originates at each triple point, resolving the von Neumann triple point paradox.

Key words. triple point paradox, von Neumann paradox, self-similar solutions

AMS subject classifications. 65M06, 35L65, 76L05

DOI. 10.1137/070698567

1. Introduction.

Consider a plane normal shock in an inviscid, compressible, perfect gas which hits a fixed wedge with half angle θ_w, as depicted in Figure 1. For a given upstream state with density $\rho = \rho_r$, velocity $\mathbf{u} = (u, v) = (0, 0)$, and pressure $p = p_r$, the fluid properties downstream of a *fast* (i.e., $u + c$) shock are given by

$$\frac{\rho_l}{\rho_r} = \frac{(\gamma+1)\,M^2}{2+(\gamma-1)\,M^2}, \qquad \frac{u_l}{c_r} = \frac{2}{\gamma+1}\left(M - \frac{1}{M}\right), \qquad \frac{p_l}{p_r} = \frac{2\gamma}{\gamma+1}M^2 - \frac{\gamma-1}{\gamma+1}, \tag{1.1}$$

where γ is the ratio of specific heats and $M > 1$ is the shock Mach number, defined as the shock speed given by the Rankine–Hugoniot relations divided by the upstream sound speed $c_r = \sqrt{\gamma p_r/\rho_r}$. Following interaction of the shock with the wedge wall, a number of self-similar reflection patterns are possible, depending on the values of M and θ_w. The simplest pattern is *regular reflection*, in which there is a single reflected shock, depicted in Figure 1(a). For small wedge angles or strong shocks, regular reflection is replaced by *Mach reflection.* The simplest type of Mach reflection is called single Mach reflection, in which the incident and reflected shocks move off the wedge and a single shock called the Mach stem extends down to the wall; see Figure 1(b). The point where the three shocks meet is called the triple point, and a contact discontinuity also originates there.

[*] Received by the editors July 30, 2007; accepted for publication (in revised form) January 2, 2008; published electronically April 23, 2008.

http://www.siam.org/journals/siap/68-5/69856.html

[†] Fields Institute, Toronto, ON M5T 3J1, Canada, and Department of Mathematics, University of Houston, Houston, TX 77204 (atesdall@fields.utoronto.ca, bkeyfitz@fields.utoronto.ca). The first author's research was supported by National Science Foundation grant DMS 03-06307, NSERC grant 312587-05, and the Fields Institute. The third author's research was supported by National Science Foundation grant DMS 03-06307, Department of Energy grant DE-FG02-03ER25575, and NSERC grant 312587-05.

[‡] Department of Mathematics, University of Houston, Houston, TX 77204 (sanders@math.uh.edu). This author's research was supported by National Science Foundation grant DMS 03-06307.

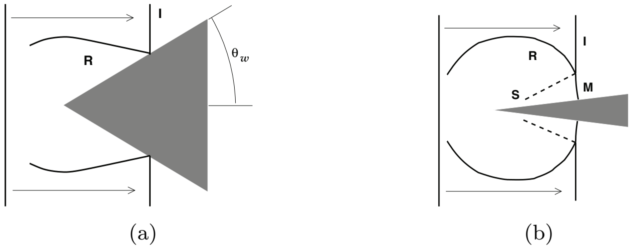


Fig. 1. *A planar shock moving from left to right impinges on a wedge. After contact, I indicates the incident shock and R indicates the reflected shock. Regular reflection is depicted in* (a) *and Mach reflection in* (b). *In* (b), *the dotted line S indicates a contact discontinuity and M is the Mach stem.*

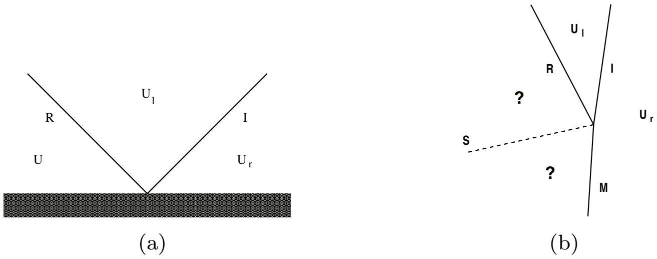


Fig. 2. *An enlargement of the incident and reflected shock intersections in Figure* 1. *Regular reflection is depicted in* (a) *and Mach reflection in* (b). *The constant states upstream and downstream of the incident shock are denoted by U_r and U_l. Whether or not constant states indicated by the question marks in* (b) *exist depends on the strength of I.*

The basic equations that describe regular and Mach reflection were formulated by von Neumann in 1943 [10] and are known as two-shock and three-shock theory, respectively. His analysis is based on the assumption that Mach reflection solutions can be locally approximated by constant states separated by plane discontinuities (this is also true, in a finite neighborhood, for flows that are supersonic at the reflection point; see Figure 2 for an illustration of this assumption). The oblique shock relations of gasdynamics connect the constant states. To explain transition between regular and Mach reflection, von Neumann suggested several criteria. For weak incident shocks he proposed that transition occurs at the detachment angle, which is a function of M. For a shock with a given M impinging on a wedge, when the wedge angle is larger than the detachment angle there are two regular reflection solutions, one with a strong reflected shock and one with a weak reflected shock (only the weak reflected shock solution is observed experimentally). At the detachment angle the two solutions coalesce, and for smaller wedge angles there is no regularly reflected solution. This is a possible condition for transition but has not been definitively established. Several criteria for transition can be obtained using von Neumann's approach; see Henderson [7] for a detailed discussion.

Good agreement between the von Neumann theory and experiment is obtained for regular and Mach reflection for a wide range of conditions. For sufficiently weak shocks, however, application of the three-shock theory indicates that triple point solutions such as those depicted in Figure 2(b) do not exist. Transition from regular to Mach reflection is impossible, and the theory is unable to predict what kind of reflection does occur for weak shocks reflecting off thin wedges. However, experiments in which a very weak shock reflects off a thin wedge appear to show a pattern of reflection containing three shocks meeting at a triple point. This discrepancy is referred to as the von Neumann, or triple point, paradox.

Over the years, a number of ways to resolve the paradox have been proposed. A singularity could occur in the solution behind the triple point [15] or in the reflected shock curvature at the triple point [14], so that a local approximation of the solution by plane waves separated by constant states is invalid; there could be an unobserved fourth shock at the triple point [6]; or the reflected shock could decay into a continuous wave before hitting the incident shock, so that there is no triple point [4]. In 1947 Guderley [5] proposed the existence of an expansion fan and a supersonic region behind the triple point in a steady weak shock Mach reflection (the same triple point paradox occurs in the case of steady flow as in unsteady reflection off a wedge). He demonstrated that one could construct local solutions consisting of three plane shocks, an expansion fan, and a contact discontinuity meeting at a point. However, despite intensive study, no evidence of an expansion fan or a supersonic patch was seen either in experiments (see, for example, [1, 11, 14]) or in numerical solutions (see [4, 2, 15]).

The first indication that Guderley's proposed resolution might be essentially correct was contained in numerical solutions of shock reflection problems for the unsteady transonic small disturbance equations (UTSDE) in [8] and the compressible Euler equations in [18]. Solutions containing a supersonic patch embedded in the subsonic flow directly behind the triple point in a weak shock Mach reflection were presented there. Subsequently, Zakharian et al. [19] found a supersonic region in a numerical solution of the compressible Euler equations. The supersonic region in all of these solutions is extremely small, explaining why it had never been observed before, experimentally or numerically.

The supersonic patches obtained in the solutions in [8, 18, 19] appeared to confirm Guderley's four wave solution. The patch indicates that it is plausible for an expansion fan to be an unobserved part of the observed three-shock confluence, since the flow must be supersonic for an expansion wave to occur. However, these solutions are not sufficiently well resolved to show the structure of the flow inside the supersonic region. In [16] numerical solutions were obtained of a problem for the UTSDE that describes the reflection of weak shocks off thin wedges, with the equations formulated in special self-similar variables. The advantage of using self-similar coordinates is that the point of interest remains fixed on the computational grid, and a steady self-similar solution is obtained by letting a pseudotime $t \to \infty$. In a parameter range for which regular reflection is impossible, the solutions contain a remarkably complex structure: there is a sequence of triple points and tiny supersonic patches immediately behind the leading triple point, formed by the reflection of weak shocks and expansion fans between the sonic line and the Mach shock. A centered expansion fan originates at each triple point. It was shown that the triple points with expansion fans observed numerically are consistent with theory and resolve the paradox. The term *Guderley Mach reflection* was chosen in [9] to name this new reflection pattern.

Following the detection of Guderley Mach reflection in [16], a problem for the nonlinear wave system that is analogous to the reflection of weak shocks off thin wedges was studied numerically in [17]. The nonlinear wave system is a simple 3×3 hyperbolic system which resembles the Euler equations, but is not obtained from them via a limit, and which has no known physical relevance. It is obtained from the isentropic Euler equations by dropping the momentum transport terms from the momentum equations, and it has a characteristic structure similar to that of the compressible

Euler equations: nonlinear acoustic waves coupled (weakly) with linearly degenerate waves. In a parameter range where regular reflection is not possible, a numerical solution of this system formulated in self-similar variables was obtained which again contains a sequence of triple points in a tiny region behind the leading triple point, with a centered expansion fan originating at each triple point. This solution is very similar in pattern to those obtained for the UTSDE. The discovery of Guderley Mach reflection in a solution of this system leads one to expect that a sequence of supersonic patches and triple points is a generic feature of two-dimensional Riemann problems for some class of hyperbolic conservation laws. This class is possibly characterized by "acoustic waves," as defined in [3]. The compressible Euler system for gasdynamics is another member of this class, suggesting that weak shock solutions of the Euler equations—the subject of the present work—would contain Guderley Mach reflection solutions as well.

The numerical solutions in [8, 18, 19] were obtained by solving an initial value problem for the unsteady equations. The problem of inviscid shock reflection at a wedge is self-similar, and there are advantages to solving the problem in self-similar, rather than unsteady, variables. In the unsteady formulation local grid refinement near the triple point is difficult, because any waves which are present initially move through the numerical domain, requiring the refined region to move as well. Solutions of the self-similar equations are stationary, making local grid refinement easier to implement. Also, in self-similar variables a global grid continuation procedure can be used in which a partially converged solution on a coarse grid is interpolated onto a fine grid and then driven to convergence on the fine grid. Procedures for solving the UTSDE in self-similar variables were developed in [16] and extended to apply to the nonlinear wave system in [17]. The procedures used to solve the nonlinear wave system have been applied, with only slight modification, in the present work to obtain solutions of the full Euler system.

In this paper we present high resolution numerical solutions of the shock reflection problem for the full Euler equations computed in self-similar coordinates. Our most highly resolved solution shows that Guderley Mach reflection occurs at a set of parameter values where Mach reflection is impossible: there is a sequence of tiny supersonic patches and triple points behind the leading triple point in a weak shock Mach reflection. This numerical solution is remarkably similar to those obtained for the UTSDE in [16] and for the nonlinear wave system in [17].

Experimental confirmation of these results is challenging simply because the computed structure is so small and weak. Nevertheless, recent experimental evidence appears to confirm that Guderley Mach reflection occurs when a weak shock reflects off a thin wedge. Skews and Ashworth in [12] modified an existing shock tube in order to obtain Mach stem lengths more than an order of magnitude larger than those possible with conventional shock tubes. They present photographic images of shock reflection experiments that clearly show an expansion wave behind the triple point in a weak shock Mach reflection, a terminating shocklet, and evidence of a second terminating shocklet. The supersonic region is extremely small, as predicted by the computations in [16, 17] and the present work. Further experimental improvements and data acquisition are underway.

This paper is organized as follows. In section 2 we describe the shock reflection problem for the full Euler equations. In section 3 we discuss our approach to solving this problem numerically. The numerical results obtained are presented in section 4. In section 5 we discuss questions raised by our results. Finally, we summarize our findings in section 6.

2. The shock reflection problem for the Euler equations.

We consider a problem for the full Euler equations that describes the reflection of a shock wave off a wedge. The shock reflection problem consists of the compressible Euler equations,

$$\begin{aligned}\rho_t + (\rho u)_x + (\rho v)_y &= 0,\\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y &= 0,\\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y &= 0,\\ (\rho e)_t + ((\rho e + p)u)_x + ((\rho e + p)v)_y &= 0,\end{aligned} \tag{2.1}$$

with piecewise constant Riemann data consisting of two states separated by a discontinuity located at $x = 0$. Here, ρ is the density, u and v are the x and y components of velocity, respectively, p is the pressure, and e is the energy. We use an ideal gas equation of state,

$$p = (\gamma - 1)\rho\left(e - \frac{1}{2}(u^2 + v^2)\right),$$

where the ratio of specific heats γ is taken to be 1.4. The initial data correspond to a vertical plane shock hitting the corner of a wedge at $t = 0$. Letting $U(x, y, t) = (\rho, u, v, p)$,

$$U(x, y, 0) = \begin{cases} U_R \equiv (\rho_R, 0, 0, p_R) & \text{if } x > 0,\\ U_L \equiv (\rho_L, u_L, 0, p_L) & \text{if } x < 0,\end{cases} \tag{2.2}$$

where the left- and right-hand states are connected by the Rankine–Hugoniot jump conditions for a shock with Mach number M. The boundary condition on the wedge wall is

$$\mathbf{u} \cdot \mathbf{n} = 0, \tag{2.3}$$

where $\mathbf{u} = (u, v)$ and $\mathbf{n}$ is the unit normal vector at the wall. The shock propagates to the right into stationary gas with speed Mc, where c is the sound speed in the fluid ahead of the shock at state U_R. The location of the incident shock is given by

$$x = (Mc)t. \tag{2.4}$$

For a gas with a given equation of state, there are two parameters in the shock reflection problem: the wedge angle θ and the strength of the incident shock, which we parameterize by the shock Mach number M. For sufficiently small Mach numbers and small wedge angles, neither Mach reflection nor regular reflection solutions exist (see [7, 10]).

3. The numerical method.

The problem (2.1)–(2.3) is self-similar, so the solution depends only on the similarity variables

$$\xi = \frac{x}{t}, \qquad \eta = \frac{y}{t}.$$

We write (2.1) in the form

$$U_t + F_x + G_y = 0, \tag{3.1}$$

where

$$U = (\rho, \rho u, \rho v, \rho e), \quad F = (\rho u, \rho u^2 + p, \rho uv, \rho ue + up),$$
$$\text{and} \quad G = (\rho v, \rho uv, \rho v^2 + p, \rho ve + vp).$$

Writing (3.1) in terms of ξ, η, and a pseudotime variable $\tau = \log t$, we obtain

$$U_\tau - \xi U_\xi - \eta U_\eta + F_\xi + G_\eta = 0. \tag{3.2}$$

As $\tau \to +\infty$, solutions of (3.2) converge to a pseudosteady, self-similar solution that satisfies

$$-\xi U_\xi - \eta U_\eta + F_\xi + G_\eta = 0. \tag{3.3}$$

Equation (3.3) is hyperbolic when $c^2 < \bar{u}^2 + \bar{v}^2$, corresponding to supersonic flow in a self-similar coordinate frame, and of mixed type when $c^2 > \bar{u}^2 + \bar{v}^2$, corresponding to subsonic flow, where $\bar{u} = u - \xi, \bar{v} = v - \eta$. Here, $c = \sqrt{\gamma p/\rho}$ is the local sound speed. The sonic line is given by

$$\bar{u}^2 + \bar{v}^2 = c^2. \tag{3.4}$$

By abuse of notation, we have referred to the locus of transition points between $c^2 < \bar{u}^2 + \bar{v}^2$ and $c^2 > \bar{u}^2 + \bar{v}^2$ as the sonic line, whether the flow is continuous there or not. We define a local self-similar Mach number $\bar{M}^2 = \frac{\bar{u}^2+\bar{v}^2}{c^2}$. When $\bar{M} > 1$, the flow is supersonic, and when $\bar{M} < 1$, the flow is subsonic.

In order to solve (3.2) numerically, we write it in conservative form as

$$U_\tau + (F - \xi U)_\xi + (G - \eta U)_\eta + 2U = 0. \tag{3.5}$$

In these self-similar variables, the full Euler system has the form of the unsteady equations (3.1) with modified flux functions and a lower-order source term.

An essential feature of our numerical method is the use of local grid refinement in the area of the apparent triple point. We designed a sequence of successively refined, nonuniform, logically rectangular finite volume grids. See Figure 3 for a diagram of the computational domain. We use grid continuation, in which partially converged coarse grid solutions are interpolated onto more refined grids and converged on the refined grids. For each grid, inside a given box surrounding the triple point, uniform grid spacing is used. Outside of this box, the grid is exponentially stretched in both grid directions.

The basic finite volume scheme is quite standard. Each grid cell, Ω, is a quadrilateral, and using $\vec{\nu} = (\nu_\xi, \nu_\eta)$ to denote the normal vector to a typical side of Ω, numerical fluxes are designed to be consistent with

$$\widetilde{F}(U) = (F(U) - \xi U)\,\nu_\xi + (G(U) - \eta U)\,\nu_\eta = \begin{pmatrix} \nu_\xi \rho u + \nu_\eta \rho v - \bar{\xi}\,\rho \\ \nu_\xi(\rho u^2 + p) + \nu_\eta \rho u v - \bar{\xi}\,\rho u \\ \nu_\xi \rho u v + \nu_\eta(\rho v^2 + p) - \bar{\xi}\,\rho v \\ \nu_\xi(\rho u e + u p) + \nu_\eta(\rho v e + v p) - \bar{\xi}\,\rho e \end{pmatrix},$$

where $\bar{\xi} = (\vec{\xi} \cdot \vec{\nu})$ and $\vec{\xi} = (\xi, \eta)$. Since $\vec{\xi}$ varies, our numerical flux formulae evaluate $\vec{\xi}$ frozen at the midpoint of each cell side. We use essentially the same numerical scheme as in [17], a high-order scheme based on the Roe numerical flux. High-order accuracy is achieved by using piecewise quadratic reconstruction limited in characteristic variables, together with the Roe flux

$$H_{Roe} = \frac{1}{2}(\widetilde{F}(U_l) + \widetilde{F}(U_r) - R\Lambda L\,(U_r - U_l)),$$

where $\Lambda = \mathrm{diag}(|-\bar{\xi} - c|, |-\bar{\xi}|, |-\bar{\xi} + c|)$, and R and L are the matrices of right and left eigenvectors to the Jacobian of $\widetilde{F}$. As in [17], we simplify the Roe approach by

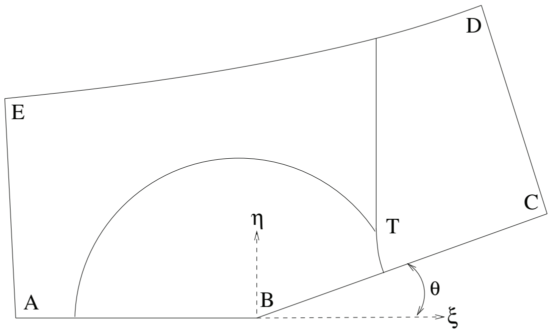


FIG. 3. *A schematic diagram of the computational domain. ABC is the wall; BC is the wedge, with angle θ. CDEA is the far field numerical boundary. The incident shock is perpendicular to AB. The incident (above T), reflected (left of T), and Mach (below T) shocks meet at the triple point T.*

evaluating R and L at the midpoint $U_{Roe} = \frac{1}{2}(U_l + U_r)$, which for the Euler equations is only an approximation to the Roe average. To avoid spurious expansion shocks, artificial dissipation on the order of $|U_r - U_l|$ is appended to the diagonal part of the Roe dissipation matrix in a field-by-field manner.

Time integration is accomplished using Heun's method, which can be written in two step predictor-corrector form (using overbars to denote predicted values) as

$$\begin{aligned}\frac{U^{\overline{n+1}} - U^n}{\Delta\tau} + \frac{1}{|\Omega|}\int_{\partial\Omega} H^n_{Roe}\, ds + 2U^n &= 0,\\ \frac{2U^{n+1} - U^{\overline{n+1}} - U^n}{\Delta\tau} + \frac{1}{|\Omega|}\int_{\partial\Omega} H^{\overline{n+1}}_{Roe}\, ds + 2U^{\overline{n+1}} &= 0.\end{aligned}$$

3.1. The grid and boundary conditions. We computed solutions of the problem (2.1)–(2.3) in the finite computational domain shown schematically in Figure 3. We use a nonuniform grid with a locally refined area of uniformly spaced grid very close to the triple point, as illustrated in Figure 4. The grid is defined by a conformal map of the form $z = w^\alpha$, so it is orthogonal with a singularity at the ramp apex $x = y = 0$. The refined uniform grid area is so small that it is obscured in the main plot shown in Figure 4. The inset plots show enlargements of the grid in the indicated rectangular regions, and the smaller inset plot contains a small superimposed box which delineates the refined uniform grid region. The grid is stretched exponentially from the edge of the uniform grid region to the outer numerical boundaries and the wall, with a stretching factor of 1%.

We use a sequence of such grids, with each grid corresponding to a level of grid refinement. The uniform grid region of each grid is refined by a factor of two in both x/t and y/t in relation to the uniform grid region of the previous grid. We obtain solutions on coarse grids, interpolate these onto more refined grids, and converge the solutions on the refined grids. We repeat this process until no further change is observed in the solution near the apparent triple point and *grid continue* to a steady state. This process is illustrated in Figure 5, which shows a coarse grid (dashed lines) overlaid with a refined grid (solid lines) in the uniform grid region of both grids. A solution is obtained on the coarse grid, and the computation on the coarse grid is stopped. This solution is interpolated onto the refined grid, and the computation

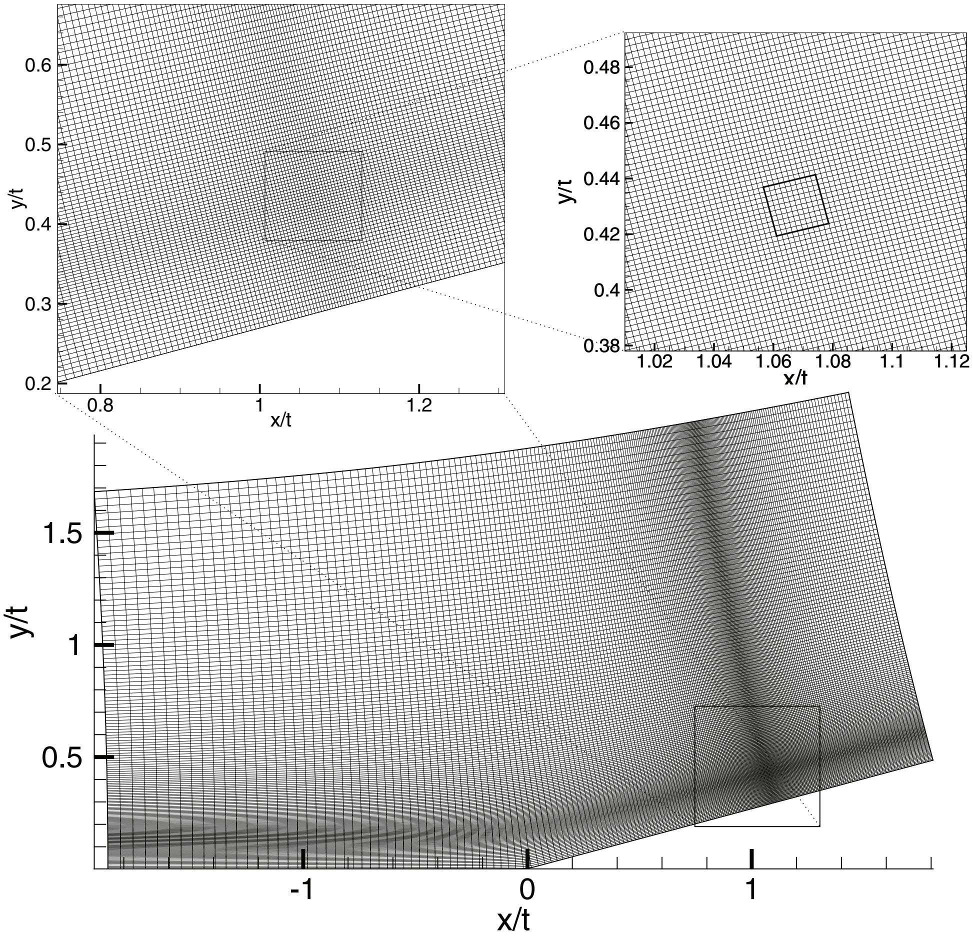


FIG. 4. *The grid structure, illustrating the region of uniform local refinement, which is outlined by the small box superimposed on the grid in the inset plot at the upper right. Outside of this small box, the grid is stretched. The main plot shows the entire grid. The locally refined region in this very coarse grid contains* 10×10 *grid cells, with* $\Delta\xi = \Delta\eta = 0.002$. *The locally refined region in our most refined grid is shown in Figure* 7(a).

is resumed on the refined grid. We found that bilinear interpolation gave the best results, while higher-order methods such as biquadratic interpolation resulted in large overshoots at the shocks.

Every grid in the sequence is designed so that the refined uniform grid region surrounds the apparent triple point as it appears in the currently available solution (the solution obtained with the previous grid). As the grids are refined and the shocks become better resolved, the triple point location can be determined more precisely, and the refined grid area can be repositioned and reduced in size. In fact, the refined uniform grid region depicted in the coarse grid shown in Figure 4 is more than 1000 times as large as the refined uniform grid region in our finest grid. The total number of grid cells in our finest grid is approximately six million, of which $300 \times 1000 = 3 \times 10^5$ ($\Delta\xi = \Delta\eta = 1 \times 10^{-6}$) are devoted to the local refinement.

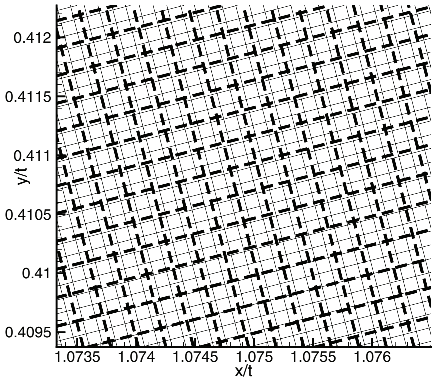


FIG. 5. *A solution on a coarse grid such as the one depicted in Figure* 4 *is interpolated onto a refined grid and converged on the refined grid. The locally refined region of the coarse grid (dashed lines) in this example has* $\Delta\xi = \Delta\eta = 0.00025$*, and the fine grid (solid lines) has* $\Delta\xi = \Delta\eta = 0.000125$*. The region shown in the plot is in the locally refined region of both grids.*

On the wall boundary ABC in Figure 3 we impose reflecting boundary conditions, equivalent to the physical no-flow condition (2.3). In addition, we require numerical boundary conditions on the outer computational boundaries, which we determine as follows.

The incident shock location (2.4) in self-similar variables is

$$\xi = Mc,$$

where c is the sound speed ahead of the shock. Boundary data on the left, right, and top are given to exactly agree with this shock, so that

$$U(\xi, \eta) = \begin{cases} U_R, & \xi > Mc, \\ U_L, & \xi < Mc, \end{cases} \tag{3.6}$$

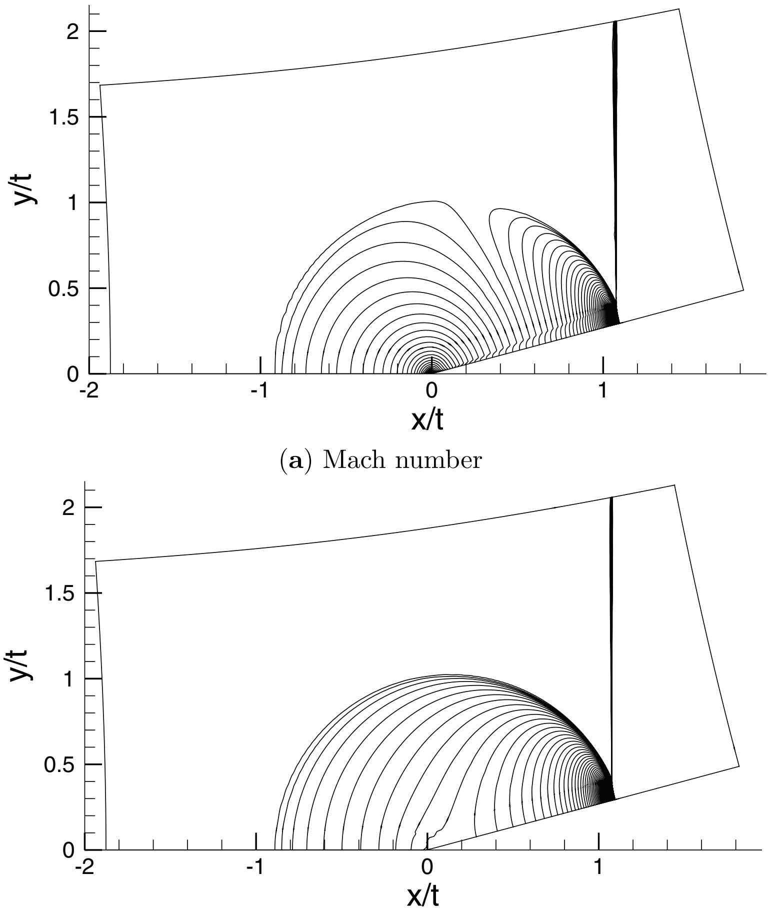
where the fluid properties U_L behind the shock are obtained from the Rankine–Hugoniot conditions. We use (3.6) as a boundary condition for (3.5) on $CDEA$.

4. Numerical results. We computed numerical solutions of (2.1)–(2.3) for a shock Mach number M equal to 1.075 and a wedge angle θ equal to 15 degrees. These data correspond to parameter $a \approx 1/2$ in the UTSDE model used in [16]. This problem is well outside the range for which regular reflection can occur. However, Mach reflection is also not possible for shocks this weak, and so this example illustrates a classic triple point paradox. In our computations we used $\rho_R = 1.4$ and $p_R = 1$ in (2.2) and determined the values U_L behind the shock from the Rankine–Hugoniot conditions. Table 4.1 gives the initial values of the fluid variables. We give our finest grid results in the plots which follow. Figure 6(a)–(b) shows a numerical solution that gives an overall picture of the shock reflection. The plots in (a) and (b) show Mach number and pressure contours, respectively, as functions of $(x/t, y/t)$. Here, we refer to the local Mach number of the solution, not to the shock Mach number M. The numerical solution appears to show a simple Mach reflection, with three shocks meeting at a triple point. Ahead of the incident shock, the pressure is equal to 1 and

Table 4.1

Left and right states for an incident shock with Mach number $M = 1.075$ and $\gamma = 1.4$.

	ρ	u	v	p
Right	1.4	0	0	1
Left	1.57697	0.12064	0	1.18156



(**a**) Mach number

(**b**) Pressure

Fig. 6. *Contour plots over the full numerical domain, showing what appears to be a Mach reflection—three shocks and a contact discontinuity (visible as a jump in the Mach contours in* (a)*) meeting in a point. The Mach contour spacing in* (a) *and the pressure contour spacing in* (b) *are both* 0.002. *The full grid contains* 2250×2710 *finite volume cells.*

the Mach number is 0. As the shock moves it induces a flow in the fluid behind it and a corresponding increase in pressure. The reflected shock is much weaker than the incident shock and decreases in strength as it moves away from the apparent triple point. The Mach shock increases in strength as it moves away from the triple point, reaching a maximum at the wall where it becomes a normal shock. Here, the pressure and the induced flow velocity are largest. A very weak contact discontinuity can be seen in the Mach contours in Figure 6(a). This is not visible in the plot in (b) because pressure does not jump across a contact discontinuity.

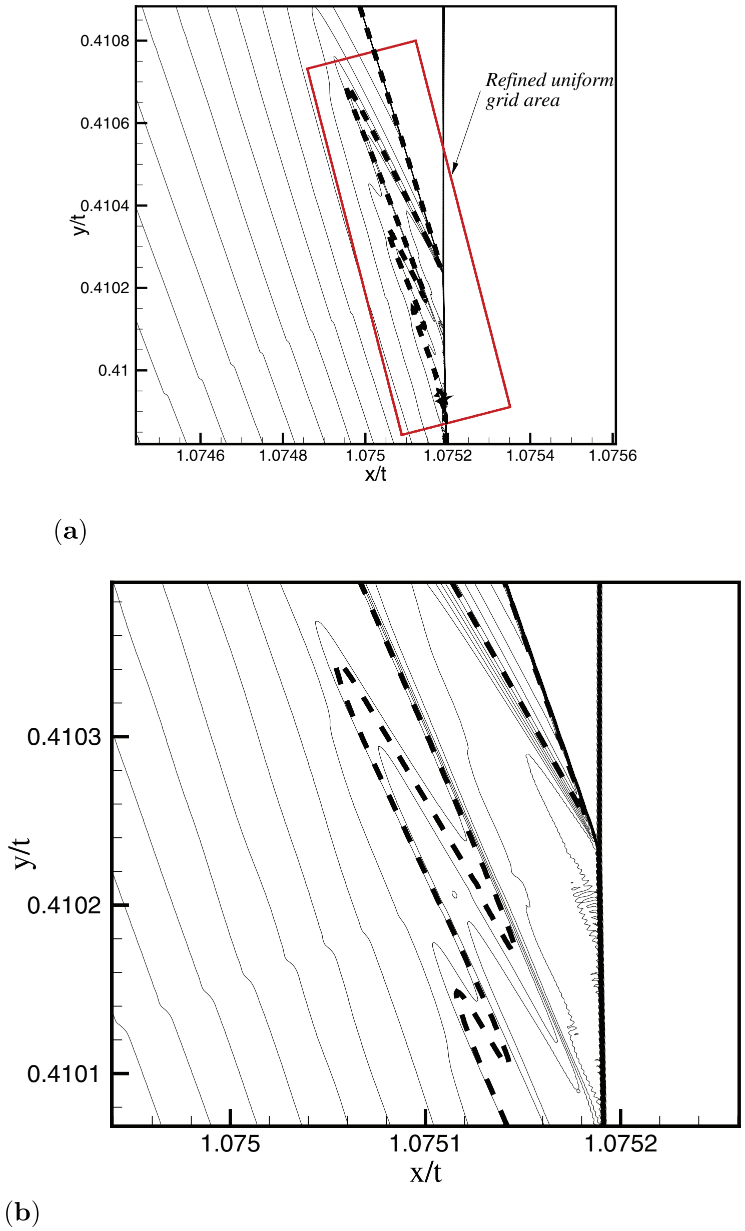


Fig. 7. *The solution near the triple point for* $M = 1.075$ *and* $\theta = 15$ *degrees. The Mach contour spacing is* 0.001 *in* (a) *and* 0.0005 *in* (b). *The dashed line in both plots is the sonic line. The refined uniform grid is contained within the box shown in* (a) *and has* 300×1000 *cells* ($\Delta\xi = \Delta\eta = 1 \times 10^{-6}$). *Two reflected shock/expansion wave pairs are clearly visible, with indications of a third. A contact discontinuity appears as a very weak jump in Mach number.*

In Figure 7(a), we show Mach number contours in the most refined region near the apparent triple point. The refined uniform grid, as indicated, is approximately aligned with the reflected shock. The dashed line in the figure is the numerically

Table 4.2

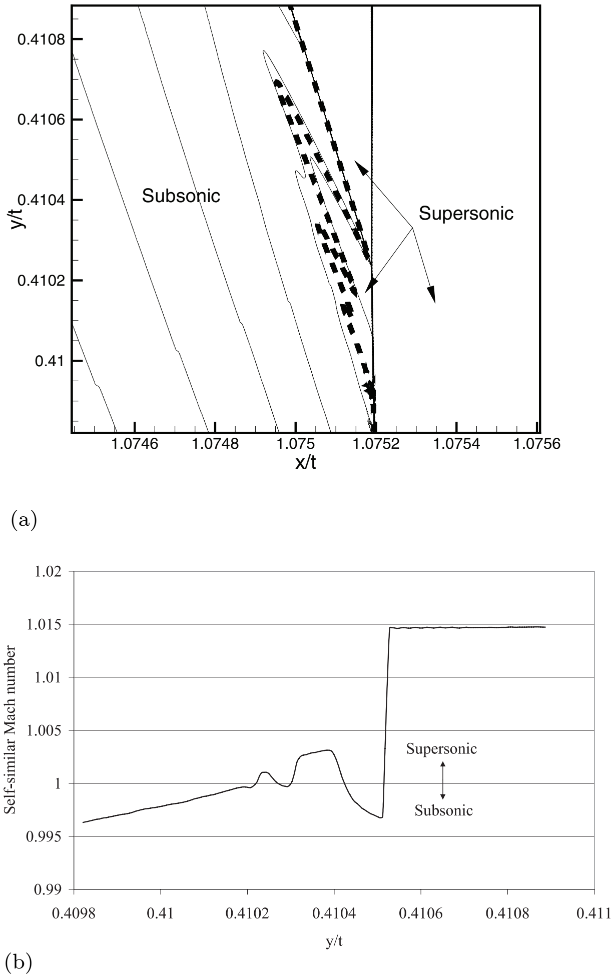
Approximate values of the reflected shock strengths for the three reflected shocks visible in Figure 7, beginning with the leading reflected shock, from the numerical data. For each shock, ρ_1 and ρ_0 denote the approximate values of ρ ahead of and behind the shock, respectively.

Shock	ρ_1	ρ_0	$[\rho]$
1	1.577	1.596	0.019
2	1.592	1.596	0.004
3	1.594	1.596	0.002

computed location of the sonic line, (3.4). (This sonic line is displayed more clearly in Figure 8(a).) Flow to the right of this line is supersonic, and the figure shows that the solution contains a small region of supersonic flow behind the triple point. There is an expansion fan centered at the leading triple point, but it cannot be seen clearly at this level of magnification. To show this expansion fan more clearly, in Figure 7(b) we show an enlargement of the solution, using more closely spaced contours, in a tiny region near the confluence of the incident, reflected, and Mach shocks. Behind the leading triple point, there is a sequence of very weak shocks that intersect the Mach shock, forming a sequence of triple points, with a very weak expansion fan centered at each triple point. Each shock-expansion wave pair in the sequence is smaller and weaker than the one preceding it. Three reflected shocks appear to be visible in the plots in Figure 7(a)–(b). Their approximate strengths, beginning with the leading reflected shock, are given in Table 4.2. The jump $[\rho]$ in ρ across a reflected shock is measured near the point where the flow behind the shock is sonic. This point is very close to the corresponding triple point on the Mach shock, as shown in Figure 7. In principle, a contact discontinuity originates at each triple point. However, the only contact discontinuity that is strong enough to be resolved numerically is the one at the leading triple point.

To depict the regions of supersonic and subsonic flow in a Guderley Mach reflection, we plot widely spaced Mach contours and the sonic line near the triple point in Figure 8(a). In the plot in (b), we give a cross section of Mach number $\bar{M}$ taken vertically through the region shown in the plot in (a), at a location slightly to the left of the Mach shock. The height $\Delta(y/t)$ of the supersonic region behind the triple point is approximately 0.00075, and the width $\Delta(x/t)$ is approximately 0.0001. Here, the height $\Delta(y/t)$ is a numerical estimate of the difference between the maximum value of y/t on the sonic line and the minimum value of y/t at the rear sonic point on the Mach shock. The width $\Delta(x/t)$ is an estimate of the width of the supersonic region at the value of y/t corresponding to the leading triple point. The height of the supersonic region is approximately 0.6% of the length of the Mach shock.

We found that a certain minimum grid resolution was necessary to resolve the supersonic region behind the triple point. As we refined the grid beyond this minimum resolution, a detailed flowfield structure became visible in the supersonic region. Figure 9 shows Mach number contours for a sequence of solutions computed on successively refined grids. In Figure 9(a)–(b), the sonic line appears fairly smooth. The supersonic patch appears to be shock-free. After two further grid refinements, each by a factor of two in both x/t and y/t (Figure 9(c)), a shock is visible behind the leading triple point. Our finest grid solution is shown in the plot in Figure 9(d). Two shocks are visible behind the leading triple point. Further refinement of the grid resulted in almost no observable change in the solution, as shown in the plot in Figure 10, an indication of grid convergence. At resolutions lower than the one shown in Figure 9(a), the supersonic region disappears entirely, and the sonic line runs down the inside of



(a)

(b)

Fig. 8. *The supersonic and subsonic regions near the triple point. The dashed line in* (a) *is the sonic line. It delineates the supersonic patch within the subsonic zone behind the triple point; the Mach contour spacing is* 0.0025. *In* (b) *a vertical cross section of $\bar{M}$ is taken at the location $x/t = 1.0751$, slightly to the left of the incident shock/Mach stem. The large jump is the leading reflected shock. Note the crossings at $\bar{M} = 1$, indicating jumps across weak reflected shocks or smooth transitions across the sonic line.*

the reflected shock, through the triple point, and down the Mach shock.

Figure 11 illustrates the size and location of the region where extreme local grid refinement is performed. The refined grid area is too small to be visible in the main plot shown in Figure 11. The inset figures show enlargements of the solution contained within the small rectangular box centered about the apparent triple point, as indicated. The solution shown in the smaller inset figure also contains a small box centered at the apparent triple point, indicating the approximate size and location of the region shown in Figures 7(a) and 8(a).

To further explore the wedge angle–shock strength parameter range in which the triple point paradox occurs, we also computed a solution of the shock reflection

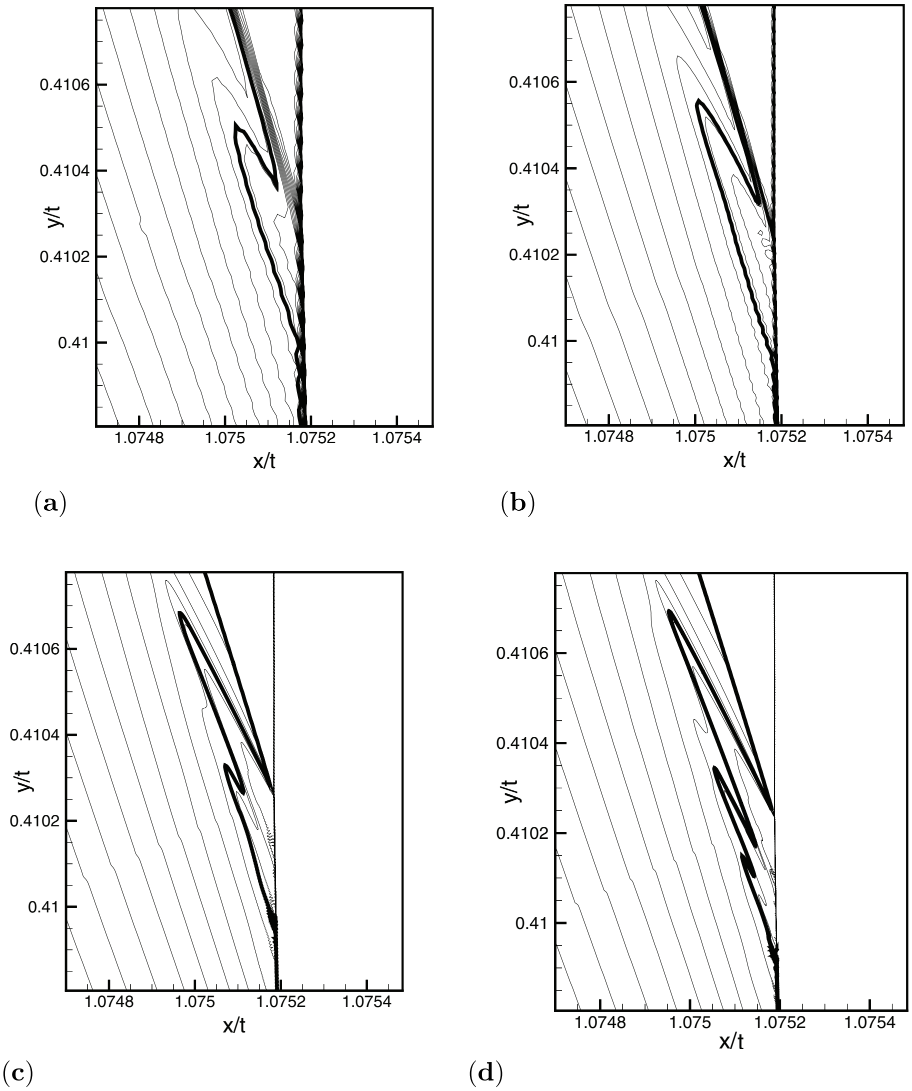


Fig. 9. *A sequence of contour plots illustrating the effect of increasing grid resolution on the numerical solution. The figures show Mach contours in the refined grid area near the triple point, with a Mach contour spacing of* 0.001. *The heavy line is the sonic line. The mesh size used in the refined uniform grid area is* $\Delta\xi = \Delta\eta = 1.6 \times 10^{-5}$ *in* (a), $\Delta\xi = \Delta\eta = 8 \times 10^{-6}$ *in* (b), $\Delta\xi = \Delta\eta = 2 \times 10^{-6}$ *in* (c), *and* $\Delta\xi = \Delta\eta = 1 \times 10^{-6}$ *in* (d). *The area of the refined uniform grid in* (c) *and* (d) *is depicted in Figure* 7(a)*; the refined uniform grids in* (a) *and* (b) *are slightly larger than the region shown. In* (a), *the refined uniform grid contains* 64×64 *grid cells. A supersonic region is visible as a bump in the sonic line, but it is poorly resolved. In* (b), *the refined uniform grid contains* 128×128 *grid cells. The supersonic region appears to be smooth. In* (c), *the refined uniform grid area contains* 150×500 *grid cells. There is a shock wave behind the leading triple point. In* (d), *the refined uniform grid area contains* 300×1000 *grid cells. Two shock waves are visible behind the leading triple point. The result of further refinement of the grid in* (d) *is shown in Figure* 10.

problem with M equal to 1.04, wedge angle θ equal to 11.5 degrees, and ratio of specific heats γ equal to 5/3. This choice of γ corresponds to shock reflection in a

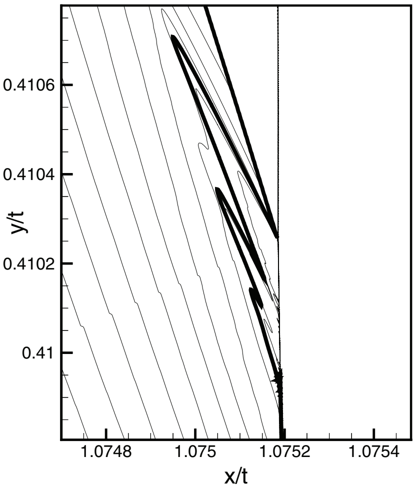


FIG. 10. *Additional refinement of the grid used to obtain the solution in Figures* 7, 8, and 9(d) *by a factor of two in both x/t and y/t* ($\Delta\xi = \Delta\eta = 5 \times 10^{-7}$) *results in little change in the solution near the triple point. The Mach contours are plotted at the same levels of Mach number as the plots in Figure* 9, *and the size of the region shown is the same. The heavy line is the sonic line.*

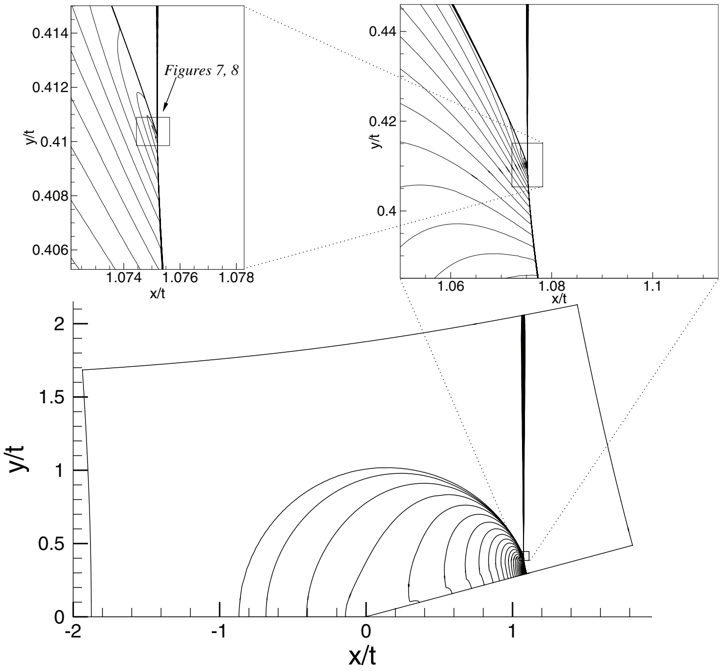


FIG. 11. *An illustration of the approximate size and location of the region shown in the plots in Figures* 7(a) *and* 8(a)*, which is contained in the small rectangular box shown in the smallest inset figure. The plot shows contour lines of* ρ *(density).*

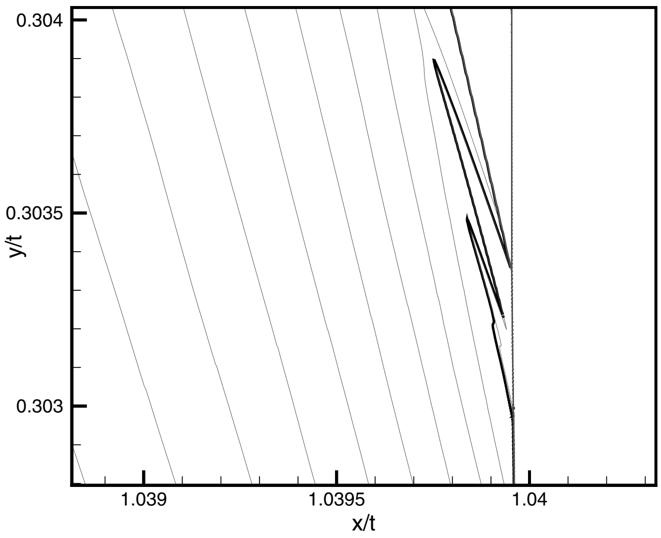


FIG. 12. *A contour plot of Mach number near the apparent triple point for* $M = 1.04$*,* $\theta = 11.5$ *degrees, and* $\gamma = 5/3$*. The heavy line is the sonic line. The number of grid points in the full grid is approximately* 11 *million, of which* $800 \times 2000 = 1.6$ *million (*$\Delta\xi = \Delta\eta = 5 \times 10^{-7}$*) are devoted to the local refinement. The refined uniform grid is contained in the region shown in the plot.*

monatomic gas ($\gamma = 1.4$ corresponds to a diatomic gas such as air). These data again correspond to parameter $a \approx 1/2$ in the UTSDE model used in [16]. Figure 12 shows Mach number contours and the sonic line in the neighborhood of the apparent triple point. Just as in our solution for $M = 1.075$, $\theta = 15$ degrees, and $\gamma = 1.4$, there is a sequence of triple points, reflected shocks, and expansion fans behind the leading triple point. Two reflected shock/expansion wave pairs are evident from the shape of the sonic line, with a slight indication of a third. At this lower shock strength, twice the grid refinement in both directions was required to obtain a solution comparable to that obtained for a $M = 1.075$ incident shock. The incident, reflected, and Mach shocks at the leading triple point are so weak that no contact discontinuity is visible.

5. Discussion.

These numerical results display a structure that is remarkably similar to the solutions of the shock reflection problem for the UTSDE model in [16] (compare Figures 7(a) and (b) with Figures 5 and 6 of [16], for example) and to its analogue for the nonlinear wave system in [17] (see Figure 6, p. 331). In all three cases, a weak shock reflection in a parameter range where regular reflection is impossible results in a sequence of triple points and supersonic patches in a tiny region behind the leading triple point, with an expansion fan originating at each triple point. The results presented here appear to confirm the validity of the UTSDE as a model for weak shock reflection. In addition, it now appears that this solution structure may occur generically in a class of conservation laws that includes the physically important Euler equations of gasdynamics.

An important feature of the numerical solution is the small size of the supersonic region. In our solution for a shock with $M = 1.075$ impinging on a 15 degree ramp, the height of the supersonic region is approximately 0.6% of the length of the Mach shock. This can be compared to the results in [16]. The UTSDE model used there depends on a single order-one transonic similarity parameter $a = \theta/\sqrt{2(M^2 - 1)}$. Solutions were obtained over a range of values of a, and the supersonic regions found in the solutions varied in height from approximately 0.05% to 3% of the length of the Mach shock, depending on the value of a. Our $M = 1.075/15°$ problem corresponds to $a = 0.5$, a value for which the height of the supersonic region in [16] was approximately 1.9% of the height of the Mach shock, somewhat larger than the region obtained in the present work. The UTSDE are an asymptotic reduction of the Euler equations in the limit of weak shocks and thin wedges (see [8]), and we do not expect exact agreement between solutions of the problem for the asymptotic equations and the problem for the Euler equations. For wedge angles closer to 0 and Mach numbers closer to 1, we would expect closer agreement with the solutions in [16]. The computation we performed with $M = 1.04$, $\theta = 11.5$ degrees displayed in Figure 12 serves as a check of this statement. The height of the supersonic region in our solution for this choice of parameters is approximately 1% of the length of the Mach shock, indeed closer to the figure of 1.9% obtained in [16].

The supersonic regions behind the triple point in our solutions for $M = 1.075$, $\theta = 15°$ and $M = 1.04$, $\theta = 11.5°$ are much larger in height than in width. Defining an aspect ratio $\Delta(y/t)/\Delta(x/t)$, these solutions have aspect ratios of approximately 8:1 and 9:1, respectively. This quantity agrees closely with the solutions in [16]. There, solutions obtained over a range of values of a contain supersonic regions with aspect ratios from approximately 2.75:1 to 8.5:1. The solution in [16] with $a = 1/2$ has an aspect ratio of 8:1, approximately the same as the solutions presented here, which correspond to this value of a. Although the supersonic regions obtained in the present work are smaller in size than the one obtained in [16] for $a = 1/2$, the shape of the regions obtained agrees quite closely.

Table 4.2 gives an indication of how the reflected shock strength decays in the sequence of shocks/expansions which comprise a Guderley Mach reflection. From the table, the strength of the first three reflected shocks is approximately in the ratio $9.5 : 2 : 1$. This is quite similar to the UTSDE result for $a = 1/2$ in [16]. There, four reflected shocks were visible in the solution; the strengths of the first three were in the approximate ratio $8 : 2 : 1$. For the nonlinear wave system solution in [17], the ratio was approximately $12 : 3 : 1$. We do not know precisely how the sequence of supersonic patches and shocks/expansions in Guderley Mach reflection decreases in size and strength, respectively, nor do we know if the sequence is finite or infinite.

The experimental results of Skews and Ashworth in [12] appear to confirm the existence of the Guderley Mach reflection structure reported here. The experiments were carried out on a 15° ramp with incident shock Mach numbers ranging from 1.05 to 1.1. The size of the expansion wave and terminating shocklet which were observed behind the leading triple point in experiments with measured Mach numbers of $M = 1.069$ and $M = 1.084$ was estimated to be less than 2% of the length of the Mach stem, a figure which, again, is somewhat larger than the wave structure observed numerically in the present work. The incident shock wave that is generated by the shock tube apparatus used in [12] is only approximately planar, however, and this may be one reason for the discrepancy. In addition, density gradients, which are visualized by the schlieren photo-optical technique used in [12], persist well beyond the supersonic patch into the subsonic region, making it difficult to estimate the extent of the supersonic patch from schlieren photographs. Nevertheless, the structure found in the experiments is very similar to the numerically computed Guderley Mach reflection solution. More recent experimental results [13] show more convincing evidence: the expansion fan and first terminating shocklet observed under conditions corresponding to $a \approx 1/2$ are more clearly visible, and the region appears to have an aspect ratio similar to the value of approximately 8:1 obtained in the solutions presented here.

Guderley's resolution was largely correct: a fourth wave, a centered expansion fan, originates at the triple point, although Guderley did not have any evidence that this is what actually occurs, nor did he suggest that there might be, in fact, a sequence of expansion fans and triple points. It is interesting, as noted in [12], that experimental observations of weak shock reflections off thin wedges show that not only does an apparent Mach reflection occur but that the slip line disappears or becomes ill defined. Figures 6(a) and 7(a) show that the slip line still exists in a weak shock reflection, but that it is extremely weak, making it difficult to observe experimentally.

6. Conclusion. We have presented numerical evidence of a sequence of triple points, each containing a centered expansion fan, in solutions of a shock reflection problem for the full Euler equations. This result is in agreement with previous numerical solutions of shock reflection problems for the UTSDE and the nonlinear wave system. The present work provides further evidence that the reflection pattern we call Guderley Mach reflection occurs when a weak shock reflects off a thin wedge.

REFERENCES

[1] W. Bleakney and A. H. Taub, *Interaction of shock waves*, Rev. Modern Phys., 21 (1949), pp. 584–605.

[2] M. Brio and J. K. Hunter, *Mach reflection for the two-dimensional Burgers equation*, Phys. D, 60 (1992), pp. 194–207.

[3] S. Čanić and B. L. Keyfitz, *Quasi-one-dimensional Riemann problems and their role in self-similar two-dimensional problems*, Arch. Rational Mech. Anal., 144 (1998), pp. 233–258.

[4] P. Colella and L. F. Henderson, *The von Neumann paradox for the diffraction of weak shock waves*, J. Fluid Mech., 213 (1990), pp. 71–94.

[5] K. G. Guderley, *Considerations of the Structure of Mixed Subsonic-Supersonic Flow Patterns*, Air Material Command Tech. Report F-TR-2168-ND, ATI 22780, GS-AAF-Wright Field 39, U.S. Wright-Patterson Air Force Base, Dayton, OH, 1947.

[6] L. F. Henderson, *On a class of multi-shock intersections in a perfect gas*, Aero. Q., 17 (1966), pp. 1–20.

[7] L. F. Henderson, *Regions and boundaries for diffracting shock wave systems*, Z. Angew. Math. Mech., 67 (1987), pp. 73–86.

[8] J. K. Hunter and M. Brio, *Weak shock reflection*, J. Fluid Mech., 410 (2000), pp. 235–261.

[9] J. K. Hunter and A. M. Tesdall, *Weak shock reflection*, in A Celebration of Mathematical Modeling, D. Givoli, M. Grote, and G. Papanicolaou, eds., Kluwer Academic Press, New York, 2004, pp. 93–112.

[10] J. von Neumann, *Collected Works*, Vol. 6, Pergamon Press, New York, 1963.

[11] A. Sasoh, K. Takayama, and T. Saito, *A weak shock wave reflection over wedges*, Shock Waves, 2 (1992), pp. 277–281.

[12] B. Skews and J. Ashworth, *The physical nature of weak shock wave reflection*, J. Fluid Mech., 542 (2005), pp. 105–114.

[13] B. Skews, *private communication.*

[14] J. Sternberg, *Triple-shock-wave intersections*, Phys. Fluids, 2 (1959), pp. 179–206.

[15] E. G. Tabak and R. R. Rosales, *Focusing of weak shock waves and the von Neumann paradox of oblique shock reflection*, Phys. Fluids, 6 (1994), pp. 1874–1892.

[16] A. M. Tesdall and J. K. Hunter, *Self-similar solutions for weak shock reflection*, SIAM J. Appl. Math., 63 (2002), pp. 42–61.

[17] A. M. Tesdall, R. Sanders, and B. L. Keyfitz, *The triple point paradox for the nonlinear wave system*, SIAM J. Appl. Math., 67 (2006), pp. 321–336.

[18] E. Vasil'ev and A. Kraiko, *Numerical simulation of weak shock diffraction over a wedge under the von Neumann paradox conditions*, Comput. Math. Math. Phys, 39 (1999), pp. 1335–1345.

[19] A. Zakharian, M. Brio, J. K. Hunter, and G. Webb, *The von Neumann paradox in weak shock reflection*, J. Fluid Mech., 422 (2000), pp. 193–205.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1378–1403

AN ASYMPTOTIC FACTORIZATION METHOD FOR INVERSE ELECTROMAGNETIC SCATTERING IN LAYERED MEDIA*

ROLAND GRIESMAIER†

Abstract. We consider the inverse problem to reconstruct the number and the positions of a collection of finitely many small perfectly conducting scatterers buried within the lower halfspace of an unbounded two-layered background medium from near field measurements of time harmonic electromagnetic fields. For this purpose we first study the direct scattering problem and derive an asymptotic expansion of the scattered field, as the size of the scatterers tends to zero. Integral equation methods and a factorization of the corresponding near field measurement operator are applied to prove this expansion. In the second part of this work we use the asymptotic expansion to justify a noniterative reconstruction algorithm, which is a combination of factorization methods and MUSIC-type methods. We illustrate the feasibility of this method by a numerical example.

Key words. inverse scattering, Maxwell's equations, small scatterers, layered media, asymptotic expansions

AMS subject classifications. 35C20, 78A46, 35Q60

DOI. 10.1137/060677021

1. Introduction.

We consider a simple but fully three dimensional model for the electromagnetic exploration of perfectly conducting objects buried within the lower half-space of an unbounded two-layered background medium. In possible applications, such as, e.g., humanitarian demining or, more generally, the exploration of the ground's subsurface to detect and identify buried objects, the two layers would correspond to air and soil. Moving a set of electric devices parallel to the surface of ground to generate a time harmonic field, the induced field is measured within the same devices. The goal is to retrieve information about buried scatterers from these data.

This work originated in the project [23] on humanitarian demining. In the course of this project mathematical methods for analyzing data obtained from standard off-the-shelf metal detectors have been developed. The aim of the project has been to reduce the number of false alarms produced by such devices used for humanitarian demining.

In mathematical terms, we consider an inverse scattering problem for Maxwell's equations in a two-layered background medium. An iterative method for such a problem was recently proposed by Delbary et al. [17]. Among the so-called qualitative methods (see Cakoni and Colton [10]), the linear sampling method was studied by Gebauer et al. [20] and by Cakoni, Fares, and Haddar [11]. Moreover, the factorization method was applied by Kirsch [29] and by Gebauer, Hanke, and Schneider [21]. In numerical experiments these methods turned out to be quite sensitive to noise. This is of course due to the ill-posedness of the inverse problem.

In order to handle this ill-posedness it is generally advisable to incorporate all available a priori knowledge about the measurement device and the scatterers and

*Received by the editors December 7, 2006; accepted for publication (in revised form) January 6, 2007; published electronically April 23, 2008. Part of this work has been supported by the German Federal Ministry of Education and Research in the BMBF-project [23] under grant 01 RX 0302.

http://www.siam.org/journals/siap/68-5/67702.html

†Institut für Mathematik, Johannes Gutenberg-Universität, 55099 Mainz, Germany (griesmaier@math.uni-mainz.de).

to determine very specific features. Standard off-the-shelf metal detectors used for humanitarian demining work at relatively low frequencies around 20 kHz; cf., e.g., [22]. In vacuum this corresponds to wavelengths of approximately 15 km. Thus the typical size of the objects of interest, which is only a few centimeters, is very small with respect to the wavelength of the incident field. We use this a priori knowledge to justify a noniterative reconstruction method that determines only the number and the position of the unknown scatterers but is more robust against noise in the data.

This method is a generalization of a method which was originally developed for electrical impedance tomography by Brühl, Hanke, and Vogelius [8]. It is based on an asymptotic expansion of the scattered field on the measurement device as the size of the scatterers tends to zero. A similar reconstruction method was recently investigated by Ammari et al. [2] for homogeneous background media and by Iakovleva et al. [24] for two-layered background media. In contrast to the present investigation, these works study a discrete measurement array, which can be considered as a special case of the measurement device studied here. We expect that the theoretical results obtained for nondiscrete measurement devices can be applied to even more realistic models for the measurement process; cf., e.g., [17]. Moreover, the asymptotic expansions of the scattered field were obtained only formally in [2] and [24]. Here we give a rigorous justification of these formulas for two-layered background media. For bounded background domains related formulas were rigorously proven by Ammari et al. [6, 4, 5], and these results were extended to unbounded homogeneous media and plane wave incident fields by Ammari and Volkov [7]. But this analysis applies neither to layered media nor to near field measurements such as considered here.

Our proof of the asymptotic formula employs a factorization of the near field measurement operator that maps magnetic dipole distributions on the measurement device to the corresponding scattered field on the same device. We apply layer potential techniques to describe the three operators occurring in this factorization and expand them separately as the size of the scatterers tends to zero. Then these expansions are combined to calculate the leading order term in the asymptotic expansion of the scattered field. This generalizes the approach we used in [1] for a boundary value problem in electrostatics. By contrast, in [6, 4, 5, 7] variational methods were applied.

Then, we derive a characterization of the location of the scatterers in terms of the range of the leading order term of the asymptotic expansion of the near field measurement operator, similar to range criteria known from factorization methods, introduced first by Kirsch [28], and MUSIC-type methods, applied first to inverse scattering problems by Devaney [18]. We use a MUSIC-type strategy to implement this range criterion numerically; basically, MUltiple SIgnal Classification is a method of characterizing the range of finite rank operators on Hilbert spaces; see Cheney [12].

The article is organized as follows. After introducing some notation in the next section we describe our model and define the measurement operator in section 3. In section 4 we derive a factorization of this operator, and in section 5 we collect some facts concerning boundary integral operators arising in electromagnetic scattering theory for layered background media. Sections 6 and 7 are devoted to the asymptotic expansion of the measurement operator. Then, in section 9 we derive a characterization of the scatterers in terms of a range criterion, and in section 10 we comment on how to implement this criterion numerically. Finally, we present numerical results.

2. Preliminaries.

We introduce our notation and recall some facts concerning function spaces used in the context of Maxwell's equations. For further details we refer the reader to [9, 31, 32]. Suppose $D \subset \mathbb{R}^3$ is a bounded domain of class $C^{2,\alpha}$, $0 < \alpha < 1$. Denote by $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ the usual Cartesian basis of $\mathbb{R}^3$, by $\boldsymbol{x} = (x_1, x_2, x_3)^\top$ a generic point in $\mathbb{R}^3$, and by $\boldsymbol{\nu}$ the unit outward normal to ∂D. Throughout let $\boldsymbol{x} \cdot \boldsymbol{y}$ and $\boldsymbol{x} \times \boldsymbol{y}$ be the scalar product and the vector product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3$, respectively, and let $|\boldsymbol{x}|$ denote the Euclidean norm of $\boldsymbol{x}$. The standard complex valued Sobolev spaces $H^r(D)$, $H^r_{\mathrm{loc}}(\mathbb{R}^3 \setminus \overline{D})$ for any $r \in \mathbb{R}$ and $H^s(\partial D)$ for $s \in [-2, 2]$ are defined on D, $\mathbb{R}^3 \setminus \overline{D}$ and on the boundary ∂D, respectively; see [30]. Let $\gamma_0 : H^r(D) \to H^{r-1/2}(\partial D)$, $1/2 < r \leq 2$, be the standard trace operator. We also need the spaces $\boldsymbol{H}(\mathbf{curl}, D)$, $\boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}, \mathbb{R}^3 \setminus \overline{D})$, $\boldsymbol{H}(\mathrm{div}, D)$, and $\boldsymbol{H}_{\mathrm{loc}}(\mathrm{div}, \mathbb{R}^3 \setminus \overline{D})$ of (locally) square integrable vector fields with (locally) square integrable curl and divergence, respectively.

The surface gradient $\nabla_{\partial D}$ and the surface vector curl $\mathbf{curl}_{\partial D}$ are defined on ∂D in the usual way by a localization argument. The adjoint operators of $-\nabla_{\partial D}$ and $\mathbf{curl}_{\partial D}$ are the surface divergence $\mathrm{div}_{\partial D}$ and the surface scalar curl $\mathrm{curl}_{\partial D}$, respectively. We introduce the Hilbert space $\boldsymbol{H}_t^{-1/2}(\partial D)$ of tangential vector fields in $H^{-1/2}(\partial D)^3$ and the Hilbert spaces $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$ of vector fields in $\boldsymbol{H}_t^{-1/2}(\partial D)$ with surface divergence and surface scalar curl in $H^{-1/2}(\partial D)$, respectively. The space $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$ is naturally identified with the dual space of $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$. We denote the corresponding duality pairing by $\langle \boldsymbol{b}, \boldsymbol{a} \rangle_{\partial D} = \int_{\partial D} \boldsymbol{b} \cdot \boldsymbol{a} \, \mathrm{d}s$ for any $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{b} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$.

For any regular vector field $\boldsymbol{u}$ we define the normal trace $\gamma_n(\boldsymbol{u}) := \boldsymbol{u}|_{\partial D} \cdot \boldsymbol{\nu}$, the tangential trace $\gamma_t(\boldsymbol{u}) := \boldsymbol{\nu} \times \boldsymbol{u}|_{\partial D}$, and the projection on the tangent plane $\pi_t(\boldsymbol{u}) := (\boldsymbol{\nu} \times \boldsymbol{u}|_{\partial D}) \times \boldsymbol{\nu}$. Furthermore, let $r(\boldsymbol{a}) := \boldsymbol{\nu} \times \boldsymbol{a}$ for any regular vector field $\boldsymbol{a}$ on ∂D. Then γ_n, γ_t, π_t, and r can be extended to continuous linear, surjective operators

$$\gamma_n : \boldsymbol{H}(\mathrm{div}, D) \to H^{-1/2}(\partial D), \qquad \gamma_t : \boldsymbol{H}(\mathbf{curl}, D) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D),$$
$$\pi_t : \boldsymbol{H}(\mathbf{curl}, D) \to \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D), \qquad r : \boldsymbol{H}_t^{-1/2}(\partial D) \to \boldsymbol{H}_t^{-1/2}(\partial D).$$

The extension of r is an isomorphism with $r^{-1} = r^\top = -r$, which maps $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ to $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$ and vice versa. For $\boldsymbol{u} \in \boldsymbol{H}(\mathbf{curl}, D)$ we have $\gamma_t(\boldsymbol{u}) = r(\pi_t(\boldsymbol{u}))$ and $\pi_t(\boldsymbol{u}) = -r(\gamma_t(\boldsymbol{u}))$. We note that for $\boldsymbol{a} \in \boldsymbol{H}_t^{-1/2}(\partial D)$,

$$\mathrm{div}_{\partial D}\, \boldsymbol{a} = \mathrm{curl}_{\partial D}\, r(\boldsymbol{a}) \quad \text{and} \quad \mathrm{curl}_{\partial D}\, \boldsymbol{a} = -\,\mathrm{div}_{\partial D}\, r(\boldsymbol{a}). \tag{2.1}$$

Furthermore, for $f \in H^1(D)$,

$$\nabla_{\partial D} \gamma_0(f) = \pi_t(\nabla f) \quad \text{and} \quad \mathbf{curl}_{\partial D} \gamma_0(f) = -r(\nabla_{\partial D} \gamma_0(f)) = -\gamma_t(\nabla f). \tag{2.2}$$

Finally, for $\boldsymbol{u} \in \boldsymbol{H}(\mathbf{curl}, D)$, it holds that

$$-\,\mathrm{div}_{\partial D}\, \gamma_t(\boldsymbol{u}) = \mathrm{curl}_{\partial D}\, \pi_t(\boldsymbol{u}) = \gamma_n(\mathbf{curl}\boldsymbol{u}). \tag{2.3}$$

Throughout we let scalar operators operate on vectors componentwise and vector operators on matrices column by column. For Banach spaces X and Y we denote by $\mathcal{L}(X, Y)$ the set of all bounded linear operators on X to Y. We write $\mathcal{L}(X)$ for $\mathcal{L}(X, X)$. Moreover, in our estimates we shall use a generic constant C.

3. The mathematical setting.

We decompose the space $\mathbb{R}^3 = \mathbb{R}^3_+ \cup \Sigma_0 \cup \mathbb{R}^3_-$ in a hyperplane $\Sigma_0 := \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_3 = 0\}$ corresponding to the surface of the ground, and the two halfspaces $\mathbb{R}^3_+ := \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_3 > 0\}$ and $\mathbb{R}^3_- := \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_3 < 0\}$

above and below Σ_0 representing air and ground, respectively. For convenience we set $\mathbb{R}^3_0 := \mathbb{R}^3 \setminus \Sigma_0$. We assume that both halfspaces are filled with homogeneous materials with dielectricity ε and permeability μ given by

$$\varepsilon(\boldsymbol{x}) := \begin{cases} \varepsilon_+, & \boldsymbol{x} \in \mathbb{R}^3_+, \\ \varepsilon_-, & \boldsymbol{x} \in \mathbb{R}^3_-, \end{cases} \qquad \mu(\boldsymbol{x}) := \begin{cases} \mu_+, & \boldsymbol{x} \in \mathbb{R}^3_+, \\ \mu_-, & \boldsymbol{x} \in \mathbb{R}^3_-, \end{cases}$$

and we require that ε_+ as well as $\mu_\pm$ are positive numbers, whereas ε_- may be complex with positive real and nonnegative imaginary parts to allow for soil materials that are conducting. The associated (discontinuous) wavenumber is $k := \omega\sqrt{\varepsilon\mu}$, where we assume $\omega > 0$. If $\varepsilon_- \notin \mathbb{R}$, then k is taken to have positive imaginary part. Throughout we investigate radiating solutions of the time harmonic Maxwell system

$$\mathbf{curl}\boldsymbol{H} + \mathrm{i}\,\omega\varepsilon\boldsymbol{E} = 0, \quad \mathbf{curl}\boldsymbol{E} - \mathrm{i}\,\omega\mu\boldsymbol{H} = 0 \tag{3.1}$$

in the exterior of some compact set $C \subset \mathbb{R}^3$. By this we understand, cf., e.g., [16, 31], solutions $\boldsymbol{E}, \boldsymbol{H} \in \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}, \mathbb{R}^3 \setminus C)$ which obey the integral radiation condition

$$\int_{\partial B_R(0)} \left| \frac{\boldsymbol{x}}{R} \times \boldsymbol{H}(\boldsymbol{x}) + \Big(\frac{\varepsilon(\boldsymbol{x})}{\mu(\boldsymbol{x})}\Big)^{1/2} \boldsymbol{E}(\boldsymbol{x}) \right|^2 \mathrm{d}s(\boldsymbol{x}) = o(1) \qquad \text{as } R \to \infty, \tag{3.2}$$

where $B_R(0) := \{\boldsymbol{x} \in \mathbb{R}^3 \mid |\boldsymbol{x}| < R\}$ denotes the ball of radius $R > 0$ around the origin.

For layered medium we have to distinguish between the electric and the magnetic dyadic Green's functions. The electric dyadic Green's function $\mathbb{G}^e$ is the radiating (distributional) solution of

$$\mathbf{curl}_x \frac{1}{\mu(\boldsymbol{x})} \mathbf{curl}_x\, \mathbb{G}^e(\boldsymbol{x}, \boldsymbol{y}) - \omega^2 \varepsilon(\boldsymbol{x}) \mathbb{G}^e(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\mu(\boldsymbol{x})} \delta(\boldsymbol{x} - \boldsymbol{y})\, \mathbb{I}_3, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3,$$

where $\mathbb{I}_3$ denotes the 3×3 identity matrix. Note that we are using $\boldsymbol{x}$ as an independent variable and $\boldsymbol{y}$ denotes the position of the source. The magnetic dyadic Green's function $\mathbb{G}^m$ fulfills the same equation, but ε and μ have to be swapped. From the derivation of these Green's tensors in [31, pp. 318–327] (cf. also [34, 17, 33]), we find that $\mathbb{G}^e$ and $\mathbb{G}^m$ can be written as

$$\mathbb{G}^{e/m}(\boldsymbol{x}, \boldsymbol{y}) = \Pi^{e/m}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{k(\boldsymbol{x})^2} \nabla_x \operatorname{div}_x \Pi^{e/m}(\boldsymbol{x}, \boldsymbol{y})$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3_0$, $\boldsymbol{x} \neq \boldsymbol{y}$. Here the (matrix valued) functions Π^e and Π^m are given by

$$\Pi^{e/m}(\boldsymbol{x}, \boldsymbol{y}) := \Phi_{k(\boldsymbol{x})}(\boldsymbol{x} - \boldsymbol{y})\, \mathbb{I}_3 + F^{e/m}(\boldsymbol{x}, \boldsymbol{y}), \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3_0,\ \boldsymbol{x} \neq \boldsymbol{y}, \tag{3.3}$$

where Φ_{k_+} and Φ_{k_-} denote the fundamental solution for the scalar Helmholtz equation in homogeneous medium with wavenumber k_+ and k_-, respectively; cf. [15, p. 16]. The functions Π^e and Π^m solve

$$(\Delta_x + k(\boldsymbol{x})^2) \Pi^{e/m}(\boldsymbol{x}, \boldsymbol{y}) = -\delta(\boldsymbol{x} - \boldsymbol{y})\, \mathbb{I}_3, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3_0,$$

and so F^e and F^m solve

$$(\Delta_x + k(\boldsymbol{x})^2) F^{e/m}(\boldsymbol{x}, \boldsymbol{y}) = 0, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3_0.$$

Applying a regularity result due to Weber [35, Thm. 2.9], we find that $\mathbb{G}^{e/m}|_{\mathbb{R}^3_\pm}(\cdot,\boldsymbol{y}) \in C^\infty(\overline{\mathbb{R}^3_\pm} \setminus \{\boldsymbol{y}\})$ for $\boldsymbol{y} \in \mathbb{R}^3_0$. Thus, $F^e(\cdot,\boldsymbol{y})$ and $F^m(\cdot,\boldsymbol{y})$ are smooth functions in $\mathbb{R}^3_0$ for $\boldsymbol{y}$ in any compact subset of $\mathbb{R}^3_0$.

Using Maxwell's equations and integration by parts the following reciprocity relations can be proven; cf. also [16, 24, 13]:

$$\mu(\boldsymbol{y})\mathbb{G}^e(\boldsymbol{x},\boldsymbol{y}) = \mu(\boldsymbol{x})\mathbb{G}^{e^\top}(\boldsymbol{y},\boldsymbol{x}) \qquad \text{for } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^3_0,\ \boldsymbol{x} \neq \boldsymbol{y}, \tag{3.4a}$$

$$\varepsilon(\boldsymbol{y})\mathbb{G}^m(\boldsymbol{x},\boldsymbol{y}) = \varepsilon(\boldsymbol{x})\mathbb{G}^{m^\top}(\boldsymbol{y},\boldsymbol{x}) \qquad \text{for } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^3_0,\ \boldsymbol{x} \neq \boldsymbol{y}, \tag{3.4b}$$

$$k^2(\boldsymbol{y})\mathbf{curl}_x\mathbb{G}^e(\boldsymbol{x},\boldsymbol{y}) = k^2(\boldsymbol{x})(\mathbf{curl}_y\mathbb{G}^m)^\top(\boldsymbol{y},\boldsymbol{x}) \qquad \text{for } \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^3_0,\ \boldsymbol{x} \neq \boldsymbol{y}. \tag{3.4c}$$

We denote by $\Sigma_d := \{\boldsymbol{x} \in \mathbb{R}^3_+ \mid \boldsymbol{x}\cdot\boldsymbol{e}_3 = d\} \subset \mathbb{R}^3_+$ the hyperplane parallel to the surface of the ground at height $d > 0$ and assume that measurements and excitations are restricted to an open bounded sheet $\mathcal{M} \subset \Sigma_d$ supporting the device. A time harmonic excitation, given by a magnetic dipole density $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M}) := L^2(\mathcal{M})^3$ on $\mathcal{M}$, leads to a primary electromagnetic field $(\boldsymbol{E}^i, \boldsymbol{H}^i)$ satisfying (3.1) in $\mathbb{R}^3 \setminus \mathcal{M}$, where the magnetic field has the form

$$\boldsymbol{H}^i = k_+^2 \int_{\mathcal{M}} \mathbb{G}^m(\cdot,\boldsymbol{y})\boldsymbol{\varphi}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}); \tag{3.5}$$

cf., e.g., Sommerfeld [34].

We suppose that $\mathbb{R}^3_-$ contains a finite number of perfectly conducting scatterers, each of the form $D_{\delta,j} := \boldsymbol{z}_j + \delta B_j$, where B_j is a bounded domain of class $C^{2,\alpha}$, $0 < \alpha < 1$, containing the origin, such that all components of B_j are simply connected, and their boundaries are connected, $1 \le j \le m$. The points $\boldsymbol{z}_j \in \mathbb{R}^3_-$, $1 \le j \le m$, that determine the location of the scatterers are assumed to satisfy $|\boldsymbol{z}_j - \boldsymbol{z}_l| \ge c_0$ for $j \neq l$ and $\operatorname{dist}(\boldsymbol{z}_j, \Sigma_0) \ge c_0$ for some constant $c_0 > 0$, $1 \le j, l \le m$. The value of $0 < \delta \le 1$, the common order of magnitude of the diameters of the scatterers, is assumed to be small enough such that the scatterers are disjoint and compactly contained in $\mathbb{R}^3_-$. So the total collection of scatterers takes the form $D_\delta := \bigcup_{j=1}^m (\boldsymbol{z}_j + \delta B_j)$. The perfect conductor sitting in D_δ induces a secondary field $(\boldsymbol{E}^s, \boldsymbol{H}^s)$ which is a radiating solution of (3.1) in $\mathbb{R}^3 \setminus \overline{D_\delta}$ subject to the boundary condition

$$\boldsymbol{\nu} \times \boldsymbol{E}^s = -\boldsymbol{\nu} \times \boldsymbol{E}^i \qquad \text{on } \partial D_\delta. \tag{3.6}$$

For a mathematical treatment of this direct problem we refer the reader to [16, 31, 17]. We define the (measurement) operator G_δ, which maps given excitations $\boldsymbol{\varphi}$ onto the corresponding secondary magnetic field $\boldsymbol{H}^s|_{\mathcal{M}}$ on $\mathcal{M}$, i.e.,

$$G_\delta :\ \boldsymbol{L}^2(\mathcal{M}) \to \boldsymbol{L}^2(\mathcal{M}), \quad G_\delta\boldsymbol{\varphi} := \boldsymbol{H}^s|_{\mathcal{M}}. \tag{3.7}$$

As in [20, Thm. 2.1] it can be seen that G_δ is a compact operator.

4. The factorization of G_δ.

In this section we study a factorization of the measurement operator G_δ from (3.7) similar to the one developed in [20], but here we do not restrict ourselves to tangential excitations and measurements.

Suppose $\boldsymbol{\psi} \in \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)$ and denote by $(\boldsymbol{E}^\psi, \boldsymbol{H}^\psi)$ the associated radiating solution of the exterior Maxwell boundary value problem

$$\mathbf{curl}\boldsymbol{H}^\psi + \mathrm{i}\,\omega\varepsilon\boldsymbol{E}^\psi = 0, \quad \mathbf{curl}\boldsymbol{E}^\psi - \mathrm{i}\,\omega\mu\boldsymbol{H}^\psi = 0 \qquad \text{in } \mathbb{R}^3 \setminus \overline{D_\delta}, \tag{4.1a}$$

$$\boldsymbol{\nu} \times \boldsymbol{E}^\psi = \boldsymbol{\psi} \qquad \text{on } \partial D_\delta. \tag{4.1b}$$

Uniqueness of solutions follows for $\Im\varepsilon_- = 0$ from [16, Prop. 2.5]. For $\Im\varepsilon_- > 0$ this was proven in [17, Thm. 2.1]. Existence of solutions will be shown in the next sections by reducing the boundary value problem to an integral equation of the second kind and applying Riesz–Fredholm theory. We define

$$L_\delta : \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta) \to \boldsymbol{L}^2(\mathcal{M}), \quad L_\delta \boldsymbol{\psi} := \boldsymbol{H}^{\psi}|_{\mathcal{M}}. \tag{4.2}$$

Then L_δ is a bounded linear operator. In particular, if $\boldsymbol{E}^i$ and $\boldsymbol{H}^s$ are the primary electric and secondary magnetic fields introduced in section 3, respectively, then $\boldsymbol{\psi} := -\boldsymbol{\nu} \times \boldsymbol{E}^i|_{\partial D_\delta}$ yields $\boldsymbol{H}^{\psi} = \boldsymbol{H}^s$. This means that $L_\delta : -\boldsymbol{\nu} \times \boldsymbol{E}^i|_{\partial D_\delta} \mapsto \boldsymbol{H}^s|_{\mathcal{M}}$.

We denote the standard bilinear form on $\boldsymbol{L}^2(\mathcal{M})$ by $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ and the corresponding transpose of L_δ by $L_\delta^\top : \boldsymbol{L}^2(\mathcal{M}) \to \boldsymbol{H}^{-1/2}_{\mathrm{curl}}(\partial D_\delta)$.

Proposition 4.1. *Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$. Denote by $\boldsymbol{H}^i$ and $\boldsymbol{H}^s$ the associated primary and secondary magnetic fields introduced in section* 3. *Then*

$$L_\delta^\top \boldsymbol{\varphi} = \frac{1}{\mathrm{i}\,\omega\mu_+}(\boldsymbol{\nu} \times \boldsymbol{H}|_{\partial D_\delta}) \times \boldsymbol{\nu} \qquad \text{on } \partial D_\delta, \tag{4.3}$$

where $\boldsymbol{H} = \boldsymbol{H}^i + \boldsymbol{H}^s$ is the total magnetic field.

Proof. Given $\boldsymbol{\psi} \in \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)$, let $(\boldsymbol{E}^{\boldsymbol{\psi}}, \boldsymbol{H}^{\boldsymbol{\psi}})$ be the radiating solution to (4.1). For any $\boldsymbol{y} \in \mathbb{R}^3 \setminus \overline{D_\delta}$ we have the representation formula

$$\boldsymbol{H}^{\boldsymbol{\psi}}(\boldsymbol{y}) = \int_{\partial D_\delta} \frac{\varepsilon(\boldsymbol{y})}{\varepsilon(\boldsymbol{x})}\Big(\mathbb{G}^{m\top}(\boldsymbol{x},\boldsymbol{y})(\boldsymbol{\nu} \times \mathbf{curl}\boldsymbol{H}^{\boldsymbol{\psi}})(\boldsymbol{x}) + (\mathbf{curl}_x\mathbb{G}^m)^\top(\boldsymbol{x},\boldsymbol{y})(\boldsymbol{\nu} \times \boldsymbol{H}^{\boldsymbol{\psi}})(\boldsymbol{x})\Big)\,\mathrm{d}s(\boldsymbol{x});$$

cf. [16, Prop. A.9]. Using this formula the proposition can be proven by applying (4.1), (3.5), (3.6), two times partial integration as in [31, Thm. 3.31], and (3.2). See also [20] for a corresponding result for tangential densities $\boldsymbol{\varphi}$ on $\mathcal{M}$. □

Finally, we consider the diffraction problem

$$\mathbf{curl}\boldsymbol{H}^d + \mathrm{i}\,\omega\varepsilon\boldsymbol{E}^d = 0, \quad \mathbf{curl}\boldsymbol{E}^d - \mathrm{i}\,\omega\mu\boldsymbol{H}^d = 0 \qquad \text{in } \mathbb{R}^3 \setminus \partial D_\delta, \tag{4.4a}$$

with the jump conditions

$$[(\boldsymbol{\nu} \times \boldsymbol{H}^d) \times \boldsymbol{\nu}]_{\partial D_\delta} = \boldsymbol{\chi}, \qquad [\boldsymbol{\nu} \times \boldsymbol{E}^d]_{\partial D_\delta} = 0 \qquad \text{on } \partial D_\delta. \tag{4.4b}$$

Here, $\boldsymbol{\chi} \in \boldsymbol{H}^{-1/2}_{\mathrm{curl}}(\partial D_\delta)$ is a given tangential field on ∂D_δ, and the square brackets denote the differences between the respective traces from outside and inside. We are looking for a radiating solution $(\boldsymbol{E}^d, \boldsymbol{H}^d)$ of this problem. Uniqueness of solutions has been stated in [29, Thm. 3.4] for $\Im\varepsilon_- = 0$. If $\Im\varepsilon_- > 0$ this can be shown by the same arguments as used in [33, pp. 61–63]. Existence of solutions will be shown later by writing them in terms of layer potentials. Given the solution, we define

$$F_\delta : \boldsymbol{H}^{-1/2}_{\mathrm{curl}}(\partial D_\delta) \to \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta), \quad F_\delta\boldsymbol{\chi} := \boldsymbol{\nu} \times \boldsymbol{E}^d|_{\partial D_\delta}. \tag{4.5}$$

Then F_δ is a bounded linear operator. For $\boldsymbol{\chi} = (\boldsymbol{\nu} \times \boldsymbol{H}|_{\partial D_\delta}) \times \boldsymbol{\nu}$, i.e., the tangential component of the total magnetic field corresponding to some excitation $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ as described in section 3, the solution of the diffraction problem (4.4) can be constructed from the corresponding primary and secondary fields, namely,

$$\boldsymbol{E}^d = \begin{cases} \boldsymbol{E}^s, & x \in \mathbb{R}^3 \setminus \overline{D_\delta}, \\ -\boldsymbol{E}^i, & x \in D_\delta, \end{cases} \qquad \boldsymbol{H}^d = \begin{cases} \boldsymbol{H}^s, & x \in \mathbb{R}^3 \setminus \overline{D_\delta}, \\ -\boldsymbol{H}^i, & x \in D_\delta. \end{cases}$$

Consequently, we have $F_\delta:\ (\boldsymbol{\nu}\times\boldsymbol{H}|_{\partial D_\delta})\times\boldsymbol{\nu}\mapsto\boldsymbol{\nu}\times\boldsymbol{E}^s|_{\partial D_\delta}=-\boldsymbol{\nu}\times\boldsymbol{E}^i|_{\partial D_\delta}$. Altogether, we obtain the mapping sequence

$$\boldsymbol{\varphi}\xmapsto{L_\delta^\top}\frac{1}{\mathrm{i}\,\omega\mu_+}(\boldsymbol{\nu}\times\boldsymbol{H}|_{\partial D_\delta})\times\boldsymbol{\nu}\xmapsto{F_\delta}-\frac{1}{\mathrm{i}\,\omega\mu_+}\boldsymbol{\nu}\times\boldsymbol{E}^i|_{\partial D_\delta}\xmapsto{L_\delta}\frac{1}{\mathrm{i}\,\omega\mu_+}\boldsymbol{H}^s|_{\mathcal{M}}.$$

This yields the following theorem; cf. [20] for a corresponding result in case of tangential densities $\boldsymbol{\varphi}$ on $\mathcal{M}$.

THEOREM 4.2. *Given L_δ from* (4.2) *and F_δ from* (4.5) *the measurement operator G_δ from* (3.7) *admits the factorization*

$$G_\delta=\mathrm{i}\,\omega\mu_+L_\delta F_\delta L_\delta^\top. \tag{4.6}$$

5. Surface potentials.

Here, we collect some results concerning boundary integral operators for electromagnetic scattering in two-layered media.

5.1. Surface potentials for homogeneous media.

First, we consider a homogeneous medium with wavenumber k_-. If $D\subset\mathbb{R}^3$ is a bounded domain of class $C^{2,\alpha}$, $0<\alpha<1$, the single layer potential with smooth density f is defined by

$$(\mathcal{S}_D^-f)(\boldsymbol{x}):=\int_{\partial D}\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})f(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}),\qquad \boldsymbol{x}\in\mathbb{R}^3\setminus\partial D.$$

Then $\mathcal{S}_D^-f$ and $\boldsymbol{\nu}\times\nabla\mathcal{S}_D^-f$ are continuous across ∂D; cf. [14, Thm. 2.12, Thm. 2.17]. It can be shown [30, Thm. 6.11] that the mapping $\mathcal{S}_D^-:H^{-1/2}(\partial D)\to H^1_{\mathrm{loc}}(\mathbb{R}^3)$ is bounded. The jump relations on ∂D remain valid for $f\in H^{-1/2}(\partial D)$, but they have to be interpreted in the sense of trace theorems.

Analogously, the vector potential with smooth tangential density $\boldsymbol{a}$ is given by

$$(\mathcal{A}_D^-\boldsymbol{a})(\boldsymbol{x}):=\int_{\partial D}\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}),\qquad \boldsymbol{x}\in\mathbb{R}^3\setminus\partial D.$$

Then $\mathcal{A}_D^-\boldsymbol{a}$, $\boldsymbol{\nu}\cdot\mathbf{curl}\mathcal{A}_D^-\boldsymbol{a}$, and $\boldsymbol{\nu}\times\mathbf{curl}\,\mathbf{curl}\mathcal{A}_D^-\boldsymbol{a}$ are continuous across ∂D; cf. [14, Thm. 2.24] and [15, Thm. 6.11]. The tangential components of $\boldsymbol{\nu}\times\mathbf{curl}\mathcal{A}_D^-\boldsymbol{a}$ are discontinuous across ∂D and satisfy the jump relation

$$\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}\mathcal{A}_D^-\boldsymbol{a}\big|_{\partial D}^{\pm}(\boldsymbol{x})=\int_{\partial D}\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}_x(\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y}))\,\mathrm{d}s(\boldsymbol{y})\pm\frac{1}{2}\boldsymbol{a}(\boldsymbol{x})$$

for $\boldsymbol{x}\in\partial D$. It can be shown [30, Thm. 6.11] that $\mathcal{A}_D^-:\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)\to\boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl},\mathbb{R}^3)$ is bounded and that the jump relations on ∂D remain valid for $\boldsymbol{a}\in\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$. Furthermore, for smooth tangential densities $\boldsymbol{a}$ we define

$$\begin{aligned}(M_D^-\boldsymbol{a})(\boldsymbol{x})&:=\int_{\partial D}\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}_{\boldsymbol{x}}(\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y}))\,\mathrm{d}s(\boldsymbol{y}), &&\boldsymbol{x}\in\partial D,\\ (N_D^-\boldsymbol{a})(\boldsymbol{x})&:=\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}\,\mathbf{curl}\int_{\partial D}\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{\nu}(\boldsymbol{y})\times\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}), &&\boldsymbol{x}\in\partial D.\end{aligned}$$

Combining results from [30, 15] and [27], it can be seen that the operators $M_D^-:\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)\to\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $N_D^-:\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)\to\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ are continuous. Moreover, M_D^- is compact, and its transpose with respect to the bilinear form $\langle\cdot,\cdot\rangle_{\partial D}$ is given by $M_D^{-\top}=rM_D^-r$. The operator N_D^- is symmetric. We also need the following identity (see [15, p. 170]:

$$\operatorname{div}\mathcal{A}_D^-\boldsymbol{a}=\mathcal{S}_D^-\operatorname{div}_{\partial D}\boldsymbol{a},\qquad \boldsymbol{a}\in\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D). \tag{5.1}$$

5.2. The potential theoretic limit. For $k_- = 0$ the expression Φ_{k_-} reduces to the fundamental solution Φ_0 of Laplace's equation. Substituting Φ_{k_-} by Φ_0 in the definitions above, we obtain integral operators

$$\mathcal{S}_D^0 : H^{-1/2}(\partial D) \to H^1_{\mathrm{loc}}(\mathbb{R}^3), \qquad \mathcal{A}_D^0 : \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}, \mathbb{R}^3),$$
$$M_D^0 : \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D), \qquad N_D^0 : \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D).$$

The corresponding mapping properties and jump relations remain valid for $k = 0$. Suppose that all components of D are simply connected and the complement of D is connected. Then the operator $\frac{1}{2}I + M_D^0$ has trivial nullspace in $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ [14, Thm. 5.5]. Therefore, we can apply Fredholm's alternative and find that $\frac{1}{2}I + M_D^0$ and $\frac{1}{2}I + {M_D^0}^\top$ are invertible on $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$, respectively. From ${M_D^0}^\top = r M_D^0 r$ we observe that for any $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ it holds that

$$\left(\frac{1}{2}I \pm {M_D^0}^\top\right) r\boldsymbol{a} = r\left(\frac{1}{2}I \mp M_D^0\right)\boldsymbol{a}. \tag{5.2}$$

Thus, $-\frac{1}{2}I + M_D^0$ and $-\frac{1}{2}I + {M_D^0}^\top$ are invertible on $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$, respectively, too.

Furthermore, given a scalar smooth function f we define

$$(K_D^0 f)(\boldsymbol{x}) := \int_{\partial D} \frac{\partial \Phi_0(\boldsymbol{x}-\boldsymbol{y})}{\partial \boldsymbol{\nu}(\boldsymbol{y})} f(\boldsymbol{y}) \,\mathrm{d}s(\boldsymbol{y}), \qquad \boldsymbol{x} \in \partial D.$$

It can be shown [32, Thm. 4.4.1] that the mapping $K_D^0 : H^{1/2}(\partial D) \to H^{1/2}(\partial D)$ is compact. The operator $-\frac{1}{2}I + K_D^0$ has trivial nullspace in $H^{1/2}(\partial D)$ [3, Lem. 2.5]. Hence, by Fredholm's alternative $-\frac{1}{2}I + K_D^0$ and $-\frac{1}{2}I + {K_D^0}^\top$ are invertible on $H^{1/2}(\partial D)$ and $H^{-1/2}(\partial D)$, respectively.

Lemma 5.1. (a) *The operators* $\pm\frac{1}{2}I + M_D^0$ *are isomorphisms on*

$$\boldsymbol{H}_{\mathrm{div},0}^{-1/2}(\partial D) := \{\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \mid \mathrm{div}_{\partial D}\, \boldsymbol{a} = 0\}.$$

(b) *For any* $f \in H^{1/2}(\partial D)$,

$$\left(\pm\frac{1}{2}I + {M_D^0}^\top\right)^{-1} \nabla_{\partial D} f = -\nabla_{\partial D}\left(\mp\frac{1}{2}I + K_D^0\right)^{-1} f. \tag{5.3}$$

Proof. Part (a) follows at once from

$$\mathrm{div}_{\partial D}\, M_D^0 \boldsymbol{a} = -{K_D^0}^\top \mathrm{div}_{\partial D}\, \boldsymbol{a}, \qquad \boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D); \tag{5.4}$$

cf. [15, p. 169]. By duality, (5.4) yields ${M_D^0}^\top \nabla_{\partial D} f = -\nabla_{\partial D} K_D^0 f$ for $f \in H^{1/2}(\partial D)$. Thus, we find

$$\left(\pm\frac{1}{2}I + {M_D^0}^\top\right)\nabla_{\partial D} f = -\nabla_{\partial D}\left(\mp\frac{1}{2}I + K_D^0\right) f,$$

which gives (5.3). □

5.3. Surface potentials for layered media. Here, we consider again the two-layered medium introduced in section 3. Let $D \subset \mathbb{R}^3_-$ be a bounded domain of class $C^{2,\alpha}$, $0 < \alpha < 1$, such that $\operatorname{dist}(D, \Sigma_0) \geq c_0$ for some constant $c_0 > 0$. We define modified vector potentials with smooth tangential density $\boldsymbol{a}$ by

$$\begin{aligned}(\mathcal{A}_D^{e/m}\boldsymbol{a})(\boldsymbol{x}) :=& \int_{\partial D} \Pi^{e/m}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y})\\ &=(\mathcal{A}_D^{-}\boldsymbol{a})(\boldsymbol{x}) + \int_{\partial D} F^{e/m}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}), \qquad \boldsymbol{x} \in \mathbb{R}^3 \setminus \partial D,\end{aligned}$$

and boundary integrals

$$(R_D^{e/m}\boldsymbol{a})(\boldsymbol{x}) := \int_{\partial D} \boldsymbol{\nu}(\boldsymbol{x}) \times \mathbf{curl}_x(F^{e/m}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y}))\,\mathrm{d}s(\boldsymbol{y}), \qquad \boldsymbol{x} \in \partial D, \tag{5.5}$$

$$\begin{aligned}(M_D^{e/m}\boldsymbol{a})(\boldsymbol{x}) :=& \int_{\partial D} \boldsymbol{\nu}(\boldsymbol{x}) \times \mathbf{curl}_x(\Pi^{e/m}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y}))\,\mathrm{d}s(\boldsymbol{y})\\ &=(M_D^{-}\boldsymbol{a})(\boldsymbol{x}) + (R_D^{e/m}\boldsymbol{a})(\boldsymbol{x}), \qquad \boldsymbol{x} \in \partial D.\end{aligned} \tag{5.6}$$

Because $F^{e/m}(\cdot,\boldsymbol{y})$ is smooth in $\mathbb{R}^3_0$ for $\boldsymbol{y}$ in any compact subset of $\mathbb{R}^3_0$, we find that $\mathcal{A}_D^{e/m}\boldsymbol{a}$ and $\boldsymbol{\nu} \times \mathbf{curl}\,\mathbf{curl}\mathcal{A}_D^{e/m}\boldsymbol{a}$ are continuous across ∂D. Furthermore,

$$\boldsymbol{\nu} \times \mathbf{curl}\mathcal{A}_D^{e/m}\boldsymbol{a}\big|_{\partial D}^{\pm} = \Big(\pm\frac{1}{2}I + M_D^{e/m}\Big)\boldsymbol{a} \qquad \text{on } \partial D.$$

The mapping $\mathcal{A}_D^{e/m} : \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{loc}}(\mathbf{curl}, \mathbb{R}^3_0)$ is continuous, and the operators $R_D^{e/m} : \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $M_D^{e/m} : \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ are compact. Assume that the exterior of D is connected and that the wavenumber k_- is not an interior Maxwell eigenvalue for D. Then the operator $\frac{1}{2}I + M_D^m$ has trivial nullspace in $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$. This can be proven in essentially the same way as [14, Thm. 4.23] for homogeneous medium and continuous densities. So we can apply Fredholm's alternative to obtain that $\frac{1}{2}I + M_D^m$ and $\frac{1}{2}I + M_D^{m\top}$ are invertible on $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$, respectively.

6. First estimates. In the following two sections we consider the case of a single scatterer, i.e., $D_\delta = \boldsymbol{z} + \delta B$. Multiple scatterers will be studied in section 8.

We often have to deal with changes of coordinates; thus we introduce the following notation. Given $\boldsymbol{a} \in C(\partial D_\delta)^3$ and $\boldsymbol{b} \in C(\partial B)^3$ we define $\hat{\boldsymbol{a}}, (\boldsymbol{a})^\wedge \in C(\partial B)^3$ and $\check{\boldsymbol{b}}, (\boldsymbol{b})^\vee \in C(\partial D_\delta)^3$ by

$$(\boldsymbol{a})^\wedge(\boldsymbol{\xi}) := \hat{\boldsymbol{a}}(\boldsymbol{\xi}) := \boldsymbol{a}(\delta\boldsymbol{\xi} + \boldsymbol{z}) \quad \text{and} \quad (\boldsymbol{b})^\vee(\boldsymbol{x}) := \check{\boldsymbol{b}}(\boldsymbol{x}) := \boldsymbol{b}\Big(\frac{\boldsymbol{x}-\boldsymbol{z}}{\delta}\Big) \tag{6.1}$$

for $\boldsymbol{\xi} \in \partial B$ and $\boldsymbol{x} \in \partial D_\delta$, respectively. This notation is also applied to functions from Sobolev spaces.

For arbitrary bounded domains $D \subset \mathbb{R}^3$ of class $C^{2,\alpha}$, $0 < \alpha < 1$, we use the following norms on $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)$:

$$\|\boldsymbol{a}\|_{\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D)} := \inf_{\boldsymbol{u}\in\boldsymbol{H}(\mathbf{curl},D),\,\gamma_t(\boldsymbol{u})=\boldsymbol{a}} \|\boldsymbol{u}\|_{\boldsymbol{H}(\mathbf{curl},D)} \qquad \text{for } \boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D),$$

$$\|\boldsymbol{b}\|_{\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D)} := \inf_{\boldsymbol{u}\in\boldsymbol{H}(\mathbf{curl},D),\,\pi_t(\boldsymbol{u})=\boldsymbol{b}} \|\boldsymbol{u}\|_{\boldsymbol{H}(\mathbf{curl},D)} \qquad \text{for } \boldsymbol{b} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D).$$

A simple calculation (cf. [1, Lem. 4.1] for a similar result) yields the following scaling properties of these norms under changes of coordinates as in (6.1). Suppose $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$, $\boldsymbol{b} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta)$, and assume $0 < \delta \leq 1$. Then

$$\delta^{\frac{3}{2}} \|\hat{\boldsymbol{a}}\|_{\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)} \leq \|\boldsymbol{a}\|_{\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)} \leq \delta^{\frac{1}{2}} \|\hat{\boldsymbol{a}}\|_{\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)}, \tag{6.2a}$$

$$\delta^{\frac{3}{2}} \|\hat{\boldsymbol{b}}\|_{\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)} \leq \|\boldsymbol{b}\|_{\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta)} \leq \delta^{\frac{1}{2}} \|\hat{\boldsymbol{b}}\|_{\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)}. \tag{6.2b}$$

In order to derive the asymptotic expansion in section 7 we need to expand the fundamental solution $\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y}) = \Phi_{k_-}(\delta(\boldsymbol{\xi}-\boldsymbol{\eta}))$ for $\boldsymbol{x} = \delta\boldsymbol{\xi}+\boldsymbol{z} \neq \delta\boldsymbol{\eta}+\boldsymbol{z} = \boldsymbol{y} \in \partial D_\delta$, i.e., $\boldsymbol{\xi} \neq \boldsymbol{\eta} \in \partial B$, as $\delta \to 0$. Expanding $e^{\mathrm{i}\,k_-\delta|\boldsymbol{\xi}-\boldsymbol{\eta}|}$ in a power series we obtain the following formulas:

$$\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y}) = \frac{1}{\delta}\left(\Phi_0(\boldsymbol{\xi}-\boldsymbol{\eta}) + \frac{\mathrm{i}\,k_-\delta}{4\pi} + \mathcal{O}(\delta^2)\right) \qquad \text{as } \delta \to 0, \tag{6.3}$$

$$\nabla_x \Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y}) = \frac{1}{\delta^2}\left(\nabla_\xi \Phi_0(\boldsymbol{\xi}-\boldsymbol{\eta}) - \frac{k_-^2\delta^2}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|} + \mathcal{O}(\delta^3)\right) \qquad \text{as } \delta \to 0. \tag{6.4}$$

Remark 6.1 (eigenvalues). In section 5.3 we had to assume that the wavenumber k_- is not an interior Maxwell eigenvalue for the bounded domain D to obtain invertibility of the operators $\frac{1}{2}I + M_D^m$ and $\frac{1}{2}I + M_D^{m\top}$. Interior Maxwell eigenvalues for D are wavenumbers κ such that Maxwell's equations (3.1) in D with homogeneous boundary condition $\boldsymbol{\nu} \times \boldsymbol{E}|_{\partial D} = 0$ on ∂D have a nontrivial solution. If $\Im\kappa > 0$, it is well known that solutions to the interior Maxwell boundary value problem are unique (cf. [31, Thm. 4.17]), and thus κ is no eigenvalue. On the other hand, there is a discrete set of real eigenvalues $\kappa_j > 0$, $j \in \mathbb{N}$, for D accumulating only at infinity; cf. [31, Thm. 4.18].

Let $\{k_j\}_{j\in\mathbb{N}}$ be the set of interior Maxwell eigenvalues corresponding to the reference domain B. By a change of coordinates in the variational formulation of the eigenvalue problem (see [31, p. 96]), we find that $\{\delta^{-1}k_j\}_{j\in\mathbb{N}}$ is the set of eigenvalues corresponding to the domain $D_\delta = z + \delta B$, $0 < \delta \leq 1$. Therefore, we can assume henceforth in the derivation of the asymptotic expansion without loss of generality that δ is small enough so that $k_- \notin \{\delta^{-1}k_j\}_{j\in\mathbb{N}}$, i.e., that k_- is no interior Maxwell eigenvalue for the domains D_δ considered hereafter.

In the next lemma we investigate the scaling properties of the operator $M_{D_\delta}^m$.

LEMMA 6.2. *For $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$ we have*

$$M_{D_\delta}^m \boldsymbol{a} = (M_B^0 \hat{\boldsymbol{a}})^\vee + (E_M^m \hat{\boldsymbol{a}})^\vee.$$

Here E_M^m is a bounded linear operator, which is $\mathcal{O}(\delta^2)$ in $\mathcal{L}(\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B))$ as $\delta \to 0$, independent of $\boldsymbol{a}$.

Proof. Let $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$ and $\boldsymbol{a}_j$, $j \in \mathbb{N}$, be smooth tangential vector fields with smooth surface divergence on ∂D_δ so that $\boldsymbol{a}_j$ converges to $\boldsymbol{a}$ in $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$. For fixed $j \in \mathbb{N}$ and $\boldsymbol{x} \in D_\delta$ we observe by a change of variables $\boldsymbol{\xi} := \frac{\boldsymbol{x}-\boldsymbol{z}}{\delta}$ and $\boldsymbol{\eta} := \frac{\boldsymbol{y}-\boldsymbol{z}}{\delta}$ that

$$(M_{D_\delta}^0 \boldsymbol{a}_j)(\boldsymbol{x}) = \int_{\partial B} \boldsymbol{\nu}(\boldsymbol{\xi}) \times \frac{1}{\delta}\mathbf{curl}_\xi \left(\hat{\boldsymbol{a}}_j(\boldsymbol{\eta}) \frac{1}{4\pi\delta|\boldsymbol{\xi}-\boldsymbol{\eta}|}\right) \delta^2 \,\mathrm{d}s(\boldsymbol{\eta}) = (M_B^0 \hat{\boldsymbol{a}}_j)(\boldsymbol{\xi}),$$

i.e., $M_{D_\delta}^0 \boldsymbol{a}_j = (M_B^0 \hat{\boldsymbol{a}}_j)^\vee$. From (6.4) we find

$$\nabla_x(\Phi_{k_-} - \Phi_0)(\boldsymbol{x}-\boldsymbol{y}) = \frac{1}{\delta^2}\left(-\frac{k_-^2\delta^2}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|} + \mathcal{O}(\delta^3)\right)$$

for $\boldsymbol{x} \neq \boldsymbol{y}$ as $\delta \to 0$. So, again by a change of coordinates we obtain for $\boldsymbol{x} \in \partial D_\delta$

$$\begin{aligned}\big((M^-_{D_\delta} - M^0_{D_\delta})\boldsymbol{a}_j\big)(\boldsymbol{x}) &= \int_{\partial D_\delta} \boldsymbol{\nu}(\boldsymbol{x}) \times \big(\nabla_x(\Phi_{k_-} - \Phi_0)(\boldsymbol{x}-\boldsymbol{y}) \times \boldsymbol{a}_j(\boldsymbol{y})\big)\,\mathrm{d}s(\boldsymbol{y})\\ &= \delta^2 \int_{\partial B} \boldsymbol{\nu}(\boldsymbol{\xi}) \times \bigg(\Big(-\frac{k_-^2}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|} + \mathcal{O}(\delta)\Big) \times \hat{\boldsymbol{a}}_j(\boldsymbol{\eta})\bigg)\,\mathrm{d}s(\boldsymbol{\eta}) =: (E^-_M \hat{\boldsymbol{a}}_j)(\boldsymbol{\xi}).\end{aligned}$$

The kernel of E^-_M is pseudohomogeneous of class -2 (cf. [32, pp. 168–175]), and hence E^-_M is continuous from $H^{-1/2}(\partial B)^3$ into $H^{3/2}(\partial B)^3$; cf. [32, Thm. 4.3.2]. So E^-_M is also continuous from $\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)$ to $\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)$; in particular, it is $\mathcal{O}(\delta^2)$ in $\mathcal{L}(\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B))$ as $\delta \to 0$. Thus, recalling the continuity properties of the operators $M^-_{D_\delta}$, M^0_B, and E^-_M and letting $j \to \infty$, we obtain

$$M^-_{D_\delta}\boldsymbol{a} = (M^0_B \hat{\boldsymbol{a}})^\vee + (E^-_M \hat{\boldsymbol{a}})^\vee.$$

Recalling (5.6) it remains to estimate the norm of $R^m_{D_\delta}\boldsymbol{a}$. For this purpose, we denote by $\tilde{R}^m_{D_\delta}\boldsymbol{a}$ the extension of $R^m_{D_\delta}$ to $\boldsymbol{H}(\mathbf{curl}, D_\delta)$ (with respect to the trace operator γ_t), which is obtained canonically from (5.5) via

$$\tilde{R}^m_{D_\delta}\boldsymbol{a} := \int_{\partial D_\delta} \mathbf{curl}_x F^m(\cdot, \boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}) \qquad \text{in } D_\delta.$$

Then, because F^m is smooth near the center of the scatterer,

$$\begin{aligned}\|R^m_{D_\delta}\boldsymbol{a}\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)} &= \inf_{\boldsymbol{u}\in \boldsymbol{H}(\mathbf{curl},D_\delta),\, \gamma_t(\boldsymbol{u}) = R^m_{D_\delta}\boldsymbol{a}} \|\boldsymbol{u}\|^2_{\boldsymbol{H}(\mathbf{curl},D_\delta)} \leq \|\tilde{R}^m_{D_\delta}\boldsymbol{a}\|^2_{\boldsymbol{H}(\mathbf{curl},D_\delta)}\\ &= \int_{D_\delta}\Big|\int_{\partial D_\delta} \mathbf{curl}_x F^m(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y})\Big|^2\,\mathrm{d}\boldsymbol{x}\\ &\quad + \int_{D_\delta}\Big|\mathbf{curl}_x \int_{\partial D_\delta} \mathbf{curl}_x F^m(\boldsymbol{x},\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y})\Big|^2\,\mathrm{d}\boldsymbol{x}\\ &\leq \int_{D_\delta}\Big(\|\mathbf{curl}_x F^m(\boldsymbol{x},\cdot)\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{curl}}(\partial D_\delta)}\\ &\qquad\quad + \|\mathbf{curl}_x\,\mathbf{curl}_x F^m(\boldsymbol{x},\cdot)\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{curl}}(\partial D_\delta)}\Big)\|\boldsymbol{a}\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)}\,\mathrm{d}\boldsymbol{x}\\ &\leq C\delta^3\|\boldsymbol{a}\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)}\int_{D_\delta} 1\,\mathrm{d}\boldsymbol{x} \leq C\delta^6\|\boldsymbol{a}\|^2_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)}.\end{aligned}$$

Using (6.2) we find

$$\|(R^m_{D_\delta}\boldsymbol{a})^\wedge\|_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)} \leq \delta^{-\frac{3}{2}}\|R^m_{D_\delta}\boldsymbol{a}\|_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)} \leq C\delta^2\|\hat{\boldsymbol{a}}\|_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)}.$$

Thus, we define

$$E^m_M\boldsymbol{b} := E^-_M\boldsymbol{b} + (R^m_{D_\delta}\check{\boldsymbol{b}})^\wedge, \qquad \boldsymbol{b} \in \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B),$$

and obtain the desired result. ∎

For $\boldsymbol{a} \in \boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial D_\delta)$ Lemma 6.2 yields

$$\Big(\frac{1}{2}I + M^m_{D_\delta}\Big)\boldsymbol{a} = \Big(\Big(\frac{1}{2}I + M^0_B + E^m_M\Big)\hat{\boldsymbol{a}}\Big)^\vee.$$

Thus,

$$(6.5)\quad \Big(\frac{1}{2}I+M^m_{D_\delta}\Big)^{-1}\boldsymbol{a} = \Big(\Big(\frac{1}{2}I+M^0_B+E^m_M\Big)^{-1}\hat{\boldsymbol{a}}\Big)^\vee = \Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\hat{\boldsymbol{a}}\Big)^\vee+(\tilde{E}^m_M\hat{\boldsymbol{a}})^\vee,$$

where $\tilde{E}^m_M$ is a bounded linear operator, which is $\mathcal{O}(\delta^2)$ in $\mathcal{L}(\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B))$ as $\delta \to 0$, independent of $\boldsymbol{a}$.

7. Asymptotic expansion. In this section we expand the three operators L_δ, $L_\delta^\top$, and F_δ occurring in the factorization (4.6) of the measurement operator G_δ separately as the inhomogeneity size δ tends to zero. Then, we use these expansions to calculate the leading order term in the asymptotic expansion of G_δ.

First, we consider the exterior Maxwell boundary value problem (4.1) and study the asymptotic behavior of the operator L_δ from (4.2). A radiating solution of this problem is given by

$$\begin{aligned}
\boldsymbol{E}^\psi &:= \frac{\varepsilon_-}{\varepsilon}\mathbf{curl}\mathcal{A}^m_{D_\delta}\Big(\frac{1}{2}I+M^m_{D_\delta}\Big)^{-1}\boldsymbol{\psi} && \text{in } \mathbb{R}^3_0\setminus\overline{D_\delta},\\
\boldsymbol{H}^\psi &:= -\,\mathrm{i}\,\omega\varepsilon_-\int_{\partial D_\delta}\mathbb{G}^m(\cdot,\boldsymbol{y})\Big(\Big(\frac{1}{2}I+M^m_{D_\delta}\Big)^{-1}\boldsymbol{\psi}\Big)(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}) && \text{in } \mathbb{R}^3_0\setminus\overline{D_\delta}.
\end{aligned}$$

By Taylor expansion we obtain for $\boldsymbol{x}\in\mathcal{M}$, $\boldsymbol{z}\in\mathbb{R}^3_-$ with $\mathrm{dist}(\boldsymbol{z},\Sigma_0)\geq c_0$ for some constant $c_0>0$, and $\boldsymbol{\eta}\in\partial B$ as $\delta\to 0$ that

$$\mathbb{G}^m(\boldsymbol{x},\delta\boldsymbol{\eta}+\boldsymbol{z}) = \mathbb{G}^m(\boldsymbol{x},\boldsymbol{z})+\delta\sum_{l=1}^3\frac{\partial\mathbb{G}^m}{\partial y_l}(\boldsymbol{x},\boldsymbol{z})\eta_l+\mathcal{O}(\delta^2).$$

Thus, by a change of coordinates, applying (6.5) we have

$$\begin{aligned}
\boldsymbol{H}^\psi(\boldsymbol{x}) &= -\,\mathrm{i}\,\omega\varepsilon_-\delta^2\int_{\partial B}\mathbb{G}^m(\boldsymbol{x},\delta\boldsymbol{\eta}+\boldsymbol{z})\Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\hat{\boldsymbol{\psi}}\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})+\mathcal{O}(\delta^4)\\
&= -\,\mathrm{i}\,\omega\varepsilon_-\delta^2\mathbb{G}^m(\boldsymbol{x},\boldsymbol{z})\int_{\partial B}\Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\hat{\boldsymbol{\psi}}\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})\\
&\quad -\,\mathrm{i}\,\omega\varepsilon_-\delta^3\int_{\partial B}\sum_{l=1}^3\eta_l\frac{\partial\mathbb{G}^m}{\partial y_l}(\boldsymbol{x},\boldsymbol{z})\Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\hat{\boldsymbol{\psi}}\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})+\mathcal{O}(\delta^4)
\end{aligned}$$

for $\boldsymbol{x}\in\mathcal{M}$ as $\delta\to 0$. The last term on the right-hand side is bounded by $C\delta^4\|\hat{\boldsymbol{\psi}}\|_{\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)}$, where the constant C is independent of δ and $\boldsymbol{\psi}$, uniformly for $\boldsymbol{x}\in\mathcal{M}$. We define $L_0:\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)\to\boldsymbol{L}^2(\mathcal{M})$,

$$(7.1)\qquad L_0\boldsymbol{a} := -\,\mathrm{i}\,\omega\varepsilon_-\mathbb{G}^m(\cdot,\boldsymbol{z})\int_{\partial B}\Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\boldsymbol{a}\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}),$$

and $L_1:\boldsymbol{H}^{-1/2}_{\mathrm{div}}(\partial B)\to\boldsymbol{L}^2(\mathcal{M})$,

$$(7.2)\qquad L_1\boldsymbol{a} := -\,\mathrm{i}\,\omega\varepsilon_-\int_{\partial B}\sum_{l=1}^3\eta_l\frac{\partial\mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z})\Big(\Big(\frac{1}{2}I+M^0_B\Big)^{-1}\boldsymbol{a}\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}).$$

Then L_0 and L_1 are bounded linear operators, and we have the following asymptotic behavior.

Proposition 7.1. *For all $\boldsymbol{\psi} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$,*

$$L_\delta \boldsymbol{\psi} = \delta^2 L_0 \hat{\boldsymbol{\psi}} + \delta^3 L_1 \hat{\boldsymbol{\psi}} + E_L \hat{\boldsymbol{\psi}}, \tag{7.3}$$

where E_L is a bounded linear operator, which is $\mathcal{O}(\delta^4)$ in $\mathcal{L}(\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B), \boldsymbol{L}^2(\mathcal{M}))$ as $\delta \to 0$, independent of $\boldsymbol{\psi}$.

Next we consider the asymptotic behavior of the operator $L_\delta^\top$ from (4.3). Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ and $\boldsymbol{\psi} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$. For $X \in \{B, D_\delta\}$ we denote by $\langle \cdot, \cdot \rangle_{\partial X}$ the duality pairing between $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial X)$ and $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial X)$. Using (7.3) we obtain

$$\begin{aligned}\langle L_\delta^\top \boldsymbol{\varphi}, \boldsymbol{\psi} \rangle_{\partial D_\delta} &= \langle \boldsymbol{\varphi}, L_\delta \boldsymbol{\psi} \rangle_{\mathcal{M}} = \langle \delta^2 L_0^\top \boldsymbol{\varphi} + \delta^3 L_1^\top \boldsymbol{\varphi} + E_L^\top \boldsymbol{\varphi}, \hat{\boldsymbol{\psi}} \rangle_{\partial B} \\ &= \langle (L_0^\top \boldsymbol{\varphi})^\vee + \delta (L_1^\top \boldsymbol{\varphi})^\vee + \delta^{-2} (E_L^\top \boldsymbol{\varphi})^\vee, \boldsymbol{\psi} \rangle_{\partial D_\delta},\end{aligned}$$

where $L_0^\top, L_1^\top, E_L^\top : \boldsymbol{L}^2(\mathcal{M}) \to \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)$ are the dual operators of L_0, L_1, and E_L, respectively. Because by duality $E_L^\top$ is $\mathcal{O}(\delta^4)$ in $\mathcal{L}(\boldsymbol{L}^2(\mathcal{M}), \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B))$, we obtain the following asymptotic behavior.

Proposition 7.2. *For all $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$,*

$$L_\delta^\top \boldsymbol{\varphi} = (L_0^\top \boldsymbol{\varphi})^\vee + \delta (L_1^\top \boldsymbol{\varphi})^\vee + \delta^{-2} (E_L^\top \boldsymbol{\varphi})^\vee,$$

where $E_L^\top$ is a bounded linear operator, which is $\mathcal{O}(\delta^4)$ in $\mathcal{L}(\boldsymbol{L}^2(\mathcal{M}), \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B))$ as $\delta \to 0$, independent of $\boldsymbol{\varphi}$.

Now we calculate the operators $L_0^\top$ and $L_1^\top$ explicitly. Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ and $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$. Recalling the definition of the operator L_0 from (7.1) we find

$$\begin{aligned}\langle \boldsymbol{\varphi}, L_0 \boldsymbol{a} \rangle_{\mathcal{M}} &= \int_{\mathcal{M}} \Big(-\mathrm{i}\,\omega \varepsilon_- \mathbb{G}^m(\boldsymbol{x}, \boldsymbol{z}) \int_{\partial B} \Big(\Big(\frac{1}{2} I + M_B^0 \Big)^{-1} \boldsymbol{a} \Big)(\boldsymbol{\eta}) \,\mathrm{d}s(\boldsymbol{\eta}) \Big) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \,\mathrm{d}s(\boldsymbol{x}) \\ &= \Big(-\mathrm{i}\,\omega \varepsilon_- \int_{\mathcal{M}} \mathbb{G}^{m\top}(\boldsymbol{x}, \boldsymbol{z}) \boldsymbol{\varphi}(\boldsymbol{x}) \,\mathrm{d}s(\boldsymbol{x}) \Big) \cdot \int_{\partial B} \Big(\Big(\frac{1}{2} I + M_B^0 \Big)^{-1} \boldsymbol{a} \Big)(\boldsymbol{\eta}) \,\mathrm{d}s(\boldsymbol{\eta}).\end{aligned}$$

Recalling (3.4b) and (3.5), we obtain

$$\begin{aligned}\langle \boldsymbol{\varphi}, L_0 \boldsymbol{a} \rangle_{\mathcal{M}} &= \frac{1}{\mathrm{i}\,\omega \mu_+} \boldsymbol{H}^i(\boldsymbol{z}) \cdot \int_{\partial B} \Big(\Big(\frac{1}{2} I + M_B^0 \Big)^{-1} \boldsymbol{a} \Big)(\boldsymbol{\eta}) \,\mathrm{d}s(\boldsymbol{\eta}) \\ &= \int_{\partial B} \frac{1}{\mathrm{i}\,\omega \mu_+} \Big(\Big(\frac{1}{2} I + M_B^{0\top} \Big)^{-1} \pi_t \big(\boldsymbol{H}^i(\boldsymbol{z}) \big) \Big)(\boldsymbol{\xi}) \cdot \boldsymbol{a}(\boldsymbol{\xi}) \,\mathrm{d}s(\boldsymbol{\xi}),\end{aligned}$$

where π_t denotes the projection on the tangent plane to ∂B. Therefore, we have

$$L_0^\top \boldsymbol{\varphi} = \frac{1}{\mathrm{i}\,\omega \mu_+} \Big(\frac{1}{2} I + M_B^{0\top} \Big)^{-1} \pi_t (\boldsymbol{H}^i(\boldsymbol{z})). \tag{7.4}$$

In the same way we obtain from (7.2) that

$$\langle \boldsymbol{\varphi}, L_1 \boldsymbol{a} \rangle_{\mathcal{M}} = \int_{\partial B} \frac{1}{\mathrm{i}\,\omega \mu_+} \Big(\Big(\frac{1}{2} I + M_B^{0\top} \Big)^{-1} \pi_t \Big(\sum_{l=1}^{3} \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z}) \Big) \Big)(\boldsymbol{\xi}) \cdot \boldsymbol{a}(\boldsymbol{\xi}) \,\mathrm{d}s(\boldsymbol{\xi}).$$

Here, η_l denotes the lth component of the surface variable on ∂B. Thus we have

$$L_1^\top \boldsymbol{\varphi} = \frac{1}{\mathrm{i}\,\omega \mu_+} \Big(\frac{1}{2} I + M_B^{0\top} \Big)^{-1} \pi_t \Big(\sum_{l=1}^{3} \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z}) \Big). \tag{7.5}$$

We return to the diffraction problem (4.4) and the operator F_δ from (4.5). Given $\boldsymbol{\chi} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta)$, we define

$$\begin{aligned} \boldsymbol{E}^d &:= -\frac{1}{\mathrm{i}\,\omega\varepsilon}\mathbf{curl}\frac{\mu_-}{\mu}\mathbf{curl}\mathcal{A}^e_{D_\delta}(\boldsymbol{\nu}\times\boldsymbol{\chi}) && \text{in } \mathbb{R}^3_0 \setminus \partial D_\delta, \\ \boldsymbol{H}^d &:= \frac{\mu_-}{\mu}\mathbf{curl}\mathcal{A}^e_{D_\delta}(\boldsymbol{\nu}\times\boldsymbol{\chi}) && \text{in } \mathbb{R}^3_0 \setminus \partial D_\delta. \end{aligned}$$

Then $(\boldsymbol{E}^d, \boldsymbol{H}^d)$ is a radiating solution to (4.4), and recalling (3.3), (5.1), and (2.1) we find

$$\begin{aligned} \boldsymbol{\nu}\times\boldsymbol{E}^d\big|_{\partial D_\delta}(\boldsymbol{x}) &= -\frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}_x\,\mathbf{curl}_x\int_{\partial D_\delta}\Pi^e(\boldsymbol{x},\boldsymbol{y})(\boldsymbol{\nu}\times\boldsymbol{\chi})(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}) \\ &= \mathrm{i}\,\omega\mu_-\boldsymbol{\nu}(\boldsymbol{x})\times\int_{\partial D_\delta}\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{\nu}\times\boldsymbol{\chi})(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}) \\ &\quad+\frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}(\boldsymbol{x})\times\int_{\partial D_\delta}\nabla_x\Phi_{k_-}(\boldsymbol{x}-\boldsymbol{y})(\mathrm{curl}_{\partial D_\delta}\boldsymbol{\chi})(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y}) \\ &\quad-\frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}(\boldsymbol{x})\times\mathbf{curl}_x\,\mathbf{curl}_x\int_{\partial D_\delta}F^e(\boldsymbol{x},\boldsymbol{y})(\boldsymbol{\nu}\times\boldsymbol{\chi})(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y})\,. \end{aligned}$$

Remark 7.3. The previous formula employs a slight abuse of notation, because pointwise evaluation is not defined for elements of $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$. However, we included the argument for better readability.

Define $P_{D_\delta} : \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial D_\delta)$ by

$$P_{D_\delta}\boldsymbol{a} := -\frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}\times\mathbf{curl}\,\mathbf{curl}\int_{\partial D_\delta}F^e(\cdot,\boldsymbol{y})(\boldsymbol{\nu}\times\boldsymbol{a})(\boldsymbol{y})\,\mathrm{d}s(\boldsymbol{y})$$

for $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta)$. Then we can see as in the proof of Lemma 6.2 that

$$\|(P_{D_\delta}\boldsymbol{a})^\wedge\|_{\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)} \le C\delta^2\|\hat{\boldsymbol{a}}\|_{\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)}.$$

Therefore, by change of coordinates, applying (6.3) and (6.4), we obtain

$$\begin{aligned} (\boldsymbol{\nu}\times\boldsymbol{E}^d\big|_{\partial D_\delta})^\wedge(\boldsymbol{\xi}) &= \delta^{-1}\frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}(\boldsymbol{\xi})\times\int_{\partial B}\nabla_{\boldsymbol{\xi}}\Phi_0(\boldsymbol{\xi}-\boldsymbol{\eta})(\mathrm{curl}_{\partial B}\,\hat{\boldsymbol{\chi}})(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}) \\ &\quad+\delta\,\mathrm{i}\,\omega\mu_-\boldsymbol{\nu}(\boldsymbol{\xi})\times\int_{\partial B}\Phi_0(\boldsymbol{\xi}-\boldsymbol{\eta})(\boldsymbol{\nu}\times\hat{\boldsymbol{\chi}})(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}) \\ &\quad+\delta\,\mathrm{i}\,\omega\mu_-\boldsymbol{\nu}(\boldsymbol{\xi})\times\int_{\partial B}\frac{1}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|}(\mathrm{curl}_{\partial B}\,\hat{\boldsymbol{\chi}})(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})+\mathcal{O}(\delta^2) \end{aligned}$$

as $\delta \to 0$. The $\mathcal{O}(\delta^2)$-term in (6.3) and the $\mathcal{O}(\delta^3)$-term in (6.4) define pseudohomogeneous kernels of class -3 (cf. [32, pp. 168–174]); i.e., the corresponding integral operators are continuous from $H^{-1/2}(\partial B)^3$ into $H^{5/2}(\partial B)^3$ (cf. [32, Thm. 4.4.1]). Thus, these operators are also continuous from $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)$ into $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$, and together with the (constant) $\mathcal{O}(\delta)$-term in (6.3) they lead to terms of order $\mathcal{O}(\delta^2)$ in $\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$ in the asymptotic expansion of $\boldsymbol{\nu}\times\boldsymbol{E}^d|_{\partial D_\delta}$ as $\delta \to 0$.

We define $F_0 : \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$,

$$(F_0\boldsymbol{a})(\boldsymbol{\xi}) := \frac{1}{\mathrm{i}\,\omega\varepsilon_-}\boldsymbol{\nu}(\boldsymbol{\xi})\times\int_{\partial B}\nabla_{\boldsymbol{\xi}}\Phi_0(\boldsymbol{\xi}-\boldsymbol{\eta})(\mathrm{curl}_{\partial B}\,\boldsymbol{a})(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}), \tag{7.6}$$

and $F_1 : \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B) \to \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$,

$$(7.7)\quad (F_1\boldsymbol{a})(\boldsymbol{\xi}) := \mathrm{i}\,\omega\mu_-\big(\boldsymbol{\nu}\times\mathcal{A}_B^0(\boldsymbol{\nu}\times\boldsymbol{a})\big|_{\partial B}\big)(\boldsymbol{\xi}) + \mathrm{i}\,\omega\mu_-\boldsymbol{\nu}(\boldsymbol{\xi})\times\int_{\partial B}\frac{1}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|}(\mathrm{curl}_{\partial B}\,\boldsymbol{a})(\boldsymbol{\eta})\;\mathrm{d}s(\boldsymbol{\eta}).$$

Note that $-\mathrm{i}\,\omega\varepsilon_- F_0 = N_B^0$. Thus, F_0 and the first part of F_1 are bounded. Because the kernel of the second part of F_1 is homogeneous of class -2 (cf. [32, sec. 4.3.2]), the second part of F_1 is continuous also. We obtain the following asymptotic behavior.

Proposition 7.4. *For all $\boldsymbol{\chi}\in\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial D_\delta)$,*

$$F_\delta\boldsymbol{\chi} = \delta^{-1}(F_0\hat{\boldsymbol{\chi}})^\vee + \delta(F_1\hat{\boldsymbol{\chi}})^\vee + (E_F\hat{\boldsymbol{\chi}})^\vee,$$

where E_F is a bounded linear operator, which is $\mathcal{O}(\delta^2)$ in $\mathcal{L}(\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B), \boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B))$ as $\delta\to 0$, independent of $\boldsymbol{\chi}$.

Next we consider the boundary value problem of finding $\boldsymbol{u}\in\boldsymbol{H}(\mathbf{curl}, B)$ such that

$$(7.8a)\qquad \mathbf{curl}\,\mathbf{curl}\boldsymbol{u} = 0 \qquad \text{in } B,$$

$$(7.8b)\qquad \operatorname{div}\boldsymbol{u} = 0 \qquad \text{in } B,$$

$$(7.8c)\qquad \boldsymbol{\nu}\times\boldsymbol{u} = \boldsymbol{c} \qquad \text{on } \partial B,$$

where $\boldsymbol{c}\in\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$ is a given tangential function. We show that (7.8) has at most one solution and use this fact to prove that $F_0L_0^\top = 0$ on $\boldsymbol{L}^2(\mathcal{M})$ and $L_0F_0 = 0$ on $\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)$.

Lemma 7.5. *Let $\boldsymbol{c}\in\boldsymbol{H}_{\mathrm{div}}^{-1/2}(\partial B)$. Then the boundary value problem* (7.8) *has at most one solution in $\boldsymbol{H}(\mathbf{curl}, B)$.*

Proof. Using integration by parts we find for any solution $\boldsymbol{u}\in\boldsymbol{H}(\mathbf{curl}, B)$ of (7.8) with homogeneous boundary condition $\boldsymbol{c} = 0$ that

$$\begin{aligned}0 &= \int_B \mathbf{curl}\,\mathbf{curl}\boldsymbol{u}(\boldsymbol{x})\cdot\overline{\boldsymbol{u}(\boldsymbol{x})}\;\mathrm{d}\boldsymbol{x}\\ &= \int_B |\mathbf{curl}\boldsymbol{u}(\boldsymbol{x})|^2\;\mathrm{d}\boldsymbol{x} + \big\langle\gamma_t(\mathbf{curl}\boldsymbol{u}), \overline{\pi_t(\boldsymbol{u})}\big\rangle_{\partial B} = \int_B |\mathbf{curl}\boldsymbol{u}(\boldsymbol{x})|^2\;\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Hence, $\mathbf{curl}\boldsymbol{u} = 0$ in B, and because the boundaries of all components of B are assumed to be connected, we obtain from [31, Thms. 3.41 and 3.42] a scalar potential $p\in H^1(B)$ with $\gamma_0(p) = 0$ on ∂B such that $\boldsymbol{u} = \nabla p$. Finally, because $\Delta p = \operatorname{div}\boldsymbol{u} = 0$ in B by (7.8b), we have $p = 0$ in B. Hence, $\boldsymbol{u} = \nabla p = 0$ in B. □

Proposition 7.6. *Let $\boldsymbol{\varphi}\in\boldsymbol{L}^2(\mathcal{M})$ and $\boldsymbol{a}\in\boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)$. Then $F_0L_0^\top\boldsymbol{\varphi} = 0$ and $L_0F_0\boldsymbol{a} = 0$.*

Proof. Given $\boldsymbol{\varphi}\in\boldsymbol{L}^2(\mathcal{M})$, by (7.4) and (7.6) we find that on ∂B

$$F_0L_0^\top\boldsymbol{\varphi} = \frac{1}{\omega^2\varepsilon_-\mu_+}\boldsymbol{\nu}\times\mathbf{curl}\,\mathbf{curl}\mathcal{A}_B^0\Big(\boldsymbol{\nu}\times\Big(\frac{1}{2}I + {M_B^0}^\top\Big)^{-1}\pi_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big),$$

where $\boldsymbol{H}^i(\boldsymbol{z})$ is given by (3.5). An easy computation applying (5.2) shows that

$$(7.9)\qquad \pm\boldsymbol{\nu}\times\Big(\pm\frac{1}{2}I + {M_B^0}^\top\Big)^{-1}\pi_t(\cdot) = \mp\Big(\mp\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t(\cdot)\,.$$

Therefore,

$$F_0 L_0^\top \boldsymbol{\varphi} = -\frac{1}{\omega^2 \varepsilon_- \mu_+} \boldsymbol{\nu} \times \mathbf{curl}\,\mathbf{curl} \mathcal{A}_B^0 \Big(-\frac{1}{2} I + M_B^0\Big)^{-1} \gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big).$$

Now let

$$\boldsymbol{u} := \mathbf{curl} \mathcal{A}_B^0 \Big(-\frac{1}{2} I + M_B^0\Big)^{-1} \gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big) \qquad \text{in } B;$$

then $\boldsymbol{u}$ is a solution to (7.8) with $\boldsymbol{c} = \gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)$. From Lemma 7.5 we obtain that $\boldsymbol{u} = \boldsymbol{H}^i(\boldsymbol{z})$ is constant in B. Hence,

$$F_0 L_0^\top \boldsymbol{\varphi} = -\frac{1}{\omega^2 \varepsilon_- \mu_+} \gamma_t(\mathbf{curl} \boldsymbol{u}) = 0 \qquad \text{on } \partial B.$$

Because F_0 is symmetric, also $L_0 F_0 \boldsymbol{a} = 0$ for each $\boldsymbol{a} \in \boldsymbol{H}_{\mathrm{curl}}^{-1/2}(\partial B)$. □

Recalling Theorem 4.2, we can we put our results together and obtain the following asymptotic expansion of the measurement operator G_δ.

THEOREM 7.7. *Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$; then*

$$G_\delta \boldsymbol{\varphi} = \mathrm{i}\,\omega \mu_+ \delta^3 \left(L_0 F_1 L_0^\top \boldsymbol{\varphi} + L_1 F_0 L_1^\top \boldsymbol{\varphi}\right) + \mathcal{O}(\delta^4)$$

in $\boldsymbol{L}^2(\mathcal{M})$ as $\delta \to 0$. More precisely, the last term on the right-hand side is bounded by $C\delta^4 \|\boldsymbol{\varphi}\|_{\boldsymbol{L}^2(\mathcal{M})}$, where the constant C is independent of δ and $\boldsymbol{\varphi}$.

The proof of this theorem follows straightforwardly from the previous propositions and Theorem 4.2. We refer the reader to [1, Thm. 5.9] for a similar proof in the electrostatic case.

Finally, for $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$, we are going to calculate $L_0 F_1 L_0^\top \boldsymbol{\varphi}$ and $L_1 F_0 L_1^\top \boldsymbol{\varphi}$ explicitly.

LEMMA 7.8. *For each $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ we have*

$$F_0 L_1^\top \boldsymbol{\varphi} = -\frac{1}{\omega^2 \varepsilon_- \mu_+} \gamma_t\big(\mathbf{curl} \boldsymbol{H}^i(\boldsymbol{z})\big) \qquad \text{on } \partial B. \tag{7.10}$$

Proof. Given $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$, by (7.5) and (7.6), applying (7.9) we find

$$\begin{aligned} F_0 L_1^\top \boldsymbol{\varphi} &= \frac{1}{\omega^2 \varepsilon_- \mu_+} \boldsymbol{\nu} \times \mathbf{curl}\,\mathbf{curl} \mathcal{A}_B^0 \bigg(\boldsymbol{\nu} \times \Big(\frac{1}{2} I + M_B^{0\,\top}\Big)^{-1} \pi_t \Big(\sum_{l=1}^3 \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z})\Big)\bigg) \\ &= -\frac{1}{\omega^2 \varepsilon_- \mu_+} \boldsymbol{\nu} \times \mathbf{curl}\,\mathbf{curl} \mathcal{A}_B^0 \Big(-\frac{1}{2} I + M_B^0\Big)^{-1} \gamma_t \Big(\sum_{l=1}^3 \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z})\Big). \end{aligned}$$

Let

$$\boldsymbol{u} := \mathbf{curl} \mathcal{A}_B^0 \Big(-\frac{1}{2} I + M_B^0\Big)^{-1} \gamma_t \Big(\sum_{l=1}^3 \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z})\Big) \qquad \text{in } B;$$

then $\boldsymbol{u}$ is a solution to (7.8) with $\boldsymbol{c} = \gamma_t(\sum_{l=1}^3 \eta_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z}))$. From Lemma 7.5 we obtain $\boldsymbol{u}(\boldsymbol{\xi}) = \sum_{l=1}^3 \xi_l \frac{\partial \boldsymbol{H}^i}{\partial y_l}(\boldsymbol{z})$ for a.e. $\boldsymbol{\xi} \in B$. An easy calculation shows that therefore $\mathbf{curl} \boldsymbol{u} = \mathbf{curl} \boldsymbol{H}^i(\boldsymbol{z})$ in B, which ends the proof. □

Lemma 7.9. *For each $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ we have*

$$F_1 L_0^\top \boldsymbol{\varphi} = -\frac{\mu_-}{\mu_+}\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big) \qquad \text{on } \partial B. \tag{7.11}$$

Proof. Given $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$, by (7.4) and (7.7), applying (2.1) and (7.9) we find that on ∂B

$$\begin{aligned}(F_1 L_0^\top \boldsymbol{\varphi})(\boldsymbol{\xi}) =& -\frac{\mu_-}{\mu_+}\Big(\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\xi})\\ &+\frac{\mu_-}{\mu_+}\boldsymbol{\nu}(\boldsymbol{\xi}) \times \int_{\partial B}\frac{1}{8\pi}\frac{\boldsymbol{\xi}-\boldsymbol{\eta}}{|\boldsymbol{\xi}-\boldsymbol{\eta}|}\Big(\mathrm{div}_{\partial B}\Big(\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}).\end{aligned} \tag{7.12}$$

By Lemma 5.1(a), $-\frac{1}{2}I + M_B^0$ is an isomorphism on $\boldsymbol{H}^{-1/2}_{\mathrm{div},0}(\partial B)$. Therefore, because by (2.3) it holds that $\mathrm{div}_{\partial B}\,\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big) = -\gamma_n\big(\mathbf{curl}\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\big) = 0$, we find

$$\mathrm{div}_{\partial B}\Big(\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big) = 0.$$

Hence, the second term on the right-hand side of (7.12) vanishes, and we obtain the desired result. □

Definition 7.10. *For a bounded $C^{2,\alpha}$ domain $D \subset \mathbb{R}^3$ we define the magnetic polarizability tensor $\mathbb{M}_D^0 \in \mathbb{R}^{3\times 3}$ by $\mathbb{M}_D^0 := (m_{ij}^0)_{i,j=1}^3$ with*

$$m_{ij}^0 := -\int_{\partial D}\eta_j\Big(\Big(-\frac{1}{2}I + {K_D^0}^\top\Big)^{-1}\nu_i\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}), \qquad 1 \le i,j \le 3.$$

The electric polarizability tensor $\mathbb{M}_D^\infty \in \mathbb{R}^{3\times 3}$ corresponding to the domain D is given by $\mathbb{M}_D^\infty := (m_{ij}^\infty)_{i,j=1}^3$ with

$$m_{ij}^\infty := \int_{\partial D}\eta_j\Big(\Big(\frac{1}{2}I + {K_D^0}^\top\Big)^{-1}\nu_i\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta}), \qquad 1 \le i,j \le 3.$$

The magnetic and the electric polarizability tensor are symmetric and positive definite matrices; cf. [3, 19].

Proposition 7.11. *For each $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ we have*

$$L_1 F_0 L_1^\top \boldsymbol{\varphi} = \frac{1}{\mathrm{i}\,\omega\mu_+}\frac{\mu_-}{\mu_+}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z})\mathbb{M}_B^\infty\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}) \qquad \text{on } \mathcal{M}.$$

Proof. Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$. By (7.10) and (7.2),

$$L_1 F_0 L_1^\top \boldsymbol{\varphi} = -\frac{1}{\mathrm{i}\,\omega\mu_+}\int_{\partial B}\sum_{l=1}^3 \eta_l\frac{\partial\mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z})\Big(\Big(\frac{1}{2}I{+}M_B^0\Big)^{-1}\gamma_t\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})$$

on $\mathcal{M}$. Applying (7.9), we find

$$\begin{aligned}&-\int_{\partial B}\sum_{l=1}^3 \eta_l\frac{\partial\mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z})\Big(\Big(\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})\\ &\quad= \sum_{l=1}^3\frac{\partial\mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z})\int_{\partial B}\eta_l\,\mathbb{I}_3\Big(\boldsymbol{\nu}\times\Big(-\frac{1}{2}I + {M_B^0}^\top\Big)^{-1}\pi_t\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\eta})\,\mathrm{d}s(\boldsymbol{\eta})\end{aligned}$$

on $\mathcal{M}$. Because by (2.2)

$$\pi_t\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\big) = \pi_t\big(\nabla_\eta\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\cdot\boldsymbol{\eta}\big)\big) = \nabla_{\partial B}\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\cdot\boldsymbol{\eta}\big)$$

on ∂B, where $\boldsymbol{\eta}$ denotes the coordinate function in a neighborhood of ∂B, we can apply (5.3) and (2.2) to obtain for $1 \le l \le 3$ that

$$\begin{aligned}\int_{\partial B} \eta_l\, \mathbb{I}_3 \Big(\boldsymbol{\nu} \times \Big(-\frac{1}{2}I + M_B^{0\,\top}\Big)^{-1} \pi_t\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\eta})\ \mathrm{d}s(\boldsymbol{\eta})&\\ = \int_{\partial B} \pi_t(\eta_l\, \mathbb{I}_3)^\top \Big(\mathbf{curl}_{\partial B}\Big(\frac{1}{2}I + K_B^0\Big)^{-1}\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\cdot\boldsymbol{\eta}\big)\Big)(\boldsymbol{\eta})\ \mathrm{d}s(\boldsymbol{\eta}).&\end{aligned}$$

From the duality of $\mathbf{curl}_{\partial B}$ and $\mathrm{curl}_{\partial B}$, and from (2.3), we find

$$\begin{aligned}\int_{\partial B} \pi_t(\eta_l\, \mathbb{I}_3)^\top \Big(\mathbf{curl}_{\partial B}\Big(\frac{1}{2}I + K_B^0\Big)^{-1}\big(\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\cdot\boldsymbol{\eta}\big)\Big)(\boldsymbol{\eta})\ \mathrm{d}s(\boldsymbol{\eta})&\\ = \int_{\partial B} \big(\boldsymbol{\nu}\cdot\mathbf{curl}(\eta_l\, \mathbb{I}_3)\big)^\top(\boldsymbol{\xi}) \bigg(\Big(\Big(\frac{1}{2}I + K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\xi})\cdot\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\bigg)\ \mathrm{d}s(\boldsymbol{\xi}).&\end{aligned}$$

An easy calculation reveals

$$\sum_{l=1}^{3} \frac{\partial \mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z})\mathbf{curl}_\eta(\eta_l\, \mathbb{I}_3)^\top = \big(\mathbf{curl}_y \mathbb{G}^{m\top}\big)^\top(\cdot,\boldsymbol{z}).$$

Applying (3.4b) and (3.4c), we observe

$$\big(\mathbf{curl}_y \mathbb{G}^{m\top}\big)^\top(\cdot,\boldsymbol{z}) = \frac{\mu_-}{\mu_+}\mathbf{curl}_x \mathbb{G}^e(\cdot,\boldsymbol{z}).$$

So, we find

$$\begin{aligned}\sum_{l=1}^{3} \frac{\partial \mathbb{G}^m}{\partial y_l}(\cdot,\boldsymbol{z}) \int_{\partial B} \big(\boldsymbol{\nu}\cdot\mathbf{curl}(\eta_l\, \mathbb{I}_3)\big)^\top(\boldsymbol{\xi}) \bigg(\Big(\Big(\frac{1}{2}I + K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\xi})\cdot\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\bigg)\ \mathrm{d}s(\boldsymbol{\xi})&\\ = \frac{\mu_-}{\mu_+}\mathbf{curl}_x \mathbb{G}^e(\cdot,\boldsymbol{z}) \int_{\partial B} \boldsymbol{\nu}(\boldsymbol{\xi}) \bigg(\Big(\Big(\frac{1}{2}I + K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\xi})\cdot\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\bigg)\ \mathrm{d}s(\boldsymbol{\xi})&\\ = \frac{\mu_-}{\mu_+}\mathbf{curl}_x \mathbb{G}^e(\cdot,\boldsymbol{z})\mathbb{M}_B^\infty \mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}). \qquad \square&\end{aligned}$$

PROPOSITION 7.12. *For each $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ we have*

$$L_0 F_1 L_0^\top \boldsymbol{\varphi} = \mathrm{i}\,\omega\varepsilon_- \frac{\mu_-}{\mu_+}\mathbb{G}^m(\cdot,\boldsymbol{z})\mathbb{M}_B^0 \boldsymbol{H}^i(\boldsymbol{z}).$$

Proof. Let $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$ and set

$$\boldsymbol{u} := \mathbf{curl}\mathcal{A}_B^0\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(z)\big) \qquad \text{in } B.$$

As in the proof of Proposition 7.6 we find that $\boldsymbol{u} = \boldsymbol{H}^i(\boldsymbol{z})$ in B. So we obtain

$$\gamma_n\Big(\mathbf{curl}\mathcal{A}_B^0\Big(-\frac{1}{2}I + M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(z)\big)\Big) = \gamma_n\big(\boldsymbol{H}^i(\boldsymbol{z})\big) \qquad \text{on } \partial B. \tag{7.13}$$

By (7.11) and (7.1),

$$
\begin{aligned}
&L_0 F_1 L_0^\top \boldsymbol{\varphi}\\
&= \mathrm{i}\,\omega\varepsilon_- \frac{\mu_-}{\mu_+}\mathbb{G}^m(\cdot,\boldsymbol{z})\int_{\partial B}\Big(\Big(\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\eta})\;\mathrm{d}s(\boldsymbol{\eta})
\end{aligned}
$$

on $\mathcal{M}$. Observing that $\pi_t(\mathbb{I}_3) = \nabla_{\partial B}\boldsymbol{\eta}$ on ∂B, where $\boldsymbol{\eta}$ again denotes the surface variable on ∂B, and applying (2.2), we can calculate

$$
\begin{aligned}
&\int_{\partial B}\Big(\Big(\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\eta})\;\mathrm{d}s(\boldsymbol{\eta})\\
&\quad=\int_{\partial B}\Big(\Big(\frac{1}{2}I+{M_B^0}^\top\Big)^{-1}\nabla_{\partial B}\boldsymbol{\eta}\Big)^\top(\boldsymbol{\xi})\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\xi})\;\mathrm{d}s(\boldsymbol{\xi}).
\end{aligned}
$$

Applying (5.3), the duality of $-\nabla_{\partial B}$ and $\mathrm{div}_{\partial B}$, and (2.3), we have

$$
\begin{aligned}
&\int_{\partial B}\Big(\Big(\frac{1}{2}I+{M_B^0}^\top\Big)^{-1}\nabla_{\partial B}\boldsymbol{\eta}\Big)^\top(\boldsymbol{\xi})\gamma_t\Big(\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)(\boldsymbol{\xi})\;\mathrm{d}s(\boldsymbol{\xi})\\
&=\int_{\partial B}\Big(\Big(-\frac{1}{2}I+K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\eta})\Big(-\gamma_n\Big(\mathbf{curl}\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\eta})\;\mathrm{d}s(\boldsymbol{\eta}).
\end{aligned}
$$

Finally, recalling (7.13), we obtain

$$
\begin{aligned}
&\int_{\partial B}\Big(\Big(-\frac{1}{2}I+K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\eta})\Big(-\gamma_n\Big(\mathbf{curl}\mathcal{A}_B^0\Big(-\frac{1}{2}I+M_B^0\Big)^{-1}\gamma_t\big(\boldsymbol{H}^i(\boldsymbol{z})\big)\Big)\Big)(\boldsymbol{\eta})\;\mathrm{d}s(\boldsymbol{\eta})\\
&\quad=-\int_{\partial B}\Big(\Big(-\frac{1}{2}I+K_B^0\Big)^{-1}\boldsymbol{\eta}\Big)(\boldsymbol{\eta})\big(\boldsymbol{\nu}(\boldsymbol{\eta})\cdot\boldsymbol{H}^i(\boldsymbol{z})\big)\;\mathrm{d}s(\boldsymbol{\eta})=\mathbb{M}_B^0\boldsymbol{H}^i(\boldsymbol{z}). \qquad\square
\end{aligned}
$$

From Theorem 7.7 and Propositions 7.11 and 7.12 we obtain the following corollary.

Corollary 7.13. *Let $\boldsymbol{\varphi}\in\boldsymbol{L}^2(\mathcal{M})$, and let $\boldsymbol{H}^i$ be the corresponding incident field from* (3.5). *Then*

$$
G_\delta\boldsymbol{\varphi}=\delta^3\Big(-k_-^2\mathbb{G}^m(\cdot,\boldsymbol{z})\mathbb{M}_B^0\boldsymbol{H}^i(\boldsymbol{z})+\frac{\mu_-}{\mu_+}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z})\mathbb{M}_B^\infty\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z})\Big)+\mathcal{O}(\delta^4)
$$

in $\boldsymbol{L}^2(\mathcal{M})$, as $\delta\to 0$. More precisely, the last term on the right-hand side is bounded by $C\delta^4\|\boldsymbol{\varphi}\|_{\boldsymbol{L}^2(\mathcal{M})}$, where the constant C is independent of δ and $\boldsymbol{\varphi}$.

8. Multiple scatterers.

The results of the previous sections can be extended to the practically important case of finitely many well-separated small scatterers as introduced in section 3. This generalization works in the same way as we did in [1, section 6] for the electrostatic case. Because the calculations are rather technical and no new ideas are needed, we just mention the final result and leave the details to the reader.

Let $\mathbb{M}_{B_1}^0,\dots,\mathbb{M}_{B_m}^0$ and $\mathbb{M}_{B_1}^\infty,\dots,\mathbb{M}_{B_m}^\infty$ denote the magnetic and electric polarizability tensors corresponding to $B_1,\dots,B_m$, respectively. In case of multiple scatterers Corollary 7.13 reads as follows.

Corollary 8.1. *Let $\boldsymbol{\varphi}\in\boldsymbol{L}^2(\mathcal{M})$, and let $\boldsymbol{H}^i$ be the corresponding incident field from* (3.5). *Then*

(8.1)

$$
G_\delta\boldsymbol{\varphi}=\delta^3\sum_{l=1}^m\Big(-k_-^2\mathbb{G}^m(\cdot,\boldsymbol{z}_l)\mathbb{M}_{B_l}^0\boldsymbol{H}^i(\boldsymbol{z}_l)+\frac{\mu_-}{\mu_+}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z}_l)\mathbb{M}_{B_l}^\infty\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}_l)\Big)+\mathcal{O}(\delta^4)
$$

in $\boldsymbol{L}^2(\mathcal{M})$, as $\delta \to 0$. More precisely, the last term on the right-hand side is bounded by $C\delta^4\|\boldsymbol{\varphi}\|_{\boldsymbol{L}^2(\mathcal{M})}$, where the constant C is independent of δ and $\boldsymbol{\varphi}$.

9. A characterization of the scatterers.

Using the asymptotic formula (8.1) we can now derive a characterization of the centers of the scatterers $\boldsymbol{z}_1, \dots, \boldsymbol{z}_l$ using a range criterion. For this purpose we introduce the operator $T : \boldsymbol{L}^2(\mathcal{M}) \to \boldsymbol{L}^2(\mathcal{M})$ describing the leading order term in the asymptotic expansion (8.1), given by

$$T\boldsymbol{\varphi} := \sum_{l=1}^{m}\Big(-k_-^2\,\mathbb{G}^m(\cdot,\boldsymbol{z}_l)\mathbb{M}^0_{B_l}\boldsymbol{H}^i(\boldsymbol{z}_l) + \frac{\mu_-}{\mu_+}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z}_l)\mathbb{M}^\infty_{B_l}\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}_l)\Big). \tag{9.1}$$

Because (3.5) implies that $\boldsymbol{H}^i$ depends linearly on $\boldsymbol{\varphi}$, it follows that T is linear. From Corollary 8.1 we obtain

$$G_\delta = \delta^3 T + \mathcal{O}(\delta^4) \tag{9.2}$$

as $\delta \to 0$ in $\mathcal{L}(\boldsymbol{L}^2(\mathcal{M}))$. Next we define the operator $R : \mathbb{C}^{3\times 2m} \to \boldsymbol{L}^2(\mathcal{M})$:

$$R\boldsymbol{a} := k_-^2\sum_{l=1}^{m}\Big(\mathbb{G}^m(\cdot,\boldsymbol{z}_l)\boldsymbol{a}_l + \frac{\mu_-}{\mu_+}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z}_l)\boldsymbol{a}_{m+l}\Big) \tag{9.3}$$

for $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_{2m}) \in \mathbb{C}^{3\times 2m}$, $\boldsymbol{a}_l \in \mathbb{C}^3$. Endowing $\mathbb{C}^{3\times 2m}$ with the bilinear form $\langle \boldsymbol{a}, \boldsymbol{b}\rangle_{\mathbb{C}^{3\times 2m}} := \sum_{l=1}^{2m}\boldsymbol{a}_l\cdot\boldsymbol{b}_l$ for $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_{2m})$ $\boldsymbol{b} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_{2m}) \in \mathbb{C}^{3\times 2m}$ with $\boldsymbol{a}_l$, $\boldsymbol{b}_l \in \mathbb{C}^3$, using (3.5), (3.4b), and (3.4c), we obtain

$$\begin{aligned}\langle R\boldsymbol{a}, \boldsymbol{\varphi}\rangle_{\mathcal{M}} =& \sum_{l=1}^{m}\boldsymbol{a}_l\cdot k_-^2\int_{\mathcal{M}}\mathbb{G}^{m\top}(\boldsymbol{x},\boldsymbol{z}_l)\boldsymbol{\varphi}(\boldsymbol{x})\,\mathrm{d}s(\boldsymbol{x})\\ &+\sum_{l=1}^{m}\boldsymbol{a}_{l+m}\cdot k_-^2\frac{\mu_-}{\mu_+}\int_{\mathcal{M}}(\mathbf{curl}_x\mathbb{G}^e)^\top(\boldsymbol{x},\boldsymbol{z}_l)\boldsymbol{\varphi}(\boldsymbol{x})\,\mathrm{d}s(\boldsymbol{x})\\ =&\sum_{l=1}^{m}\Big(\boldsymbol{a}_l\cdot\frac{\mu_-}{\mu_+}\boldsymbol{H}^i(\boldsymbol{z}_l) + \boldsymbol{a}_{l+m}\cdot\frac{\mu_-}{\mu_+}\mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}_l)\Big)\end{aligned}$$

for any $\boldsymbol{a} \in \mathbb{C}^{3\times 2m}$ and $\boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M})$. So, $R^\top : \boldsymbol{L}^2(\mathcal{M}) \to \mathbb{C}^{3\times 2m}$ is given by

$$R^\top\boldsymbol{\varphi} = \frac{\mu_-}{\mu_+}\big(\boldsymbol{H}^i(\boldsymbol{z}_1), \dots, \boldsymbol{H}^i(\boldsymbol{z}_m), \mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}_1), \dots, \mathbf{curl}\boldsymbol{H}^i(\boldsymbol{z}_m)\big). \tag{9.4}$$

LEMMA 9.1. (a) *R is injective.* (b) *$R^\top$ is surjective.*

Proof. (a) Suppose $\boldsymbol{a} \in \mathbb{C}^{3\times 2m}$ such that $R\boldsymbol{a} = 0$. Then

$$\tilde{\boldsymbol{H}} := k_-^2\sum_{l=1}^{m}\Big(\mathbb{G}^m(\cdot,\boldsymbol{z}_l)\boldsymbol{a}_l + \frac{\mu_-}{\mu}\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z}_l)\boldsymbol{a}_{m+l}\Big)$$

together with the associated electric field $\tilde{\boldsymbol{E}} := -1/(\mathrm{i}\,\omega\varepsilon)\mathbf{curl}\tilde{\boldsymbol{H}}$ is a radiating solution of Maxwell's equations (3.1) in $\mathbb{R}^3 \setminus \bigcup_{l=1}^m\{\boldsymbol{z}_l\}$ that satisfies $\tilde{\boldsymbol{H}}|_{\mathcal{M}} = 0$. Now we can follow the proof of [20, Thm. 3.2] and obtain $\tilde{\boldsymbol{H}} = 0$ in $\mathbb{R}^3 \setminus \bigcup_{l=1}^m\{\boldsymbol{z}_l\}$.

Let $l \in \{1, \dots, m\}$; then of course $\lim_{t\to 0}\tilde{\boldsymbol{H}}(\boldsymbol{z}_l + t\boldsymbol{b}) = 0$ for any $\boldsymbol{b} \in \mathbb{R}^3$. A short calculation shows that the singularity of $\mathbb{G}^m(\cdot,\boldsymbol{z}_l)$ in $\boldsymbol{z}_l$ is of order 3, while the singularity of $\mathbf{curl}_x\mathbb{G}^e(\cdot,\boldsymbol{z}_l)$ in $\boldsymbol{z}_l$ is of order 2. So, from $\lim_{t\to 0}\mathbb{G}^m(\boldsymbol{z}_l + t\boldsymbol{e}_3,\boldsymbol{z}_l)\boldsymbol{a}_l = 0$ it follows that $\boldsymbol{a}_l = 0$. Indeed, otherwise the singularity of $\mathbb{G}^m(\cdot, \boldsymbol{z}_l)\boldsymbol{a}_l$ at $\boldsymbol{z}_l$ would imply that $\lim_{t\to 0} |\tilde{\boldsymbol{H}}(\boldsymbol{z}_l + t\boldsymbol{e}_3)| = \infty$. Accordingly, $\lim_{t\to 0} \mathbf{curl}_x \mathbb{G}^e(\boldsymbol{z}_l + t\boldsymbol{e}_1, \boldsymbol{z}_l)\boldsymbol{a}_{m+l} = 0$ and $\lim_{t\to 0} \mathbf{curl}_x \mathbb{G}^e(\boldsymbol{z}_l + t\boldsymbol{e}_2, \boldsymbol{z}_l)\boldsymbol{a}_{m+l} = 0$ yield $\boldsymbol{a}_{m+l} = 0$. Because $l \in \{1, \dots, m\}$ was arbitrary, we are done.

(b) This part of the proof follows from part (a) and the well-known relation $\mathcal{R}(R^\top) = \mathcal{N}(R)^a$ between ranges and null spaces of dual operators with finite rank. Here, $\mathcal{N}(R)^a$ denotes the annihilator of $\mathcal{N}(R)$ in $\mathbb{C}^{3\times 2m}$. □

Comparing the formulas (9.3) and (9.4) for R and $R^\top$ and the definition (9.1) of T, we find that these operators are related by $T = RMR^\top$, where the operator $M : \mathbb{C}^{3\times 2m} \to \mathbb{C}^{3\times 2m}$ is given by

$$M\boldsymbol{a} := \frac{\mu_+}{\mu_-}\left(-\mathbb{M}^0_{B_1}\boldsymbol{a}_1, \dots, -\mathbb{M}^0_{B_m}\boldsymbol{a}_m, \frac{1}{k_-^2}\mathbb{M}^\infty_{B_1}\boldsymbol{a}_{m+1}, \dots, \frac{1}{k_-^2}\mathbb{M}^\infty_{B_m}\boldsymbol{a}_{2m}\right).$$

From the positive definiteness of the magnetic and electric polarizability tensors $\mathbb{M}^0_{B_1}, \dots, \mathbb{M}^0_{B_m}$ and $\mathbb{M}^\infty_{B_1}, \dots, \mathbb{M}^\infty_{B_m}$ we conclude that M is invertible. Taking a closer look at the range of T, we first observe that $\mathcal{R}(T) \subset \mathcal{R}(R)$. We show that this inclusion is actually an equality.

PROPOSITION 9.2. *The range of T has dimension $6m$ and is given by*

$$\mathcal{R}(T) = \operatorname{span}_{\mathbb{C}}\left\{\mathbb{G}^m(\cdot, \boldsymbol{z}_l)\boldsymbol{e}_j\,,\, \mathbf{curl}_x\mathbb{G}^e(\cdot, \boldsymbol{z}_l)\boldsymbol{e}_j \;\middle|\; j = 1, 2, 3;\ l = 1, \dots, m\right\}.$$

Proof. The surjectivity of $R^\top$ and M implies $\mathcal{R}(T) = \mathcal{R}(RMR^\top) = \mathcal{R}(R)$. The proposition is then an immediate consequence of (9.3) and Lemma 9.1(a). □

Now we present the main tool for the identification of the positions $\boldsymbol{z}_l$: the characterization of the centers of the scatterers in terms of the range of the leading order term T of the asymptotic expansion of the measurement operator G_δ.

PROPOSITION 9.3. *Let $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2) \in (\mathbb{C}^3 \times \mathbb{C}^3) \setminus \{(0,0)\}$, $\boldsymbol{z} \in \mathbb{R}^3_-$, and*

$$\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} := \big(\mathbb{G}^m(\cdot, \boldsymbol{z})\boldsymbol{d}_1 + \mathbf{curl}_x\mathbb{G}^e(\cdot, \boldsymbol{z})\boldsymbol{d}_2\big)|_{\mathcal{M}}.$$

Then, $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} \in \mathcal{R}(T)$ if and only if $\boldsymbol{z} \in \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}$.

Proof. Assume that $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} \in \mathcal{R}(T)$. As a consequence of Proposition 9.2, $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$ may be represented as

$$\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} = \sum_{l=1}^m \big(\mathbb{G}^m(\cdot, \boldsymbol{z}_l)\boldsymbol{a}_l + \mathbf{curl}_x\mathbb{G}^e(\cdot, \boldsymbol{z}_{l+m})\boldsymbol{a}_{l+m}\big) \qquad \text{on } \mathcal{M},$$

with $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{2m} \in \mathbb{C}^3$. But then both

$$\boldsymbol{H}^a := \sum_{l=1}^m \Big(\mathbb{G}^m(\cdot, \boldsymbol{z}_l)\boldsymbol{a}_l + \frac{\mu_+}{\mu}\mathbf{curl}_x\mathbb{G}^e(\cdot, \boldsymbol{z}_{l+m})\boldsymbol{a}_{l+m}\Big)$$

and

$$\boldsymbol{H}^b := \mathbb{G}^m(\cdot, \boldsymbol{z})\boldsymbol{d}_1 + \frac{\mu_+}{\mu}\mathbf{curl}_x\mathbb{G}^e(\cdot, \boldsymbol{z})\boldsymbol{d}_2,$$

together with their associated electric fields, are radiating solutions of Maxwell's equations (3.1) in $\mathbb{R}^3 \setminus (\bigcup_{l=1}^m\{\boldsymbol{z}_l\} \cup \{\boldsymbol{z}\})$ that coincide on $\mathcal{M}$. So, $\tilde{\boldsymbol{H}} := \boldsymbol{H}^a - \boldsymbol{H}^b$ together with its electric field is a radiating solution of (3.1) in $\mathbb{R}^3 \setminus (\bigcup_{l=1}^m\{\boldsymbol{z}_l\} \cup \{\boldsymbol{z}\})$ that satisfies $\tilde{\boldsymbol{H}}|_{\mathcal{M}} = 0$. Following the proof of [20, Thm. 3.2] we conclude that $\tilde{\boldsymbol{H}} = 0$ everywhere in $\mathbb{R}^3 \setminus (\bigcup_{l=1}^m\{\boldsymbol{z}_l\} \cup \{\boldsymbol{z}\})$. Thus $\boldsymbol{H}^a = \boldsymbol{H}^b$ in $\mathbb{R}^3 \setminus (\bigcup_{l=1}^m\{\boldsymbol{z}_l\} \cup \{\boldsymbol{z}\})$. This is only possible if $\boldsymbol{z} \in \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}$, and we have established the necessity of this condition. The sufficiency follows from Proposition 9.2. □

10. Determining the position of the inhomogeneities. Let $(\cdot,\cdot)_{\boldsymbol{L}^2(\mathcal{M})}$ denote the (complex) scalar product on $\boldsymbol{L}^2(\mathcal{M})$. Because G_δ is a compact operator on $\boldsymbol{L}^2(\mathcal{M})$, it admits a singular value decomposition

$$G_\delta \boldsymbol{\varphi} = \sum_{j=1}^{\infty} \sigma_j^\delta (\boldsymbol{\varphi}, \boldsymbol{v}_j^\delta)_{\boldsymbol{L}^2(\mathcal{M})} \boldsymbol{u}_j^\delta, \qquad \boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M}),$$

where $((\sigma_j^\delta)^2)_{j\in\mathbb{N}}$ are the eigenvalues of $G_\delta^* G_\delta$, written in decreasing order with multiplicity, $\sigma_j^\delta \geq 0$. Similarly, the finite rank operator T can be decomposed as

$$T\boldsymbol{\varphi} = \sum_{l=1}^{6m} \sigma_l (\boldsymbol{\varphi}, \boldsymbol{v}_l)_{\boldsymbol{L}^2(\mathcal{M})} \boldsymbol{u}_l, \qquad \boldsymbol{\varphi} \in \boldsymbol{L}^2(\mathcal{M}),$$

with $s_1 \geq s_2 \geq \cdots \geq s_{6m} > 0$. From (9.2) we obtain

$$G_\delta^* G_\delta = \delta^6 T^* T + \mathcal{O}(\delta^7)$$

in $\mathcal{L}(\boldsymbol{L}^2(\mathcal{M}))$ as $\delta \to 0$. So, applying [26, Thm. V.4.10] we get the following aymptotic formula for the singular values as $\delta \to 0$:

$$(10.1) \qquad (\sigma_j^\delta)^2 = \delta^6 \sigma_j^2 + \mathcal{O}(\delta^7), \qquad j \in \mathbb{N},$$

where we have set $\sigma_l = 0$ for $l \geq 6m$. Next, for $j \in \mathbb{N}$ and $l = 1, \ldots, 6m$, let

$$P_j^\delta : \boldsymbol{L}^2(\mathcal{M}) \to \operatorname{span}_{\mathbb{C}}\{\boldsymbol{u}_1^\delta, \ldots, \boldsymbol{u}_j^\delta\} \qquad \text{and} \qquad P_l : \boldsymbol{L}^2(\mathcal{M}) \to \operatorname{span}_{\mathbb{C}}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l\}$$

denote the orthogonal projections onto these subspaces, respectively. We can write these projections as line integrals of the resolvent of $G_\delta G_\delta^*$ and $\delta^6 T T^*$, respectively; see [26, III-(6.19)]. Then, a short calculation shows that

$$(10.2) \qquad P_l^\delta = P_l + \mathcal{O}(\delta), \qquad l = 1, \ldots, 6m,$$

as $\delta \to 0$ in $\mathcal{L}(\boldsymbol{L}^2(\mathcal{M}))$, provided that we make appropriate choices of eigenvectors $\boldsymbol{u}_l^\delta$ and $\boldsymbol{u}_l$, $l = 1, \ldots, 6m$.

In Proposition 9.3 we have seen that a point $\boldsymbol{z} \in \mathbb{R}^3_-$ coincides with one of the positions $\boldsymbol{z}_l$, $l = 1, \ldots, m$, if and only if $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} \in \mathcal{R}(T)$ or, equivalently, $(I - P_{6m})\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} = 0$. If we decompose the test function orthogonally as $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} = P_{6m}\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} + (I - P_{6m})\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$ and define the angle $\beta(\boldsymbol{z}) \in [0, \pi/2]$ by

$$\cot \beta(\boldsymbol{z}) := \frac{\|P_{6m}\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}\|_{\boldsymbol{L}^2(\mathcal{M})}}{\|(I - P_{6m})\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}\|_{\boldsymbol{L}^2(\mathcal{M})}},$$

then we have

$$\boldsymbol{z} \in \{\boldsymbol{z}_l \mid l = 1, \ldots, m\} \Longleftrightarrow \beta(\boldsymbol{z}) = 0 \Longleftrightarrow \cot \beta(\boldsymbol{z}) = \infty.$$

Unfortunately, we cannot compute $\beta(\boldsymbol{z})$, because P_{6m} corresponds to the leading order term T of the asymptotic expansion (9.2), but what we measure is the full measurement operator G_δ. However, in view of (10.2), for small values of δ the projected test function $P_{6m}\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$ is well approximated by $P_{6m}^\delta \boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$, and the projectors

P_p^δ can be computed for each $p \in \mathbb{N}$ by means of the singular value expansion of the measurement operator G_δ. Hence, for $p \in \mathbb{N}$, we define the angle $\beta_p^\delta(\boldsymbol{z}) \in [0, \pi/2]$ by

$$\cot \beta_p^\delta(\boldsymbol{z}) := \frac{\|P_p^\delta \boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}\|_{\boldsymbol{L}^2(\mathcal{M})}}{\|(I - P_p^\delta)\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}\|_{\boldsymbol{L}^2(\mathcal{M})}} = \left(\frac{\sum_{j \le p} |(\boldsymbol{u}_j^\delta, \boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}})_{\boldsymbol{L}^2(\mathcal{M})}|^2}{\sum_{j > p} |(\boldsymbol{u}_j^\delta, \boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}})_{\boldsymbol{L}^2(\mathcal{M})}|^2} \right)^{1/2} .$$

If we plot $\cot \beta_{6m}^\delta(\boldsymbol{z})$, we expect to see large values for points $\boldsymbol{z}$ which are close to the positions $\boldsymbol{z}_l$, $l = 1, \ldots, m$.

Because the number m of unknown scatterers is usually not known a priori, it has to be estimated somehow. Two different strategies are available: On the one hand, recalling (10.1), m may be estimated by looking for a gap in the set of singular values σ_l^δ of G_δ. This works if δ is small enough and the noise level is not too high. Otherwise it may give misleading results. On the other hand, we can plot $\cot \beta_p^\delta(\boldsymbol{z})$ for increasing values of p, until the number of reconstructed scatterers does not increase any more. This is reasonable, because for any subspace $U \subset \mathcal{R}(T)$ Proposition 9.3 reduces to

$$\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}} \in U \quad \Longrightarrow \quad \boldsymbol{z} \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}.$$

So, testing whether $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$ is contained in a subspace $U \subset \mathcal{R}(T)$, we can only expect to reconstruct a (possibly empty) subset of $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$. The number of reconstructed scatterers is monotonically increasing as $\dim(U)$ increases until all m scatterers are reconstructed for $\dim(U) = 6m$. Because none of the singular vectors of G_δ corresponding to singular values σ_j^δ, $j > 6m$, is expected to be exactly of the form $\boldsymbol{g}^{\boldsymbol{z},\boldsymbol{d}}$, $\boldsymbol{z} \notin \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$, the number of reconstructed scatterers should be constant for moderately sized $j > 6m$. Both strategies have been successfully tested in [8].

Finally, we show numerical results to illustrate the feasibility of the reconstruction method. We consider a two-layered background medium; the upper layer is empty ($\varepsilon_+ = \varepsilon_0 = 8.85 \cdot 10^{-12}$ Fm^{-1}, $\mu_+ = \mu_0 = 1.25 \cdot 10^{-6}$ Hm^{-1}), while the lower halfspace is filled with soil ($\varepsilon_- = \varepsilon_0(\varepsilon_r + \mathrm{i}\,\frac{\sigma}{\omega\varepsilon_0}) = 8.67 \cdot 10^{-11} + \mathrm{i}\, 5.95 \cdot 10^{-9}$ Fm^{-1}, $\mu_- = (1 + \chi)\mu_0 = 1.25 \cdot 10^{-6}$ Hm^{-1}, i.e., $\sigma = 7.5 \cdot 10^{-4}$ Sm^{-1}, $\chi = 1.9 \cdot 10^{-5}$, and $\varepsilon_r = 9.8$). The parameters for the lower halfspace are measurement data taken by Igel and Preetz [25] in the course of the project [23].

The measurement device operates on a square of size 50×50 cm^2 parallel to the surface of ground centered at $(0, 0, 10)$ cm. We simulate the measurement operator G_δ as done in [20]. For this purpose we impose magnetic dipoles with three linearly independent polarizations and a frequency of 20 kHz on a 6×6 equidistant grid on the measurement device. Then we approximate the corresponding scattered fields on the same grid using a boundary element method. The scatterers are two ellipsoids with semiaxes $(0.1, 0.2, 0.3)$ cm and $(2, 3, 1)$ cm buried at position $(-15, 15, -10)$ cm and $(15, -15, -40)$ cm, respectively. The simulated forward data contain an estimated numerical error of 4%. Additionally, we perturb the simulated scattered field by a uniformly distributed relative error of 3%.

The values of $\cot \beta_{12}^\delta(\boldsymbol{z})$ for $\boldsymbol{z} \in [-25, 25]^2 \times [-50, 0]$ cm^3 are used to visualize the location of the scatterers. The numerical implementation is essentially the same that has been used in [20] for a linear sampling method. Concerning implementation details, we refer the reader to this work; see also [8]. Figure 10.1 shows the first 20 singular values of the measurement operator G_δ and horizontal cross sections of $\cot \beta_{12}^\delta(\boldsymbol{z})$ for $z_3 = -10$ cm and $z_3 = -40$ cm, respectively. There is no distinct gap after the first 12 singurar values. One reason for this is the (numerical) error

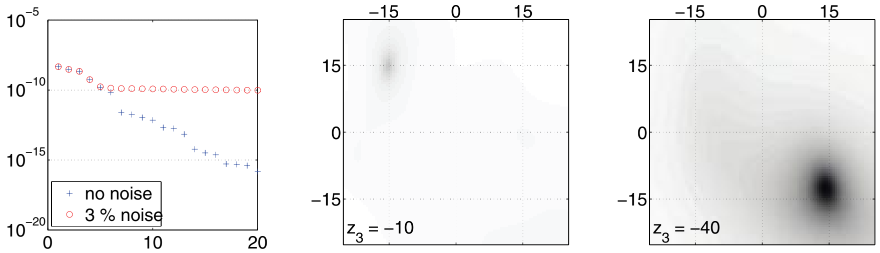


FIG. 10.1. *Singular values of G_δ and cross-sectional plots of* $\cot\beta_{12}^\delta(\boldsymbol{z})$ *for* $z_3 = -10$ *cm and* $z_3 = -40$ *cm with* 3% *noise.*

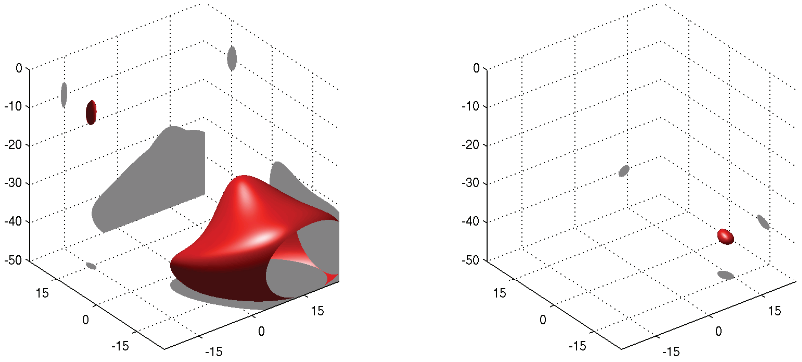


FIG. 10.2. *Isosurface plots* $\cot\beta_{12}^\delta(\boldsymbol{z}) = 25$ *and* $\cot\beta_{12}^\delta(\boldsymbol{z}) = 125$ *with* 3% *noise.*

in the forward data. Here the iterative procedure described above can be used to estimate the number of the scatterers. In the cross sectional plots the centers of the scatterers are clearly determined. Figure 10.2 shows isosurface plots $\cot\beta_{12}^\delta(\boldsymbol{z}) = 20$ and $\cot\beta_{12}^\delta(\boldsymbol{z}) = 200$. We emphasize that these visualizations should not be mistaken as reconstructions of the shape of the scatterers. These plots give just an idea of possible positions of buried scatterers; they can be expected to be inside the (red) surfaces. Our method does not allow a binary test for whether some point belongs to a scatterer or not. If we perturb the simulated forward data in this example with 5% equally distributed noise, the reconstructions of the positions of the scatterers get worse, but still two scatterers are reconstructed. For higher amounts of noise the method fails.

Note that this is only one particular numerical example that by no means covers all possible situations of interest. Comparing the method proposed here with the linear sampling method from [20], using (among others) the example above, we found that the linear sampling method is more sensitive to uncorrelated noise. Using the unperturbed simulated data, the position of the scatterers has been reconstructed by the linear sampling method. But with 3% noise in the data the linear sampling method failed. The MUSIC-type reconstruction method studied in [24] gives numerical results comparable to the results presented here, although we mention that much higher frequencies have been used in [24]. In their final implementation both methods are quite similar. Our analysis from sections 9 and 10 is meant to be an extension of [24] and a rigorous justification of both methods.

11. Conclusions. We have considered an inverse scattering problem for small scatterers in a two-layered background medium which originated in the project [23] on humanitarian demining. An asymptotic expansion of the near field measurement operator as the size of the scatterers tends to zero has been proven. We used the asymptotic formula to justify a noniterative reconstruction method that can be interpreted as an asymptotic version of a factorization method, or as a MUSIC-type method. First numerical results indicate that this method may be appropriate to detect small buried objects from sufficiently accurate measurements of the scattered field above the surface of ground. Although our results have been derived for an idealized setting, we expect that the asymptotic expansion as well as the reconstruction method can be applied to more realistic models for measurement devices used for humanitarian demining, including special coil geometries such as, e.g., the double D design considered in [17]. We intend to address this in the future.

Acknowledgments. We would like to thank Christoph Schneider for providing us with his code for the linear sampling method from [20], which we adapted to implement the reconstruction method presented here. The boundary element method we used for generating simulated measurement data for multiple scatterers is an extension of a boundary element code for single scatterers developed by Roland Potthast and his group at the University of Göttingen; see also [17]. Last but not least, the author thanks Habib Ammari, Bastian Gebauer, Martin Hanke, Andreas Kirsch, and Christoph Schneider for useful discussions and valuable hints.

REFERENCES

[1] H. Ammari, R. Griesmaier, and M. Hanke, *Identification of small inhomogeneities: Asymptotic factorization*, Math. Comput., 76 (2007), pp. 1425–1448.

[2] H. Ammari, E. Iakovleva, D. Lesselier, and G. Perrusson, *MUSIC-type electromagnetic imaging of a collection of small three-dimensional bounded inclusions*, SIAM J. Sci. Comput., 29 (2007), pp. 674–709.

[3] H. Ammari and H. Kang, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[4] H. Ammari and A. Khelifi, *Electromagnetic scattering by small dielectric inhomogeneities*, J. Math. Pures Appl., 82 (2003), pp. 749–842.

[5] H. Ammari, S. Moskow, and M. S. Vogelius, *Boundary integral formulae for the reconstruction of electric and electromagnetic inhomogeneities of small volume*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 49–66.

[6] H. Ammari, M. S. Vogelius, and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter* II. *The full Maxwell equations*, J. Math. Pures Appl., 80 (2001), pp. 769–814.

[7] H. Ammari and D. Volkov, *The leading order term in the asymptotic expansion of the scattering amplitude of a collection of finite number of dielectric inhomogeneities of small diameter*, Multiscale Computational Engineering, 3 (2005), pp. 149–160.

[8] M. Brühl, M. Hanke, and M. S. Vogelius, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[9] A. Buffa, M. Costabel, and D. Sheen, *On traces of $H(\mathrm{curl}, \Omega)$ in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.

[10] F. Cakoni and D. Colton, *Qualitative methods in inverse scattering theory. An introduction*, Interaction of Mechanics and Mathematics, Springer-Verlag, Berlin, 2006.

[11] F. Cakoni, M'B. Fares, and H. Haddar, *Analysis of two linear sampling methods applied to electromagnetic imaging of buried objects*, Inverse Problems, 22 (2006), pp. 845–867.

[12] M. Cheney, *The linear sampling method and the MUSIC algorithm*, Inverse Problems, 17 (2001), pp. 591–595.

[13] W. C. Chew, *Waves and Fields in Inhomogeneous Media*, Van Nostrand Reinhold, New York, 1990.

[14] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley & Sons, New York, 1983.

[15] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[16] P.-M. Cutzach and C. Hazard, *Existence, uniqueness and analyticity properties for electromagnetic scattering in a two-layered medium*, Math. Methods Appl. Sci., 21 (1998), pp. 433–461.

[17] F. Delbary, K. Erhard, R. Kress, R. Potthast, and J. Schulz, *Inverse electromagnetic scattering in a two-layered medium with an application to mine detection*, Inverse Problems, 24 (2008), 015002.

[18] A. J. Devaney, *Super-resolution processing of multi-static data using time reversal and MUSIC*, preprint, Department of Electrical Engineering, Northeastern University, Boston, MA, 1999.

[19] A. Friedman and M. S. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[20] B. Gebauer, M. Hanke, A. Kirsch, W. Muniz, and C. Schneider, *A sampling method for detecting buried objects using electromagnetic scattering*, Inverse Problems, 21 (2005), pp. 2035–2050.

[21] B. Gebauer, M. Hanke, and C. Schneider, *Sampling methods for low-frequency electromagnetic imaging*, Inverse Problems, 24 (2008), 015007.

[22] D. Guelle, A. Smith, A. Lewis, and T. Bloodworth, *EUR* 20837 *Metal Detector Handbook for Humanitarian Demining*, Office for Official Publications of the European Communities, Luxembourg, 2003.

[23] *HuMin/MD—Metal Detectors for Humanitarian Demining—Development Potentials in Data Analysis Methodology and Measurement,* Project Network, available online at http://www.humin-md.de/.

[24] E. Iakovleva, S. Gdoura, D. Lesselier, and G. Perrusson, *Multi-static response matrix of a 3-D inclusion in half space and MUSIC imaging*, IEEE Trans. Antennas Propagat., 55 (2007), pp. 2598–2609.

[25] J. Igel and H. Preetz, *Elektromagnetische Bodenparameter und ihre Abhängigkeit von den Bodeneigenschaften.—Zwischenbericht Projektverbund Humanitäres Minenräumen*, Technical report, Leibniz Institute of Applied Geosciences, Hannover, Germany, 2005.

[26] T. Kato, *Perturbation Theory for Linear Operators*, Grundlehren Math. Wiss. 132, Springer-Verlag, Berlin, 1966.

[27] A. Kirsch, *Surface gradients and continuity properties for some integral operators in classical scattering theory*, Math. Methods Appl. Sci., 11 (1989), pp. 789–804.

[28] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[29] A. Kirsch, *An integral equation for Maxwell's equations in a layered medium with an application to the factorization method*, J. Integral Equations Appl., 19 (2007), pp. 333–359.

[30] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[31] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, UK, 2003.

[32] J.-C. Nédélec, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Appl. Math. Sci. 144, Springer-Verlag, New York, 2001.

[33] M. Petry, *Über die Streuung Zeitharmonischer Wellen im Geschichteten Raum*, Ph.D. thesis, Georg-August-Universität zu Göttingen, Göttingen, Germany, 1993.

[34] A. Sommerfeld, *Partial Differential Equations in Physics*, Academic Press, New York, 1949.

[35] C. Weber, *Regularity theorems for Maxwell's equations*, Math. Methods Appl. Sci., 3 (1981), pp. 523–536.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1404–1422

ANALYTICAL AND NUMERICAL SOLUTIONS FOR TORSIONAL FLOW BETWEEN COAXIAL DISCS WITH HEAT TRANSFER*

DAVID O. OLAGUNJU[†], ANAND B. VYAS[†], AND SHANGYOU ZHANG[†]

Abstract. We consider nonisothermal torsional flow between two coaxial parallel plates with heat transfer at the upper rotating plate, constant temperature on the lower stationary plate, and no heat loss at the fluid-air interface. Viscous heating is modelled by a Nahme law with exponential dependence on temperature. Due to the highly nonlinear nature of the governing equations an exact solution is not feasible. Therefore we solve the problem using both numerical and perturbation methods. Specifically, analytical solutions are obtained using asymptotic expansions based on the aspect ratio and the Nahme–Griffith number, a measure of viscous heating, as perturbation parameters. The numerical solutions are obtained by a finite element method. Good agreement is found between the analytical and numerical solutions in the appropriate parameter range. In viscometric applications the torque exerted by the fluid on the lower plate is an important quantity. For isothermal flow the dimensionless torque can easily be calculated. In this paper we obtain an analytical formula that can be used to calculate nonisothermal correction to the torque.

Key words. parallel-plate flow, viscous heating, axisymmetric finite elements

AMS subject classifications. 76D07, 76M10, 65F10

DOI. 10.1137/070686871

1. Introduction.

The problem of viscous heating in viscometric flow of Newtonian and non-Newtonian liquids is of both theoretical and practical interest. An exact solution was obtained for flow of a Newtonian fluid between two infinite parallel plates by Nahme in 1940 [1]. The viscosity was modelled by an exponential function of temperature. A similar result was obtained by Kearsley [2] for the pressure gradient flow of a Newtonian fluid in a tube. Bird and Turian [3] analyzed the viscous heating problem for the flow of a Newtonian fluid between a cone and a plate. The equations for the temperature and velocity were uncoupled by assuming an isothermal velocity profile. The variational form for the temperature was then solved numerically. Isothermal boundary conditions were imposed at the physical boundaries and zero heat loss imposed at the air-liquid interface. Their analysis showed that viscous heating can lead to observable errors in the cone-and-plate viscometer. In a subsequent paper, Turian and Bird [4] extended the theoretical investigation to plane Couette flow of Newtonian fluids with temperature dependent viscosity and thermal conductivity. The thermal conductivity was assumed to be a linear function of temperature, while the viscosity was assumed to obey a Nahme law with exponential dependence on temperature. A regular perturbation solution in powers of the Brinkman number was obtained for the velocity and temperature. A perturbation solution in powers of the Brinkman number were later obtained for non-Newtonian liquids described by the power-law and Ellis models [5]. The boundary conditions were the same as those used in [4]. Exact analytical solutions have also been obtained by Martin [6] for flows between infinite concentric cylinders and infinite parallel plates for Newtonian and power-law fluids. Two types of boundary conditions were considered, one in which

*Received by the editors March 30, 2007; accepted for publication (in revised form) November 16, 2007; published electronically May 2, 2008.
http://www.siam.org/journals/siap/68-5/68687.html

[†]Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (olagunju@math.udel.edu, abvyas@math.udel.edu, szhang@udel.edu).

both surfaces were isothermal and the other in which one was isothermal and the other adiabatic. Closed-form solutions similar to those of Nahme and Kearsley were obtained for isothermal and adiabatic boundary conditions by Gavis and Laurence [7].

These early analytical studies were motivated by the need to quantify the deviation form isothermal flow in viscometers when viscous dissipation is significant and to provide simple formulas to correct for such deviations in quantities such as the torque on the stationary plate in parallel-plate flow. In recent years renewed interest in this matter has been spurred by experiments in elastic instabilities of viscoelastic fluids. It has been observed that viscous heating could lead to qualitatively and quantitatively significant deviation from isothermal theory in the stability property of a viscoelastic fluid. Experimental study of the stability of isothermal flow of viscoelastic torsional flow was first reported by Magda and Larson [8] and McKinley et al. [9]. Linear stability analysis was first carried out by Oztekin and Brown [10] for flow between parallel-plate flow in which the boundary condition at the fluid-air interface was neglected. Linear stability results for flow in a finite domain incorporating boundary conditions at the free surface have also been considered [11, 12, 13].

Experiments on the effect of viscous heating on the stability of torsional flow of viscoelastic fluids was first reported by Rothstein and McKinley [14]. The results were found to be remarkably different from those of isothermal flows. It was shown that viscous heating tended to stabilize the flow. A linear stability for the nonisothermal problem was later analyzed by Olagunju, Cook, and McKinley [15] which agrees qualitatively with their experimental results. In that paper isothermal boundary conditions on the plates were used, while those at the free surface were neglected. However, as noted by Arigo [16], it is practically impossible to control the temperature on the upper rotating cone (or plate in the case of the parallel-plate torsional viscometer). He also notes that the upper rotating cone (or plate) is cooled by convection of ambient air at 23–24 degrees Celsius. A more realistic set of boundary conditions is to treat the bottom plate as isothermal, the free surface as an insulated boundary, and the top plate as a thermal mass. The insulation boundary condition surface can be justified on the grounds that for a small gap thickness, the surface available for heat transfer to the ambience through the radial interface is practically negligible. It is hoped that this will give results that are in quantitative agreement with experiments. A heat transfer boundary condition was used for the viscoelastic Taylor–Couette problem by Al-Mubaiyedh, Suresh Kumar, and Khomani [17]. In their study, the heat transfer boundary condition was used to numerically simulate the experiments of Baumert and Muller [18, 19]. Another assumption that was made in [15] is that the parallel plates are infinite in extent. In order to obtain better agreement between theory and experiments we think that it is necessary to relax these assumptions. As a first step in this direction we study the effect of the finite geometry and the more realistic boundary conditions on the base flow. In [20], Olagunju showed that for torsional flow of a viscoelastic fluid the base flow is not always purely circumferential. He showed that viscous heating leads to secondary flows with recirculating roll cells in the base solution.

In this paper, we obtain perturbation and numerical solutions for the flow between two parallel plates of a Newtonian fluid with temperature dependent viscosity. Specifically we will assume an exponential dependence of Nahme type. As noted above this problem has been solved for flow between two infinite parallel plates [1, 4]. In this limit the problem reduces to two coupled ordinary differential equations for the temperature and azimuthal velocity. We propose solving the problem in a finite geometry with a fluid-air interface. In this case we obtain two coupled partial differential equations for the temperature and velocity. Bird and Turian (see [5]) have also analyzed the problem in a finite geometry between a cone and a plate. However, they assumed that the velocity profile was isothermal, thereby reducing the problem to a single partial differential equation for the temperature. In all previous work that we know the boundary conditions are either isothermal on both plates or isothermal on one plate and zero heat transfer on the other. We will adopt the more realistic boundary conditions described above, namely the isothermal condition on the stationary plate, heat transfer on the upper plate, and zero heat transfer at the fluid-air interface. To the best of our knowledge exact analytical solutions for this problem have not been previously reported. Having an analytical solution for this problem will enable one to estimate errors in the torque calculations due to heat transfer and edge effects if needed. This is important in viscometry. Analytical solutions can also be used to validate numerical calculations. For viscoelastic flows in which secondary flow in the base flow is weak or nonexistent the solution provided here provides an accurate approximation to the base flow needed in any linear stability analysis.

2. Governing equations.

We consider the flow of a fluid in the region between two coaxial parallel plates of radius a and separation h in which the top plate rotates at a constant angular speed ω and the bottom plate is stationary. Following Olagunju [21], the nondimensionalized equations governing the primary flow for the azimuthal velocity W and a scaled temperature Θ are given in cylindrical coordinates as

$$\frac{\partial^2 W}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 W}{\partial r^2} + \frac{1}{r}\frac{\partial W}{\partial r} - \frac{W}{r^2} \right) = \frac{\partial \Theta}{\partial z}\frac{\partial W}{\partial z} + \alpha^2 \frac{\partial \Theta}{\partial r}\left(\frac{\partial W}{\partial r} - \frac{W}{r} \right), \tag{1}$$

$$\frac{\partial^2 \Theta}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 \Theta}{\partial r^2} + \frac{1}{r}\frac{\partial \Theta}{\partial r} \right) = -\mathrm{Na}_0 \, e^{-\Theta} \left[\left(\frac{\partial W}{\partial z} \right)^2 + \alpha^2 \left(\frac{\partial W}{\partial r} - \frac{W}{r} \right)^2 \right] \tag{2}$$

with boundary conditions

$$\text{at } z = 0, \qquad W = 0, \qquad \Theta = \vartheta_w, \tag{3}$$

$$\text{at } z = 1, \qquad W = r, \qquad \frac{\partial \Theta}{\partial z} + B\,\Theta = B\vartheta_a, \tag{4}$$

$$\text{at } r = 0, \qquad W = 0, \qquad |\Theta| < \infty, \tag{5}$$

$$\text{at } r = 1, \qquad \frac{\partial W}{\partial r} - \frac{W}{r} = 0. \qquad \frac{\partial \Theta}{\partial r} = 0. \tag{6}$$

Here ϑ_w and ϑ_a are the scaled temperature at the stationary plate and the ambient. The aspect ratio $\alpha = h/a$ and the modified Biot number B are defined in section A.1 of the appendix. The Nahme–Griffith number Na_0, which is a measure of viscous heating in the fluid, is zero for isothermal flows. It is defined as $\mathrm{Na}_0 \equiv (\eta_0 \delta a^2 \omega^2)/(kT_0)$. The quantity η_0 is the isothermal viscosity, δ a thermal sensitivity parameter, k the thermal conductivity, and T_0 a reference temperature [14]. Since the equations are nonlinear finding an exact analytical solution valid for all parameter values is impractical. Therefore we will solve the equations numerically using a finite element method. We will also obtain analytical solutions using perturbation expansions in Nahme–Griffith number Na and aspect ratio α.

3. Analytical solutions.

3.1. An exact solution for $\alpha = 0$ and $B = 0$. An exact solution of equations (1)–(6) can be found for $\alpha = 0$, $B = 0$, and all values of Na_0. This corresponds to plane Couette flow with the upper plate insulated. This solution does not satisfy the boundary conditions at the fluid-air interface $r = 0$. However, we will show that it provides an excellent approximation to the solution for small α except very close to $r = 1$. In addition we will show that the torque exerted by the fluid on the lower plate is very well approximated by this exact solution when the aspect ratio α is small. An analytical formula is provided which can be used to calculate nonisothermal correction to the torque, as is often required in rheometry. To the best of our knowledge, this analytical representation for the torque has not been previously reported. Note that the exact solution obtained by Nahme [1] corresponds to $\alpha = 0$ and $B = \infty$.

Setting α and B to zero, (1)–(6) reduce to the following:

$$\frac{d^2W}{dz^2} = \frac{d\Theta}{dz}\frac{dW}{dz}, \tag{7}$$

$$\frac{d^2\Theta}{dz^2} = -\mathrm{Na}_0 e^{-\Theta}\left(\frac{dw}{dz}\right)^2. \tag{8}$$

The boundary conditions are

$$W = 0, \quad \Theta = \vartheta_w \quad \text{for} \quad z = 0 \tag{9}$$

and

$$W = r, \quad \frac{\partial\Theta}{\partial z} = 0 \quad \text{for} \quad z = 1. \tag{10}$$

It is straightforward to obtain the solution which is given by

$$W = \frac{r}{2} - \frac{1}{\mu\mathrm{Na}}\tanh[E(1-2z)], \tag{11}$$

$$\Theta = \vartheta_w + \ln\left[\left(1 + \frac{r^2\mathrm{Na}}{8}\right)\operatorname{sech}^2[E(1-2z)]\right], \tag{12}$$

where $\mathrm{Na} = \mathrm{Na}_0 e^{-\vartheta_w}$,

$$E = \tanh^{-1}(r\mu\mathrm{Na}/2),$$

and

$$\mu = \left[2\mathrm{Na}\left(1 + \frac{r^2}{8}\mathrm{Na}\right)\right]^{-\frac{1}{2}}.$$

The dimensionless torque on the lower plate is defined as

$$\mathcal{T} = \int_0^1\int_0^1\left(\frac{dW}{dz}\right)_{z=0} r^2 dr dz.$$

Using the solution obtained above, a series solution for $\mathcal{T}$ valid for $\mathrm{Na} < 2$ is given by

$$\mathcal{T} = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}(-1)^k\left(\frac{\mathrm{Na}}{2}\right)^{n+k}\frac{(n+k)!}{n!k!}\frac{1}{(2n+1)(2n+2k+4)}. \tag{13}$$

For other values of Na the integral can easily be computed numerically. These solutions will be compared to asymptotic and numerical solutions below.

3.2. Asymptotic solution for Na $\ll 1$: $\vartheta_w = \vartheta_a$. We seek a regular expansion in Nahme number for W and Θ as follows:

$$W = W_0 + \text{Na}\, W_1 + O(\text{Na}^2), \qquad \Theta = \Theta_0 + \text{Na}\, \Theta_1 + O(\text{Na}^2). \tag{14}$$

Here and in what follows $\text{Na} = \text{Na}_0 e^{-\vartheta_w}$. The governing equations for W_0 and Θ_0 are

$$\frac{\partial^2 W_0}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 W_0}{\partial r^2} + \frac{1}{r}\frac{\partial W_0}{\partial r} - \frac{W_0}{r^2} \right) = \frac{\partial \Theta_0}{\partial z}\frac{\partial W_0}{\partial z} + \alpha^2 \frac{\partial \Theta_0}{\partial r}\left(\frac{\partial W_0}{\partial r} - \frac{W_0}{r} \right), \tag{15}$$

$$\frac{\partial^2 \Theta_0}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 \Theta_0}{\partial r^2} + \frac{1}{r}\frac{\partial \Theta_0}{\partial r} \right) = 0 \tag{16}$$

with the boundary conditions

$$\text{at } z = 0, \quad W_0 = 0, \quad \Theta_0 = \vartheta_w, \tag{17}$$

$$\text{at } z = 1, \quad W_0 = r, \quad \frac{\partial \Theta_0}{\partial z} + B\, \Theta_0 = B\, \vartheta_w, \tag{18}$$

$$\text{at } r = 0, \quad W_0 = 0, \quad |\Theta_0| < \infty, \tag{19}$$

$$\text{at } r = 1, \quad \frac{\partial W_0}{\partial r} - \frac{W_0}{r} = 0, \quad \frac{\partial \Theta_0}{\partial r} = 0. \tag{20}$$

The leading order solution for $\text{Na} = 0$ gives the isothermal solution

$$W_0 = rz, \tag{21}$$

$$\Theta_0 = \vartheta_w. \tag{22}$$

Note that this solution is valid for all values of α.

The solution at order Na corresponding the first nonisothermal correction satisfies the following equations:

$$\frac{\partial^2 W_1}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 W_1}{\partial r^2} + \frac{1}{r}\frac{\partial W_1}{\partial r} - \frac{W_1}{r^2} \right) = r\frac{\partial \Theta_1}{\partial z}, \tag{23}$$

$$\frac{\partial^2 \Theta_1}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 \Theta_1}{\partial r^2} + \frac{1}{r}\frac{\partial \Theta_1}{\partial r} \right) = -r^2 \tag{24}$$

with boundary conditions

$$\text{at } z = 0, \quad W_1 = 0, \quad \Theta_1 = 0, \tag{25}$$

$$\text{at } z = 1, \quad W_1 = 0, \quad \frac{\partial \Theta_1}{\partial z} + B\Theta_1 = 0, \tag{26}$$

$$\text{at } r = 0, \quad W_1 = 0, \quad |\Theta_1| < \infty, \tag{27}$$

$$\text{at } r = 1, \quad \frac{\partial W_1}{\partial r} - \frac{W_1}{r} = 0, \quad \frac{\partial \Theta_1}{\partial r} = 0. \tag{28}$$

Note that (23)–(24) are now uncoupled. The equations can be solved exactly by separation of variables as follows:

$$\Theta_1 = -\sum_{n=1}^{\infty} \frac{\bar{\Gamma}_n}{\lambda_n^2} \left[\frac{2\alpha^3}{\lambda_n} \frac{I_0\left(\frac{\lambda_n r}{\alpha}\right)}{I_1\left(\frac{\lambda_n}{\alpha}\right)} - \frac{4\alpha^4}{\lambda_n^2} - \alpha^2 r^2 \right] \sin(\lambda_n z) \tag{29}$$

and

$$W_1 = \sum_{m=1}^{\infty} F_m(r) \sin(m\pi z), \tag{30}$$

where $\lambda_n, n = 1, 2, \dots,$ are positive solutions of the transcendental equation

$$\tan(\lambda_n) + \frac{\lambda_n}{B} = 0, \tag{31}$$

and $\Gamma_n = -\frac{1}{\alpha^2} \frac{4(1-\cos(\lambda_n))}{2\lambda_n - \sin(2\lambda_n)}$. This equation has infinitely many positive roots. Here I_n is the modified Bessel function of the first kind. This solution is also valid for all values of the aspect ratio α.

The expression for $F_m(r)$ involves complicated integrals of Bessel functions (see the appendix for details).

3.3. Asymptotic solution for Na $\ll$ 1 : $\boldsymbol{\vartheta_w \neq \vartheta_a}$. The governing equations for W_0 and Θ_0, W_1 and Θ_1 are the same as in the previous section.

At zeroth order in Na we have the isothermal solution

$$\Theta_0 = \chi z + \vartheta_w, \quad \text{where} \quad \chi \equiv \frac{B(\vartheta_a - \vartheta_w)}{1 + B}, \tag{32}$$

$$W_0 = \left(\frac{e^{\chi z} - 1}{e^{\chi} - 1} \right) r. \tag{33}$$

The equations at order Na are

$$\frac{\partial^2 W_1}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 W_1}{\partial r^2} + \frac{1}{r}\frac{\partial W_1}{\partial r} - \frac{W_1}{r^2} \right) = \frac{\chi \ r \ e^{\chi z}}{e^{\chi} - 1} \frac{\partial \Theta_1}{\partial z} + \chi \frac{\partial W_1}{\partial z}, \tag{34}$$

$$\frac{\partial^2 \Theta_1}{\partial z^2} + \alpha^2 \left(\frac{\partial^2 \Theta_1}{\partial r^2} + \frac{1}{r}\frac{\partial \Theta_1}{\partial r} \right) = -\frac{\chi^2 r^2 e^{\chi z}}{\left(e^{\chi} - 1\right)^2} \tag{35}$$

with boundary conditions

$$\text{on } z = 0, \qquad W_1 = 0, \qquad \Theta_1 = 0, \tag{36}$$

$$\text{on } z = 1, \qquad W_1 = 0, \qquad \frac{\partial \Theta_1}{\partial z} + B\Theta_1 = 0, \tag{37}$$

$$\text{on } r = 0, \qquad W_1 = 0, \qquad |\Theta_1| < \infty, \tag{38}$$

$$\text{on } r = 1, \qquad \frac{\partial W_1}{\partial r} - \frac{W_1}{r} = 0, \qquad \frac{\partial \Theta_1}{\partial r} = 0. \tag{39}$$

Although these equations can also be solved exactly the solutions are rather too complicated. Therefore we will obtain the solution as an expansion in α. This limit has applications in rheometric devices where α is typically less than 0.1. For the case $\alpha = O(1)$ the solution will be computed numerically. Because the limit $\alpha \to 0$ is singular we use the method of matched asymptotic expansion. Thus we seek an outer solution and an inner solution and then obtain a composite expansion for the solution by matching.

Outer solution. For the outer solution we expand as follows:

$$W_1 = W_{10}^o + \alpha W_{11}^o + O(\alpha^2), \qquad \Theta_1 = \Theta_{10}^o + \alpha \Theta_{11}^o + O(\alpha^2), \tag{40}$$

where the superscript (o) refers to the outer solution.

The governing equations at zeroth order in α are

$$\frac{\partial^2 W_{10}^o}{\partial z^2} = \frac{\chi\ r\ e^{\chi z}}{e^{\chi} - 1} \frac{\partial \Theta_{10}^o}{\partial z} + \chi \frac{\partial W_{10}^o}{\partial z}, \tag{41}$$

$$\frac{\partial^2 \Theta_{10}^o}{\partial z^2} = -\frac{\chi^2 r^2 e^{\chi z}}{(e^{\chi} - 1)^2} \tag{42}$$

with the corresponding boundary conditions

$$\text{at } z = 0, \quad W_{10}^o = 0, \quad \Theta_{10}^o = 0, \tag{43}$$

$$\text{at } z = 1, \quad W_{10}^o = 0, \quad \frac{\partial \Theta_{10}^o}{\partial z} + B\Theta_{10} = 0. \tag{44}$$

The solution satisfying the above governing equations and the boundary conditions are

$$\Theta_{10}^o = \frac{r^2}{(e^{\chi} - 1)^2} \left[1 - e^{\chi z} + z \frac{(\chi + B)e^{\chi} - B}{1 + B} \right], \tag{45}$$

$$\begin{aligned} W_{10}^o = {} & \frac{\chi r^3}{(e^{\chi} - 1)^3} \left[-\frac{e^{2\chi z}}{2\chi} + ((\chi + B)e^{\chi} - B) \frac{e^{\chi z}(\chi z - 1)}{(1 + B)\chi^2} \right] \\ & - \frac{\chi e^{\chi z} r^3}{(e^{\chi} - 1)^4} \left[\frac{1 - e^{2\chi}}{2\chi} + ((\chi + B)e^{\chi} - B) \frac{1 + (\chi - 1)e^{\chi}}{(1 + B)\chi^2} \right] \\ & - \frac{r^3}{(e^{\chi} - 1)^4} \left[\frac{e^{2\chi} - e^{\chi}}{2} - \frac{e^{\chi} \left((\chi + B)e^{\chi} - B \right)}{1 + B} \right]. \end{aligned} \tag{46}$$

Further, it is also determined that

$$W_{11}^o = 0, \qquad \Theta_{11}^o = 0. \tag{47}$$

Inner solution. For the inner expansion we introduce the stretched variable, $\xi \equiv \frac{1-r}{\alpha}$, and seek an expansion of the form

$$W_1^i = W_{10}^i + \alpha W_{11}^i + O(\alpha^2), \quad \Theta_1^i = \Theta_{10}^i + \alpha \Theta_{11}^i + O(\alpha^2). \tag{48}$$

The governing equations and the boundary conditions at zeroth order in α are

$$\frac{\partial^2 W_{10}^i}{\partial z^2} + \frac{\partial^2 W_{10}^i}{\partial \xi^2} - \chi \frac{\partial W_{10}^i}{\partial z} = \frac{\chi e^{\chi z}}{(e^\chi - 1)} \left(\frac{\partial \Theta_{10}^i}{\partial z} \right), \tag{49}$$

$$\frac{\partial^2 \Theta_{10}^i}{\partial z^2} + \frac{\partial^2 \Theta_{10}^i}{\partial \xi^2} = -\frac{\chi^2 e^{\chi z}}{(e^\chi - 1)^2}, \tag{50}$$

$$\text{at } z = 0, \qquad W_{10}^i = 0, \qquad \Theta_{10}^i = 0, \tag{51}$$

$$\text{at } z = 1, \qquad W_{10}^i = 0, \qquad \frac{\partial \Theta_{10}^i}{\partial z} + B\Theta_{10}^i = 0, \tag{52}$$

$$\text{at } \xi = 0, \qquad \frac{\partial W_{10}^i}{\partial \xi} = 0, \qquad \frac{\partial \Theta_{10}^i}{\partial \xi} = 0. \tag{53}$$

It is straightforward to obtain the following expressions:

$$\Theta_{10}^i = \frac{1}{(e^\chi - 1)^2} \left[1 - e^{\chi z} + z \frac{(\chi + B)e^\chi - B}{1 + B} \right], \tag{54}$$

$$\begin{aligned} W_{10}^i = {} & \frac{\chi}{(e^\chi - 1)^3} \left[-\frac{e^{2\chi z}}{2\chi} + ((\chi + B)e^\chi - B) \frac{e^{\chi z}(\chi z - 1)}{(1 + B)\chi^2} \right] \\ & - \frac{\chi e^{\chi z}}{(e^\chi - 1)^4} \left[\frac{1 - e^{2\chi}}{2\chi} + ((\chi + B)e^\chi - B) \frac{1 + (\chi - 1)e^\chi}{(1 + B)\chi^2} \right] \\ & - \frac{1}{(e^\chi - 1)^4} \left[\frac{e^{2\chi} - e^\chi}{2} - \frac{e^\chi \left((\chi + B)e^\chi - B \right)}{1 + B} \right]. \end{aligned} \tag{55}$$

The order α equations are

$$\frac{\partial^2 \Theta_{11}^i}{\partial z^2} + \frac{\partial^2 \Theta_{11}^i}{\partial \xi^2} = \frac{2\xi \chi^2 e^{\chi z}}{(e^\chi - 1)^2}, \tag{56}$$

$$\frac{\partial^2 W_{11}^i}{\partial z^2} + \frac{\partial^2 W_{11}^i}{\partial \xi^2} - \chi \frac{\partial W_{11}^i}{\partial z} = \frac{\partial W_{01}^i}{\partial z} \left(\frac{\partial \Theta_{10}^i}{\partial z} \right) + \frac{\partial W_{00}^i}{\partial z} \left(\frac{\partial \Theta_{11}^i}{\partial z} \right). \tag{57}$$

The boundary conditions are the same as above. The solution of the equations is

$$\Theta_{11}^i = -2\xi \Theta_{10}^i - 2 \sum_{n=1}^{\infty} \frac{\tilde{\Gamma}_n e^{-\frac{\lambda_n (1-r)}{\alpha}}}{\lambda_n^3} \sin(\lambda_n z), \tag{58}$$

$$\begin{aligned} W_{11}^i = {} & -3\xi W_{10}^i + \sum_{m=1}^{\infty} \frac{2\tilde{A}_m e^{-\Lambda_m \xi} e^{\frac{\chi z}{2}}}{\Lambda_m^2} \sin(m\pi z) \\ & + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{\tilde{\Gamma}_n \tilde{B}_{mn}}{\lambda_n^2 (\Lambda_m^2 - \lambda_n^2)} \left(e^{-\lambda_n \xi} - \frac{\lambda_n e^{-\Lambda_m \xi}}{\Lambda_m} \right) e^{\frac{\chi z}{2}} \sin(m\pi z), \end{aligned} \tag{59}$$

where

$$\tilde{\Gamma}_n = \frac{\chi^2 \int_0^1 e^{\chi z} \sin(\lambda_n z) \mathrm{d}z}{(e^{\chi} - 1)^2 \int_0^1 \sin^2(\lambda_n z) \mathrm{d}z} \tag{60}$$

and

$$\Lambda_m^2 = \frac{\chi^2}{4} + m^2\pi^2,$$

$$\tilde{A}_m = \frac{\chi}{(e^{\chi}-1)^3} \frac{\int_0^1 \left(-\chi e^{\chi z} + \frac{(\chi+B)e^{\chi}-B}{1+B}\right) e^{\frac{\chi z}{2}} \sin(m\pi z)\mathrm{d}z}{\int_0^1 \sin^2(m\pi z)\mathrm{d}z},$$

$$\tilde{B}_{mn} = \frac{2\chi}{e^{\chi}-1} \frac{\int_0^1 e^{\frac{\chi z}{2}} \cos(\lambda_n z) \sin(m\pi z)\mathrm{d}z}{\int_0^1 \sin^2(m\pi z)\mathrm{d}z}. \tag{61}$$

Composite solution. In order to use the Van Dyke matching principle, the outer solution for the velocity and the temperature distribution is expressed in terms of the inner variable including terms of the first order in α,

$$(\Theta^o)^i = \chi z + \vartheta_w + \mathrm{Na}(1 - 2\alpha\xi)\Theta_{10}^i + O(\mathrm{Na}^2), \tag{62}$$

$$(W^o)^i = \left(\frac{e^{\chi z}-1}{e^{\chi}-1}\right) r + \mathrm{Na}(1 - 3\alpha\xi)W_{10}^i + O(\mathrm{Na}^2). \tag{63}$$

The Van Dyke matching principle is expressed using the formulas

$$W^c = W^o + W^i - (W^o)^i, \qquad \Theta^c = \Theta^o + \Theta^i - (\Theta^o)^i. \tag{64}$$

The composite solution for temperature and velocity distribution is then given by

$$\begin{aligned}\Theta^c = \chi z + \vartheta_w + \mathrm{Na}\left(\frac{r^2}{(e^{\chi}-1)^2}\left[1 - e^{\chi z} + z\frac{(\chi+B)e^{\chi}-B}{1+B}\right]\right)\\ -2\mathrm{Na}\alpha \sum_{n=1}^{\infty} \frac{\tilde{\Gamma}_n e^{-\frac{\lambda_n(1-r)}{\alpha}}}{\lambda_n^3} \sin(\lambda_n z) + O(\mathrm{Na}^2),\end{aligned} \tag{65}$$

$$\begin{aligned}W^c = \left(\frac{e^{\chi z}-1}{e^{\chi}-1}\right) r + \mathrm{Na}W_1^o + \mathrm{Na}\alpha \sum_{m=1}^{\infty} \frac{2\tilde{A}_m e^{-\Lambda_m \frac{(1-r)}{\alpha}} e^{\frac{\chi z}{2}}}{\Lambda_m^2} \sin(m\pi z)\\ +\mathrm{Na}\alpha \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \frac{\tilde{\Gamma}_n \tilde{B}_{mn}}{\lambda_n^2(\Lambda_m^2 - \lambda_n^2)} \left(e^{-\lambda_n \frac{(1-r)}{\alpha}} - \frac{\lambda_n e^{-\Lambda_m \frac{(1-r)}{\alpha}}}{\Lambda_m}\right) e^{\frac{\chi z}{2}} \sin m\pi z.\end{aligned} \tag{66}$$

4. Numerical solution. The domain Ω for numerical computation is $0 < z < 1$ and $0 < r < 1$, shown in Figure 1. In order to apply the finite element method, we need to rewrite the two partial differential equations in variational forms. We multiply the continuity equation (1) by $rV(r,z)$ (cf. [22, 24, 25, 26]), the test function with boundary conditions specified in Figure 1. Then we apply the integration by parts to obtain

$$\begin{aligned}\int_{\Omega}\left(\frac{\partial W}{\partial z}\frac{\partial V}{\partial z} + \alpha^2\frac{\partial W}{\partial r}\frac{\partial V}{\partial r} + \alpha^2\frac{WV}{r^2}\right) r dr dz - \int_{r=1} \alpha^2 WV dz\\ = \int_{\Omega}\left(-\frac{\partial\Theta}{\partial z}\frac{\partial W}{\partial z} - \alpha^2\frac{\partial\Theta}{\partial r}\left(\frac{\partial W}{\partial r} - \frac{W}{r}\right)\right) V r dr dz.\end{aligned} \tag{67}$$

We do the same for (2), with a test function $rV(r,z)$, but having different boundary conditions shown in Figure 1:

$$\begin{aligned}\int_\Omega \left(\frac{\partial\Theta}{\partial z}\frac{\partial V}{\partial z} + \alpha^2 \frac{\partial\Theta}{\partial r}\frac{\partial V}{\partial r}\right) r dr dz - \int_{z=1} r(B\vartheta_a - B\Theta)V dr \\ = \mathrm{Na}\int_\Omega e^{-\Theta}\left(\left(\frac{\partial W}{\partial z}\right)^2 + \alpha^2\left(\frac{\partial W}{\partial r} - \frac{W}{r}\right)^2\right) V r dr dz.\end{aligned} \tag{68}$$

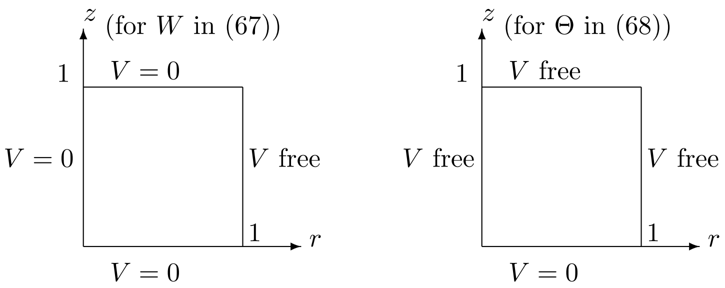


FIG. 1. *Boundary conditions for the test functions V (for W and Θ in* (67)–(68)*).*

To obtain homogeneous boundary conditions for the variational problems (67)–(68), we use the following decompositions:

$$W = W^b + W^0, \qquad W^b = rz, \tag{69}$$

$$\Theta = \Theta^b + \Theta^0, \qquad \Theta^b = \vartheta_w + z\frac{B}{1+B}(\vartheta_a - \vartheta_w). \tag{70}$$

We seek solutions W^0 and Θ^0 instead, which have homogeneous boundary conditions, also depicted in Figure 2:

$$W^0\big|_{r=0,z=0,z=1} = 0, \qquad \left.\frac{\partial W^0}{\partial r}\right|_{r=1} = \frac{W^0}{r},$$

$$\Theta^0\big|_{z=0} = 0, \qquad \left.\frac{\partial \Theta^0}{\partial r}\right|_{r=0,r=1} = 0, \qquad \left.\frac{\partial \Theta^0}{\partial z}\right|_{z=1} = -B\Theta^0.$$

That is, we will find finite element solutions W_h^0 and Θ_h^0, where h stands for the grid size.

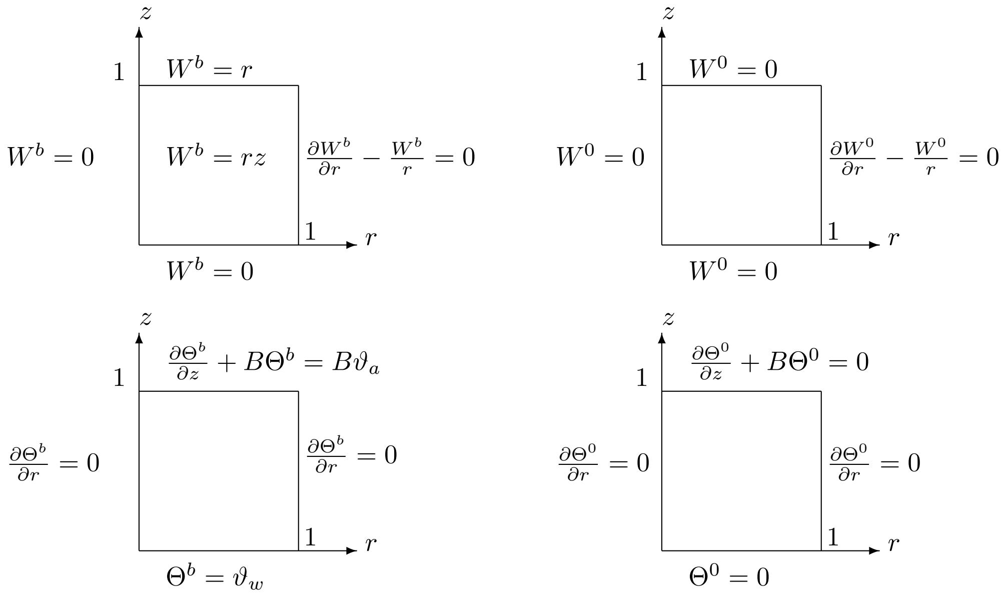
To discretize (67) and (68), due to the special domain of the unit square, one may use spectral methods (cf. [22]) or tensor product methods (cf. [23]) to get a high order approximation. To handle the nonlinearity of the coupled system, and to handle possible irregular domains in future, we use Q^k finite elements, continuous and piecewise polynomials of separate degree k or less, on uniform grids $\mathcal{K}_h = \{K \mid K = [r_i - h, r_i] \times [z_j - h, z_j],\ i, j = 1, \ldots, 1/h\}$ of Ω:

$$\mathcal{Q}_h := \left\{ V \in C(\Omega) \mid V|_K = \sum_{0 \le i,j \le k} a_{ij} r^i z^j \ \forall K \in \mathcal{K}_h \right\} \subset H^1(\Omega).$$

We use the following notation for the discrete spaces with homogeneous boundary conditions:

$$\mathcal{Q}_{h,W} := \mathcal{Q}_h \cap \{V = V(r,z) \in C(\Omega) \mid V(0,z) = V(r,0) = V(r,1) = 0\}, \tag{71}$$

$$\mathcal{Q}_{h,\Theta} := \mathcal{Q}_h \cap \{V = V(r,z) \in C(\Omega) \mid V(r,0) = 0\}. \tag{72}$$



z
1
$W^b = r$
$W^b = 0$
$W^b = rz$
$\frac{\partial W^b}{\partial r} - \frac{W^b}{r} = 0$
1
r
$W^b = 0$
z
1
$W^0 = 0$
$W^0 = 0$
$\frac{\partial W^0}{\partial r} - \frac{W^0}{r} = 0$
1
r
$W^0 = 0$
z
1
$\frac{\partial \Theta^b}{\partial z} + B\Theta^b = B\vartheta_a$
$\frac{\partial \Theta^b}{\partial r} = 0$
$\frac{\partial \Theta^b}{\partial r} = 0$
1
r
$\Theta^b = \vartheta_w$
z
1
$\frac{\partial \Theta^0}{\partial z} + B\Theta^0 = 0$
$\frac{\partial \Theta^0}{\partial r} = 0$
$\frac{\partial \Theta^0}{\partial r} = 0$
1
r
$\Theta^0 = 0$

FIG. 2. *Boundary conditions for W^0 and Θ^0 in* (69)–(70).

The finite element discretizations of (67)–(68) read as follows: Find $(W_h^0, \Theta_h^0) \in \mathcal{Q}_{h,W} \times \mathcal{Q}_{h,\Theta}$ such that

$$A_W(W_h^0, V) = F_{W,\Theta}(V) - A_W(W^b, V) \qquad \forall V \in \mathcal{Q}_{h,W}, \tag{73}$$

$$A_\Theta(\Theta_h^0, V) = G_{W,\Theta}(V) - A_\Theta(\Theta^b, V) + Bc_z(\vartheta_a, V) \qquad \forall V \in \mathcal{Q}_{h,\Theta}, \tag{74}$$

where the bilinear forms and functionals are defined by

$$A_W(U, V) = a(U, V) + \alpha^2 \left(\frac{U}{r}, \frac{V}{r}\right)_r - \alpha^2 c_r(U, V), \tag{75}$$

$$A_\Theta(U, V) = a(U, V) + Bc_z(U, V), \tag{76}$$

$$F_{W,\Theta}(V) = \left(-\frac{\partial \Theta}{\partial z}\frac{\partial W}{\partial z} - \alpha^2 \frac{\partial \Theta}{\partial r}\left(\frac{\partial W}{\partial r} - \frac{W}{r}\right), V\right)_r, \tag{77}$$

$$G_{W,\Theta}(V) = \mathrm{Na}\left(e^{-\Theta}\left[\left(\frac{\partial W}{\partial z}\right)^2 + \alpha^2\left(\frac{\partial W}{\partial r} - \frac{W}{r}\right)^2\right], V\right)_r, \tag{78}$$

$$a(U, V) = \int_\Omega \left(\frac{\partial U}{\partial z}\frac{\partial V}{\partial z} + \alpha^2 \frac{\partial U}{\partial r}\frac{\partial V}{\partial r}\right) r dr dz, \tag{79}$$

$$(U, V)_r = \int_\Omega UV r dr dz, \tag{80}$$

$$c_r(U, V) = \int_{r=1, 0 \le z \le 1} UV dz, \tag{81}$$

$$c_z(U, V) = \int_{z=1, 0 \le r \le 1} rUV dz. \tag{82}$$

We solve the nonlinear system of equations (73)–(74) numerically by a straightforward Seidel iteration. That is, initially given some guesses (both zero in computation) of W_h^0 and Θ_h^0, we generate the right-hand side of (73) and use the conjugate gradient method to solve (73) to get a new W_h^0. Then the new W_h^0 and the old Θ_h^0 would be used to generate the right-hand side vector in (74). We solve (74) again by the conjugate gradient method to get a new Θ_h^0. The next lemma shows that the two linear systems at each step described above are uniquely solvable, because both the coefficient matrices are symmetric and positive definite.

LEMMA 4.1. *For any $V \in \mathcal{Q}_{h,\Theta} \cup \mathcal{Q}_{h,W}$ and $V \neq 0$,*

$$a(V,V) > 0. \tag{83}$$

For any $V \in \mathcal{Q}_{h,W}$ and $V \neq 0$,

$$A_W(V,V) > 0. \tag{84}$$

For any $V \in \mathcal{Q}_{h,\Theta}$, $V \neq 0$, and $B \geq 0$,

$$A_\Theta(V,V) > 0. \tag{85}$$

Proof. Equations (83) and (84) are shown in [27]. Equation (85) is a corollary of (83), noting the sign of B is positive. □

ALGORITHM 4.1. *The coupled nonlinear system* (73)–(74) *is solved by the Seidel iteration with the given initial guess $W_{h,0}^0 = 0$ and $W_{\Theta,0}^0 = 0$. For $j = 1, 2, \ldots,$*

$$W_{h,j}^0 = W_{h,j-1}^0 + e_W,$$

where e_W solves the equation

$$A_W(e_W, V) = F_{W_{j-1},\Theta_{j-1}}(V) - A_W(W^b, V) - A_W(W_{h,j-1}^0, V) \qquad \forall V \in \mathcal{Q}_{h,W} \tag{86}$$

and

$$\Theta_{h,j}^0 = \Theta_{h,j-1}^0 + e_\Theta,$$

where e_Θ solves the equation

(87)

$$A_\Theta(e_\Theta, V) = G_{W_j,\Theta_{j-1}}(V) - A_\Theta(\Theta^b, V) + Bc_z(\vartheta_a, V) - A_\Theta(\Theta_{h,j-1}^0, V) \qquad \forall V \in \mathcal{Q}_{h,\Theta}.$$

Here $W_j = W^b + W_{h,j}^0$ and $\Theta_j = \Theta^b + \Theta_{h,j}^0$ for $j = 0, 1, 2, \ldots.$

A typical pair of solutions (W_h, Θ_h) is shown in Figure 3.

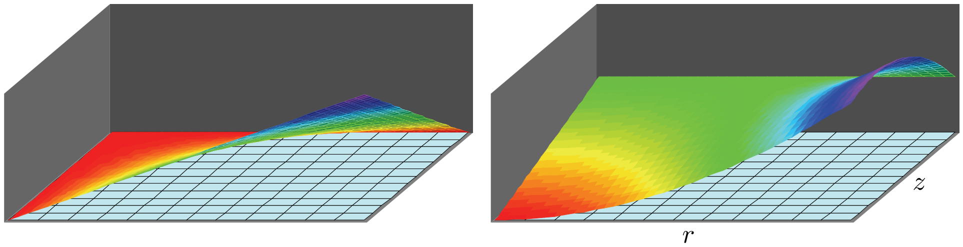


FIG. 3. *Solutions W and Θ for* (1) *and* (2) *when $\alpha = .01$, $Na = 1$, $\vartheta_w = 1.5$, $\vartheta_a = 1$, $B = 0.1$.*

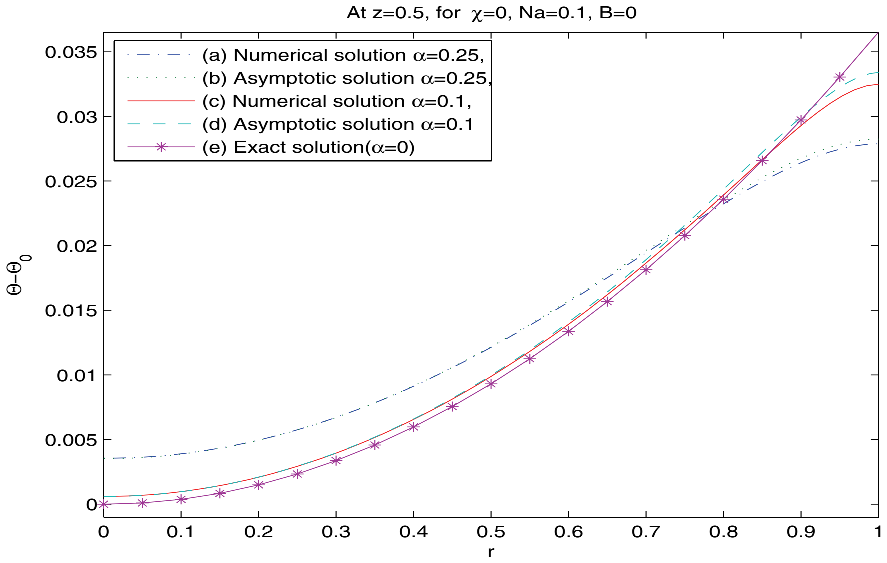


FIG. 4. *Solutions obtained by* (73)–(74), (29)–(30), *and* (11)–(12).

5. Discussion. In this section we compare the analytical solutions obtained in section 3 with the finite element solution obtained in section 4. The exact solution given in (11)–(12) is valid for $\alpha = 0$, $B = 0$, and all values of Na, the perturbation solution given in (29)–(30) is valid for all values of α and small Na, while the solution found in (65)–(66) is valid only for small α and Na. The numerical solution, on the other hand, is valid for all parameter values.

The plots in Figures 4–8 depict the deviation of the temperature and velocity from the isothermal solution. We plot $\Theta - \Theta_0$ and $W - W_0$, where Θ_0 and W_0 are the isothermal solutions. Figures 4 and 5 show the deviation of temperature at $z = 0.5$ and $r = 0.5$, respectively, for Na $= 0.1$, $B = 0$, and selected values of α. The case $B = 0$ corresponds to insulated boundary condition on the upper plate. For $\alpha = 0.1$ all three solutions agree very well except near the free surface $r = 1$. The error in the exact solution for $\alpha = 0$ arises because it does not satisfy the boundary condition at the free surface. The error between the numerical and asymptotic solutions is otherwise very small. Figure 6 shows the deviation of the velocity for the same values of the parameters. The agreement among all three solutions is again very good. In Figure 7, the deviation in temperature is shown for $\alpha = 0, 1$, $B = 1$, $\theta_w = 1.0$, $\theta_a = 1.5$ for two values of Na. While the agreement between numerical and asymptotic solutions is excellent for Na $= 0.1$, there is a small discrepancy for the case Na $= 1.0$. This is actually quite good since the perturbation expansion was truncated at order $O(\text{Na})$.

Figure 8 shows the deviation of the velocity for a large value of the Biot number B. In this limit the two plates are nearly isothermal, and we see a qualitative difference from the solution for $B = O(1)$. Specifically, the profile is symmetric about the midplane $z = 0$, whereas for smaller values of B the profile is asymmetric. Another qualitative difference is the location of the maximum temperature. On any fixed

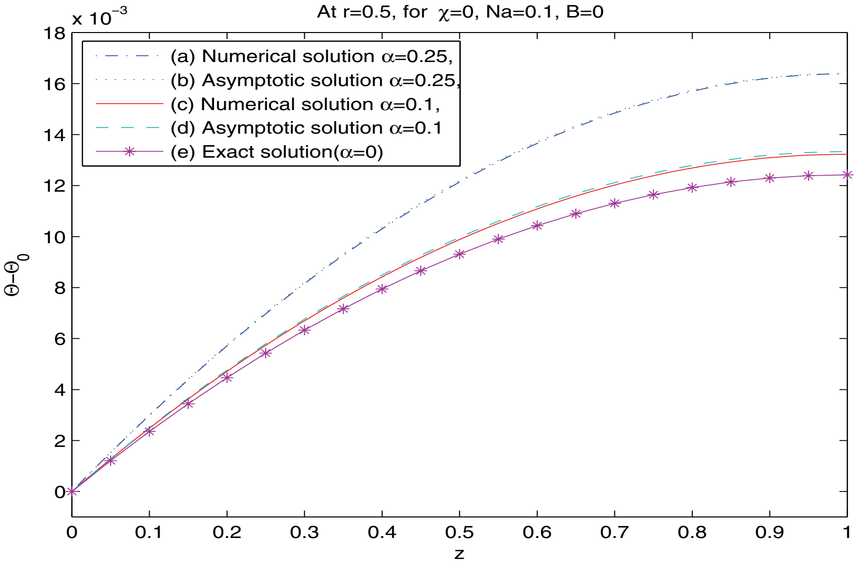


Fig. 5. *Solutions obtained by* (73)–(74), (29)–(30), *and* (11)–(12).

plane the maximum occurs at the free surface $r = 1$ when $B = 0$. As B increases the location of the maximum moves away from the free surface. From these figures we also see that the deviation in temperature and velocity from the isothermal solution is order $O(\mathrm{Na})$ when Na is small.

Finally, in Figure 9 we plot the torque on the lower stationary plate as a function of the Nahme number Na for selected values of α and B. Although the exact solution is valid only for $\alpha = 0$ and $B = 0$ the agreement with the numerical solution for $\alpha = 0.1$ and $B = 0.1$ is excellent. Thus, in applications in which the aspect ratio α is small the exact solution can be used to obtain very accurate corrections to the torque in viscometric applications. We also show the series representation for the torque equation (13), and the agreement is excellent for $\mathrm{Na} < 2$.

6. Summary. Nonisothermal torsional flow with the heat transfer boundary condition at the upper rotating plate, the isothermal boundary condition at the lower stationary plate, and the insulated boundary condition at the fluid/air interface has been analyzed. It is assumed that viscosity is an exponential function of temperature. We have obtained analytical solutions valid in the limit of small aspect ratio α and in the limit of small Nahme–Griffith number Na. The nonlinear coupled partial differential equations have also been solved numerically using the finite element method. Our results show that the asymptotic solutions agree very well with the numerical solution. For small vales of Na the deviation of temperature and velocity from the isothermal solution is small, approximately of order $O(\mathrm{Na})$. Furthermore, we show that for viscometric applications in which the aspect ratio α is typically less than 0.1, the exact solution obtained for $\alpha = 0$ and $B = 0$ gives very accurate results for the nonisothermal correction to the torque for small values of α and the Biot number B.

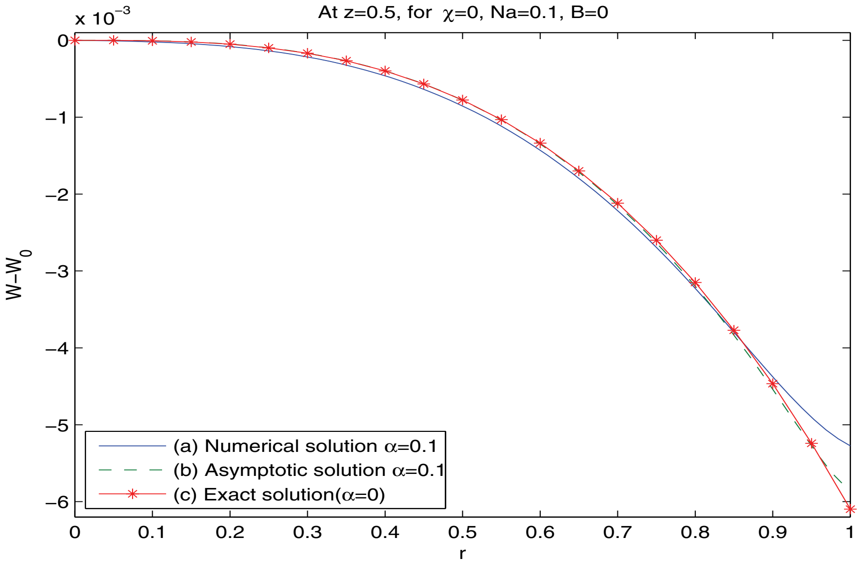


FIG. 6. *Solutions obtained by* (73)–(74), (29)–(30), *and* (11)–(12).

Appendix.

A.1. Heat transfer boundary condition. In this section, the derivation of the heat transfer boundary condition at the upper rotating plate is detailed. Following the convention adopted in Ozisik [29],

$$-k\frac{\partial \tilde{T}}{\partial \tilde{z}} = \hbar\left(\tilde{T} - \tilde{T}_a\right) \qquad \text{at the surface S,} \tag{88}$$

where $\hbar$ is the heat transfer coefficient and k is the thermal conductivity of the rotating plate. The surface S corresponds to the surface of the upper rotating plate at $z = 1$. After normalization of variables and introducing the thickness of the plate H, to procure a meaningful parameter, the Biot number $Bi = \frac{\hbar H}{k}$:

$$\frac{\partial T}{\partial z} = \frac{Bih}{H}\left(-T + T_a\right), \tag{89}$$

where h is the thickness of the gap between the two plates [28, 29]. In dimensionless form this becomes

$$\frac{\partial \Theta}{\partial z} + B\Theta = B\vartheta_a, \tag{90}$$

where $B \equiv \frac{Bi\ h}{H}$.

A.2. The function $F_m(r)$. The function $F_m(r)$ appearing in (30) satisfies the following ordinary differential equation:

$$r^2 F_m'' + rF_m' - \left(1 + \frac{m^2\pi^2 r^2}{\alpha^2}\right)F_m = \varphi(r),$$

$$\varphi = 2r^3 \sum_{n=1}^{\infty}\left[\frac{m\pi\bar{\Gamma}_n\left(1 - \cos(m\pi)\cos(\lambda_n)\right)}{m^2\pi^2 - \lambda_n^2}\right]\left[\frac{2\alpha}{\lambda_n^2}\frac{I_0(\frac{\lambda_n r}{\alpha})}{I_1(\frac{\lambda_n}{\alpha})} - \frac{r^2}{\lambda_n^2} - \frac{4\alpha^2}{\lambda_n^4}\right]. \tag{91}$$

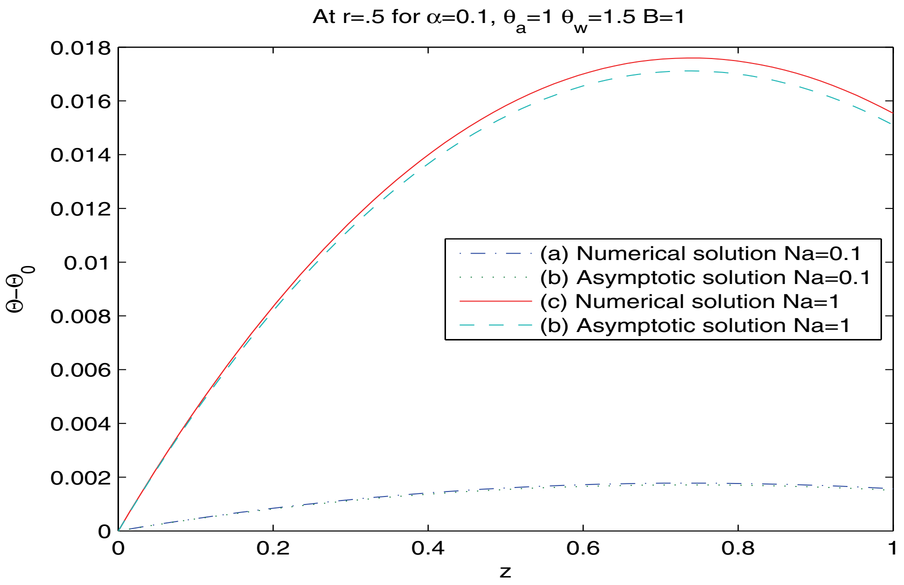


FIG. 7. *Solutions obtained by* (73)–(74), (29)–(30), *and* (11)–(12).

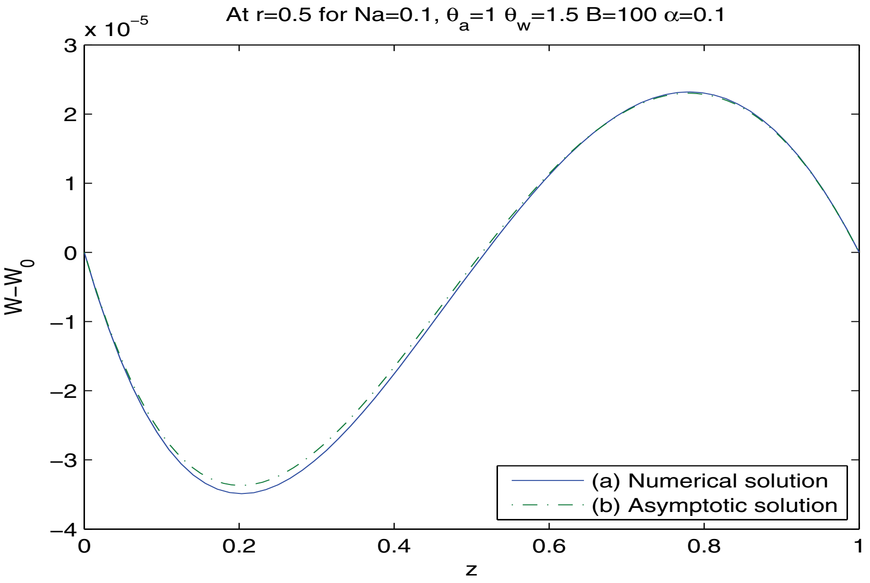


FIG. 8. *Solutions obtained by* (73)–(74), (29)–(30), *and* (11)–(12).

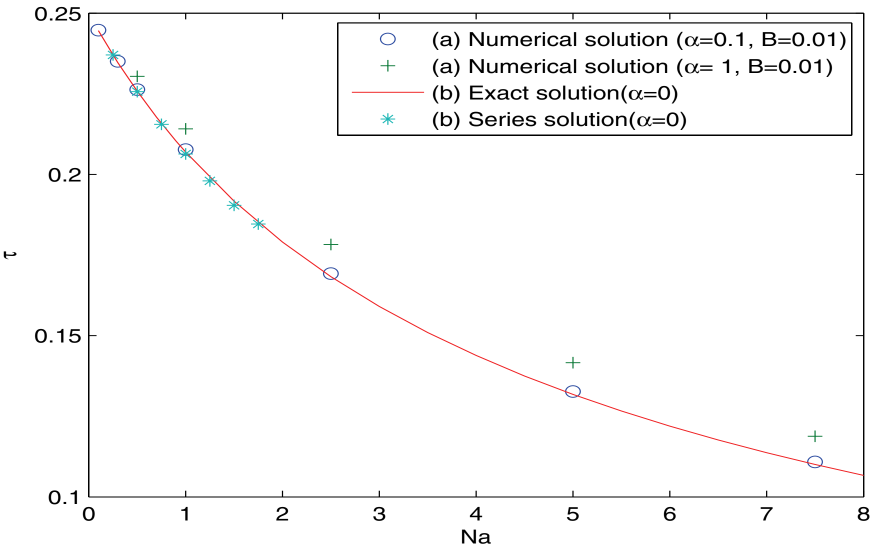


FIG. 9. *Plot of the torque $\mathcal{T}$ versus* Na.

The general solution to the above ordinary differential equation is

$$\begin{aligned} F_m(r) = {} & C_1 I_1\left(\frac{m\pi r}{\alpha}\right) + C_2 K_1\left(\frac{m\pi r}{\alpha}\right) + I_1\left(\frac{m\pi r}{\alpha}\right)\left[\int \frac{\varphi(r)K_1(\frac{m\pi r}{\alpha})}{r}\mathrm{d}r\right] \\ & - K_1\left(\frac{m\pi r}{\alpha}\right)\left[\int \frac{\varphi(r)I_1(\frac{m\pi r}{\alpha})}{r}\mathrm{d}r\right], \end{aligned} \tag{92}$$

where K_0 and K_1 are modified Bessel functions of the second kind.

The evaluation of the integrals is shown next:

$$\begin{aligned} \int \frac{\varphi(r)K_1(\frac{m\pi r}{\alpha})}{r}\mathrm{d}r = {} & 2\sum_{n=1}^{\infty} \frac{m\pi\bar{\Gamma}_n\,(1-\cos(m\pi)\cos(\lambda_n))}{\lambda_n(m^2\pi^2-\lambda_n^2)}\left[\frac{-4m\pi r\alpha^3}{\lambda_n(m^2\pi^2-\lambda_n^2)^2}\right. \\ & \times\frac{\lambda_n I_1(\frac{\lambda_n r}{\alpha})K_0(\frac{m\pi r}{\alpha}) + m\pi I_0(\frac{\lambda_n r}{\alpha})K_1(\frac{m\pi r}{\alpha})}{I_1(\frac{\lambda_n}{\alpha})} - \frac{2r^2\alpha^2}{\lambda_n(m^2\pi^2-\lambda_n^2)} \\ & \times\frac{\lambda_n I_1(\frac{\lambda_n r}{\alpha})K_1(\frac{m\pi r}{\alpha}) + m\pi I_0(\frac{\lambda_n r}{\alpha})K_0(\frac{m\pi r}{\alpha})}{I_1(\frac{\lambda_n}{\alpha})} + \frac{\alpha r^4}{m\pi}K_2\left(\frac{m\pi r}{\alpha}\right) \\ & \left. + \frac{2\alpha^2 r^3}{m^2\pi^2}K_3\left(\frac{m\pi r}{\alpha}\right) + \frac{4\alpha^3 r^2}{m\pi\lambda_n^2}K_2\left(\frac{m\pi r}{\alpha}\right)\right], \end{aligned} \tag{93}$$

$$\begin{aligned} \int \frac{\varphi(r)I_1(\frac{m\pi r}{\alpha})}{r}\mathrm{d}r = {} & 2\sum_{n=1}^{\infty} \frac{m\pi\bar{\Gamma}_n(1-\cos(m\pi)\cos(\lambda_n))}{\lambda_n(m^2\pi^2-\lambda_n^2)}\left[\frac{4m\pi r\alpha^3}{\lambda_n(m^2\pi^2-\lambda_n^2)^2}\right. \\ & \times\frac{\lambda_n I_0(\frac{m\pi r}{\alpha})I_1(\frac{\lambda_n r}{\alpha}) - m\pi I_0(\frac{\lambda_n r}{\alpha})I_1(\frac{m\pi r}{\alpha})}{I_1(\frac{\lambda_n}{\alpha})} - \frac{2r^2\alpha^2}{\lambda_n(m^2\pi^2-\lambda_n^2)} \end{aligned} \tag{94}$$

$$\times\frac{\lambda_n I_1(\frac{m\pi r}{\alpha} I_1(\frac{\lambda_n r}{\alpha}) - m\pi I_0(\frac{\lambda_n r}{\alpha}) I_0)\frac{m\pi r}{\alpha})}{I_1(\frac{\lambda_n}{\alpha})} - \frac{\alpha r^4}{m\pi} I_2\left(\frac{m\pi r}{\alpha}\right)$$
$$+ \frac{2\alpha^2 r^3}{m^2\pi^2} I_3\left(\frac{m\pi r}{\alpha}\right) - \frac{4\alpha^3 r^2}{m\pi\lambda_n^2} I_2\left(\frac{m\pi r}{\alpha}\right)\Bigg].$$

The boundary conditions $F_m = 0$ at $r = 0$ and $F_m' - \frac{F_m}{r} = 0$ at $r = 1$ are used to determine the constants in the above equation. The boundary condition on the axis causes the constant of integration C_2 to be zero. The other boundary condition is used to determine the constant of integration C_1. However, because of the complexity of the nature of the solution as judged from the above equations, the constant C_1 is shown to be determined in principle. However, the actual formulation is employed to generate plots via CAS (computer algebra system):

$$\begin{aligned} C_1 = &- \left[\int \frac{\varphi(r) K_1(\frac{m\pi r}{\alpha})}{r} \mathrm{d}r\right]_{r=1} \\ &- \frac{K_0(\frac{m\pi}{\alpha}) + K_2(\frac{m\pi}{\alpha}) - \frac{2\alpha}{m\pi} K_1(\frac{m\pi}{\alpha})}{I_0(\frac{m\pi}{\alpha}) + I_2(\frac{m\pi}{\alpha}) - \frac{2\alpha}{m\pi} I_1(\frac{m\pi}{\alpha})} \left[\int \frac{\varphi(r) I_1(\frac{m\pi r}{\alpha})}{r} \mathrm{d}r\right]_{r=1}. \end{aligned} \tag{95}$$

REFERENCES

[1] R. Nahme, *Betrage zur hydrodynamischen Theorie der Lagerreibung*, Ing. Arch., 11 (1940), pp. 191–209.

[2] E. A. Kearsley, *The viscous heating correction for viscometric flows*, Trans. Soc. Rheol., 6 (1962), pp. 253–261.

[3] R. B. Bird and R. Turian, *Viscous heating effects in a cone and plate viscometer*, Chem. Eng. Sci., 17 (1962), pp. 331–334.

[4] R. M. Turian and R. B. Bird, *Viscous heating in the cone-and-plate viscometer*-II. *Newtonian fluids with temperature-dependent viscosity and thermal conductivity*, Chem. Eng. Sci., 18 (1963), pp. 689–696.

[5] R. Turian, *Viscous heating in the cone-and-plate viscometer*-III., Chem. Eng. Sci., 20 (1965), pp. 771–781.

[6] B. Martin, *Some analytical solutions for viscometric flows of power-law fluids with heat generation and temperature dependent viscosity*, Int. J. Non-Linear Mech., 2 (1967), pp. 285–301.

[7] J. Gavis and R. L. Laurence, *Viscous heating in plane and circular flow between moving surfaces*, I & EC Fundamentals, 7 (1968), pp. 232–239.

[8] J. J. Magda and R. G. Larson, *A transition occurring in ideal elastic liquids during shearing flows*, J. Non-Newtonian Fluid Mech., 30 (1988), pp. 1–19.

[9] G. H. McKinley, J. A. Byars, R. A. Brown, and R. C. Armstrong, *Observations of elastic instability in cone-and-plate and parallel-plate flows of polyisobutylene Boger fluid*, J. Non-Newtonian Fluid Mech., 40 (1991), pp. 201–229.

[10] A. Oztekin and R. A. Brown, *Instability of a viscoelastic fluid between rotating parallel disks: Analysis of the Oldroyd-B fluid*, J. Fluid Mech., 225 (1993), pp. 473–502.

[11] A. Avagliano and N. Phan-Thien, *Torsional flow: Elastic instability in a finite domain*, J. Fluid Mech., 312 (1996), pp. 279–298.

[12] D. O. Olagunju, *On short wave elastic instability in parallel plate flow*, in Proceedings of the 1997 ASME Congress and Exposition, Rheology, and Fluid Mechanics of Nonlinear Materials, Dallas TX, 1997, pp. 243–248.

[13] Y. Renardy and M. Renardy, *A model equation for axisymmetric stability of small-gap parallel-plate flow*, J. Non-Newtonian Fluid Mech., 77 (1998), pp. 103–114.

[14] J. P. Rothstein and G. H. McKinley, *Non-isothermal modification of purely elastic flow instabilities in torsional flow of polymeric fluids*, Phys. Fluids, 13 (2001), pp. 382–396.

[15] D. O. Olagunju, L. P. Cook, and G. H. McKinley, *Effects of viscous heating on linear stability of viscoelastic cone-and-plate flow: Axisymmetric case*, J. Non-Newtonian Fluid Mech., 102 (2002), pp. 321–342.

[16] M. T. Arigo, *The Effects of Fluid Rheology on the Dynamics of Isothermal and Non-Isothermal Flows of Viscoelastic Fluids*, Doctoral Dissertation, Harvard University, Cambridge, MA, 1999.

[17] A. U. Al-Mubaiyedh, R. Sureshkumar, and B. Khomami, *Influence of energetics on the stability of viscoelastic Taylor–Couette flow*, Phys. Fluids, 11 (1999), pp. 3217–3226.

[18] B. M. Baumert and S. J. Muller, *Flow visualization of the elastic Taylor–Couette instability in Boger fluids*, Rheol. Acta, 34 (1995), pp. 147–159.

[19] B. M. Baumert and S. J. Muller, *Flow regimes in model viscoelastic fluid in a circular Couette system with independently rotating cylinders*, Phys. Fluids, 9 (1997), pp. 566–586.

[20] D. O. Olagunju, *Secondary flow in non-isothermal viscoelastic parallel-plate flow*, J. Engrg. Math., 51 (2005), pp. 325–338.

[21] D. O. Olagunju, *Analytical solutions for non-isothermal viscoelastic torsional flow in a bounded domain*, J. Non-Newtonian Fluid Mech., 112 (2003), pp. 85–100.

[22] B. Heinrich, *The Fourier-finite-element method for Poisson's equation in axisymmetric domains with edges*, SIAM J. Numer. Anal., 33 (1996), pp. 1885–1911.

[23] S. Börm and R. Hiptmair, *Analysis of tensor product multigrid*, Numer. Algorithms, 26 (2001), pp. 331–356.

[24] S. Börm and R. Hiptmair, *Multigrid computation of axisymmetric electromagnetic fields*, Adv. Comput. Math., 16 (2002), pp. 331–356.

[25] T. E. Price, *Numerically exact integration of a family of axisymmetric finite elements*, Comm. Numer. Methods Engrg., 19 (2003), pp. 253–261.

[26] J. D. Clayton and J. J. Rencis, *Numerical integration in the axisymmetric finite element formulation*, Adv. Engng. Soft., 31 (2001), pp. 137–141.

[27] S. Zhang and D. Olagunju, *Axisymmetric finite element solution of non-isothermal parallel-plate flow*, Appl. Math. Comput., 171 (2005), pp. 1081–1094.

[28] H. H. Winter, *Viscous dissipation in shear flows of molten polymers*, Adv. Heat Transfer, 13 (1977), pp. 205–267.

[29] M. N. Ozisik, *Heat Conduction*, Vol. 14, Wiley-Interscience, New York, 1980.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1423–1438

STABILITY OF SOLITARY WAVES IN A SEMICONDUCTOR DRIFT-DIFFUSION MODEL*

C. M. CUESTA† AND C. SCHMEISER‡

Abstract. We consider a macroscopic (drift-diffusion) model describing a simple microwave generator, consisting of a special type of semiconductor material that, when biased above a certain threshold voltage, generates charge waves. These waves correspond to travelling wave solutions of the model equation which, however, turn out to be unstable in a standard formulation of the travelling wave problem. Here a different formulation of this problem is considered, where an external voltage condition is applied in the form of an integral constraint. Global existence of this novel Cauchy problem is proven and the results of numerical experiments are presented, which suggest the stability of solitary waves. In addition, a small amplitude limit is considered, for which linearized orbital stability of solitary waves can be proven.

Key words. Gunn effect, drift-diffusion equation, solitary waves, global constraint

AMS subject classifications. 82D37, 35K55, 35B40

DOI. 10.1137/070690766

1. Introduction.

In this paper we consider the nondimensionalized one-dimensional semiconductor drift-diffusion model

$$\partial_t n = \partial_x(\partial_x n - v(E)\, n)\,, \tag{1.1}$$

$$\partial_x E = n - 1 \tag{1.2}$$

for $(x,t) \in \mathbb{R} \times (0,\infty)$, where $n(x,t)$ denotes the electron density and $E(x,t)$ the (negative) electric field. In the drift-diffusion equation (1.1), $v(E)$ is the field dependent drift velocity, and in the Poisson equation (1.2), the constant 1 represents the scaled constant doping concentration. The special feature of the model is the nonmonotonicity of $v(E)$, made precise below.

The system will be considered subject to the initial condition

$$n(0,x) = n_I(x) \quad \text{for all } x \in \mathbb{R}\,, \tag{1.3}$$

where initially and, thus, for all times, we assume global charge neutrality:

$$\int_{\mathbb{R}} (n_I - 1)dx = 0\,.$$

This has the consequence that the field takes the same value

$$E_\infty(t) := \lim_{|x|\to\infty} E(t,x)$$

*Received by the editors May 7, 2007; accepted for publication (in revised form) January 14, 2008; published electronically May 2, 2008. This work was partially supported by the Austrian Science Fund through project P18367.

http://www.siam.org/journals/siap/68-5/69076.html

†School of Mathematical Sciences, Division of Theoretical Mechanics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (carlota.cuesta@maths.nottingham.ac.uk). The work of this author was partially supported by the Engineering and Physical Sciences Research Council in the form of a Research Fellowship and by the Austrian Science Fund.

‡Faculty of Mathematics, University of Vienna, Nordbergstraße 15, 1090 Vienna, Austria, and Johann Radon Institute for Computational and Applied Mathematics, A-4040 Linz, Austria (christian.schmeiser@univie.ac.at). The work of this author was partially supported by the Austrian Science Fund (project W8) and from the EU funded DEASE network (contract MEST-CT-2005-021122).

at $x = \pm\infty$. Instead of prescribing $E_\infty(t)$, we leave it as an unknown and pose the integral constraint

$$\int_{\mathbb{R}} (E(t,x) - E_\infty(t))\, dx = U(t)\,, \tag{1.4}$$

where the function $U(t)$ is given for $t \geq 0$.

This problem arises from a one-dimensional model of a simple microwave generator. When biased above a certain voltage threshold, the generator produces current oscillations based on dipole charge waves travelling through the semiconductor material. This is known as the Gunn effect; see [4] and [5].

The system (1.1), (1.2) subject to (1.3), (1.4) will be motivated below by scaling arguments. We start with the unscaled equations describing the flow of electrons in a piece of homogeneous n-type semiconductor material of length L (cf. [9]),

$$\partial_t n = \partial_x (D\partial_x n - v(E)\, n)\,, \quad \text{with } t > 0,\ x \in (-L, L)\,, \tag{1.5}$$

$$\varepsilon_s \partial_x E = q(n - C), \quad \text{with } x \in (-L, L)\,. \tag{1.6}$$

This is the standard unipolar drift-diffusion model where the transport of holes is neglected. The constant parameters are the diffusivity D, the permittivity ε_s of the semiconductor material, the elementary charge q, and the donor concentration $C > 0$. Since this fixed background charge density is positive, the negatively charged electrons will dominate among the mobile charges, satisfying the omission of the positively charged holes from the model. The function v stands for the drift velocity of electrons and depends on the field, thus leading to a nonlinear coupling of the system, which is supplemented by an initial condition $n(0,x) = n_I(x)$ and by Dirichlet boundary conditions for the electron concentration:

$$n(t,-L) = n(t,L) = C \quad \text{for } t > 0\,. \tag{1.7}$$

In addition, the application of an exterior (given) voltage $\bar{U}$ is described by the integral condition

$$\int_{-L}^{L} E(t,x)\, dx = \bar{U}(t)\,. \tag{1.8}$$

For standard semiconductor materials such as silicon, measurements of the drift velocity $v(E)$ yield an odd nonlinear increasing function of the field E, almost linear for small fields, and bounded from above by a velocity saturation value v_{sat}. However, there are semiconductor materials such as *gallium arsenide* (GaAs), for which the velocity v reaches a maximum at a certain threshold value of the field E_T (cf. [13]), with the profile of v decreasing for $E > E_T$ to v_{sat}; see Figure 1. This nonmonotonicity of the velocity is responsible for the existence of pulse like solutions, namely solitary (travelling) waves, which are necessary for the Gunn effect. We are interested in studying the stability of these waves.

Using L as characteristic length, L/v_{sat} as characteristic time, v_{sat} as characteristic velocity, C as characteristic electron density, and E_T as characteristic field strength, one obtains the dimensionless equations

$$\partial_t n = \partial_x (\nu \partial_x n - n v(E))\,, \tag{1.9}$$

$$\lambda^2 \partial_x E = n - 1\,, \tag{1.10}$$

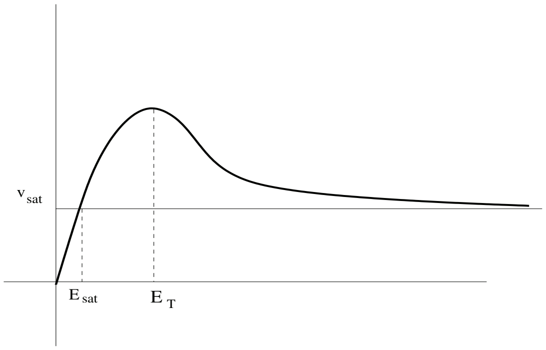


FIG. 1. *Electron drift velocity.*

subject to the conditions

$$n(t,-1) = n(t,1) = 1\,, \tag{1.11}$$

$$\int_{-1}^{1} E(t,x)\,dx = \bar{U}(t)\,, \tag{1.12}$$

where the drift velocity v is now normalized in the sense that it takes its maximum at $E = 1$ and satisfies $\lim_{E\to\infty} v(E) = 1$. The dimensionless parameters

$$\lambda^2 = \frac{\varepsilon_s E_T}{L^2 qC}\,, \quad \nu = \frac{D}{Lv_{sat}}$$

are, respectively, the square of the scaled Debye length and the relative strength of diffusive and convective terms. We are interested in the case of a high doping concentration and a long device; therefore the parameters λ^2 and ν are both small. We shall make the scaling assumption that they are of the same order of magnitude and, for simplicity, actually set $\nu = \lambda^2$.

We recall that for a given constant voltage, the homogeneous steady state solution

$$n \equiv 1, \quad E \equiv \frac{1}{2}\bar{U}$$

of (1.9), (1.10) is stable if $\bar{U} \leq 2$ ($E \leq 1$) and unstable if $\bar{U} > 2$ ($E > 1$); cf. [14], [1]. Stable solitary waves are expected to arise in the latter case. The appropriate space-time scaling for these waves is achieved by $(t,x) \to (t/\lambda^2, x/\lambda^2)$, which expands both the temporal and the spatial domains. It leads to (1.1)–(1.2), and the integral condition (1.12) becomes

$$\lambda^2 \int_{-\frac{1}{\lambda^2}}^{\frac{1}{\lambda^2}} E(t,x)\,dx = \bar{U}(t)\,. \tag{1.13}$$

In the "Gunn operation mode" we expect waves travelling through the device, whose typical length is of order one in terms of the new x-variable. Away from the wave, i.e., in most of the device, we expect an almost constant electric field, and we denote an approximation by $E_{1/\lambda^2}(t)$. The condition (1.13) can then be rewritten as

$$\lambda^2 \int_{-\frac{1}{\lambda^2}}^{\frac{1}{\lambda^2}} (E(t,x) - E_{1/\lambda^2}(t))\,dx = \bar{U}(t) - 2E_{1/\lambda^2}(t)\,. \tag{1.14}$$

Passing to the limit $\lambda^2 \to 0$ formally gives $E_\infty(t) = \bar{U}(t)/2$ with $E(t,x) \to E_\infty(t)$ as $|x| \to \infty$. In [15] Szmolyan considered the problem (1.1), (1.2) subject to this boundary condition and an initial condition for n. It is striking that, with standard linearization techniques, he proved that solitary waves are unstable in this case.

These results are rather unexpected if compared with the experimental evidence on Gunn diodes. The aim of this work is to study a reformulation of the problem, which seems to stabilize the solitary waves. Formally, the reformulation can be derived by introducing

$$U(t) := \lim_{\lambda \to 0} \frac{1}{\lambda^2} \left(\bar{U}(t) - 2E_{1/\lambda^2}(t) \right)$$

and passing to the limit in (1.14) after dividing by λ^2. Obviously, this leads to the integral condition (1.4).

In the language of asymptotic analysis, the assumption that the small parameters ν and λ^2 are of the same order of magnitude leads to a significant limit, since the small parameters can then be eliminated from the differential equations by the above rescaling. However, since the ratio λ^2/ν depends on both the device length and the doping concentration, situations where this ratio is either very small or very large can also be physically relevant. An asymptotic analysis of travelling waves in the former case can be found in [9]. It turns out that in this case all travelling wave solutions have a far-field value of the electric field close to E_{sat} (see Figure 1). This result can be seen as a (not very strong) physical justification of prescribing $U(t)$, since this is then close to prescribing the contact voltage $\bar{U}(t)$.

For convenience we introduce the unknown

$$e(t,x) := E(t,x) - E_\infty(t) = \int_{-\infty}^{x} (n(t,y) - 1)\, dy \quad \text{with } t > 0,\ x \in \mathbb{R}\,.$$

Substituting $n = \partial_x e + 1$ into (1.1) and integrating with respect to x gives the equation

$$\partial_t e = \partial_x^2 e - v(e + E_\infty)\, \partial_x e + v(E_\infty) - v(e + E_\infty)\,, \tag{1.15}$$

subject to the initial condition

$$e(0,x) = e_I(x) = \int_{-\infty}^{x} (n_I(y) - 1)\, dy\,, \tag{1.16}$$

with n_I as in (1.3), and to the integral constraint (1.4), which now simply reads

$$\int_{\mathbb{R}} e(t,x)\, dx = U(t)\,. \tag{1.17}$$

Differentiation with respect to time gives

$$U'(t) = \int_{\mathbb{R}} (v(E_\infty(t)) - v(E_\infty(t) + e(t,x)))dx\,. \tag{1.18}$$

We shall solve (1.15) subject to (1.18) instead of (1.17). This will be favorable since (1.18) can be seen as an equation for E_∞ for given $U'(t)$ and e.

The formulation of the problem will be completed by specifying the precise assumptions on the drift velocity.

Assumption 1. We assume $v \in C_B^3([0,\infty))$, $v(0) = 0$, $\operatorname{sign} v'(E) = \operatorname{sign}(1 - E)$, $\lim_{E\to\infty} v(E) = 1$, $\exists\, E_i > 1$ such that $\operatorname{sign} v''(E) = \operatorname{sign}(E - E_i)$. Finally, $v''' \geq 0$ on $(1, E_i)$.

The equation $v(E_{sat}) = 1$ uniquely defines $E_{sat} < 1$. We also introduce $\sigma_i = \sup_{E>0} |d^i v/dE^i(E)|$, $i = 1, 2, 3$.

The paper is organized as follows. In section 2 we review the existence of solitary waves but incorporate the condition (1.4) into the problem. It turns out that for all $U > 0$ there exists a unique (up to translation) solitary wave having $E_\infty < 1$. In section 3 we prove existence of solutions of (1.15)–(1.18) for positive $U(t)$. Actually there is also a restriction on the values of $U'(t)$, which is required to be in the range of the right-hand side of (1.18). The existence proof uses a fixed point argument involving the operator defined by solving the condition (1.18) (for given e). This operator is only locally Lipschitz in $L_x^1(\mathbb{R})$. This difficulty does not ensue in bounded domains; see [8]. There is still no general result on the stability of solitary waves. In section 4, however, we provide strong numerical evidence that we succeeded in stabilizing the travelling waves by the new formulation. Moreover, in section 5 we consider a *small* wave limit by imposing a small external voltage. We prove linear asymptotic stability of the limiting solitary waves. It turns out that the limit equation is the so-called conserved Fisher equation with a constant competition rate, a model of population dynamics with global regulation [11]. In particular, our proof shows linearized asymptotic stability of its stationary solutions.

2. Solitary waves. In this section we prove existence of solitary waves subject to the constraint (1.4). Let $\xi =: x - ct$ be the travelling wave variable, where $c > 0$ is the wave speed. Then a solitary wave solution $(E(\xi), n(\xi))$ of (1.1)–(1.2) is a solution of

$$\begin{aligned} n' &= n(v(E) - c) - v(E_\infty) + c\,, \\ E' &= n - 1 \end{aligned}$$

that satisfies

$$n \to 1 \quad \text{and} \quad E \to E_\infty \qquad \text{as } |\xi| \to \infty\,. \tag{2.1}$$

A straightforward computation using both differential equations leads to

$$\frac{n-1}{n} n' - (v(E) - v(E_\infty))E' = \frac{(n-1)^2}{n}(v(E_\infty) - c)\,.$$

Since the right-hand side does not change sign, integration with respect to ξ and the far-field conditions imply that a solution exists only if $c = v(E_\infty)$ holds, which we assume in the following:

$$n' = n(v(E) - v(E_\infty)), \tag{2.2}$$

$$E' = n - 1. \tag{2.3}$$

We incorporate the condition (1.4), which in the travelling wave variable reads

$$\int_{\mathbb{R}} (E(\xi) - E_\infty)\, d\xi = U, \tag{2.4}$$

where U is a given constant, and E_∞ will be determined as part of the solution of (2.1)–(2.4). The main result of this section is the following theorem.

THEOREM 2.1. *For each $U > 0$ there exists a solution (n, E, E_∞) of* (2.1)–(2.4) *which is unique up to translation in ξ and satisfies $E_{sat} < E_\infty < 1$. The far-field value E_∞ of the field is a strictly decreasing function of U, satisfying*

$$\lim_{U\to 0} E_\infty(U) = 1 \quad \textit{and} \quad \lim_{U\to\infty} E_\infty(U) = E_{sat}\,. \tag{2.5}$$

Before we prove the theorem we recall the existence result of (2.1)–(2.3) for a given value of E_∞.

LEMMA 2.2. *For every $E_\infty \in (E_{sat}, 1)$, there exists a unique (up to translation in ξ) solution (n, E) of* (2.1)–(2.3) *that satisfies $E > E_\infty$. The total charge density $n - 1$ has one simple zero, to the left of which it is positive (and negative to the right).*

This lemma is just a reformulation of the existence result that appears in [15]. The proof uses the fact that (2.2), (2.3) is a conservative system and uses the first integral relation

$$n - \log n - 1 = \int_{E_\infty}^{E} (v(y) - v(E_\infty))\, dy\,. \tag{2.6}$$

Proof of Theorem 2.1. By Lemma 2.2 it is sufficient to prove that the relation between E_∞ and U is one-to-one. With the solution (n, E) of (2.1)–(2.3) for given $E_\infty \in (E_{sat}, 1)$, we define

$$\mathcal{U}(E_\infty) := \int_{\mathbb{R}} (E(\xi) - E_\infty) d\xi\,.$$

The derivative can be written as $\mathcal{U}' := \int_{\mathbb{R}}(\hat{E}(\xi) - 1)d\xi$, where we define $\hat{E} = dE/dE_\infty$ and $\hat{n} = dn/dE_\infty$. The latter satisfy the equations

$$\hat{E}' = \hat{n}\,, \qquad \frac{n-1}{n}\hat{n} = (v(E) - v(E_\infty))\hat{E} - v'(E_\infty)(E - E_\infty)\,,$$

by differentiating (2.3) and (2.6) with respect to E_∞. Let us, without loss of generality, fix the point where $n - 1$ changes sign at $\xi = 0$, i.e., $n(0) = 1$. The second equation above implies that

$$\hat{E}(0) = v'(E_\infty)\frac{E(0) - E_\infty}{v(E(0)) - v(E_\infty)}\,.$$

The properties of v, $E_\infty < 1$, and $E > E_\infty$ imply that $\hat{E}(0) < 1$. Away from $\xi = 0$, $\hat{E}$ solves

$$\begin{aligned}\hat{E}' = {} & \frac{n}{n-1}[v(E) - v(E_\infty)](\hat{E} - 1) \\ & + \frac{n}{n-1}[v(E) - v(E_\infty) - v'(E_\infty)(E - E_\infty)]\,.\end{aligned}$$

The term in the second line is negative for large negative ξ and positive for large positive ξ. This implies $\hat{E} < 1$ for large $|\xi|$. Extrema of $\hat{E}$ away from $\xi = 0$ satisfy $\hat{E} = v'(E_\infty)\frac{E-E_\infty}{v(E)-v(E_\infty)} < 1$ analogously to the above. This shows that $\hat{E}(\xi) < 1$ for all ξ and, thus, $\mathcal{U}'(E_\infty) < 0$.

The assertion (2.5) then also follows since the amplitude of the wave tends to zero for $E_\infty \to 1$ and to infinity for $E_\infty \to E_{sat}$. ▯

3. Existence. In this section existence of solutions of (1.15), (1.16), (1.18) will be proven for given bounded $U(t) \in C_B^1(\mathbb{R}_+)$ and for initial data e_I satisfying

$$e_I \in L_x^1(\mathbb{R}) \cap L_x^\infty(\mathbb{R}), \quad e_I(x) > 0 \text{ a.e. in } x\,. \tag{3.1}$$

Clearly $U(t)$ is fixed by $U(0) = \int_{\mathbb{R}} e_I(x)\,dx > 0$ and by $U'(t)$ appearing in (1.18).

Assumption 2. There are positive constants δ and K, such that

$$0 < \delta \le U(t) \le K \quad \text{and} \quad \|e_I\|_\infty \le K\,,$$

where $\|\cdot\|_p$ denotes the norm in $L_x^p(\mathbb{R})$.

The derivative $U'(t)$ will have to be small enough as specified below. We start by the derivation of an a priori estimate.

PROPOSITION 3.1. *For solutions of* (1.15), (1.16), (1.18), $\|e(t,\cdot)\|_\infty \le C(\sigma_1)K$ *with* $C(\sigma_1) = \sqrt{2}\,\max\{2, c\sqrt{\sigma_1}\}$ *holds for all* $t \ge 0$.

Proof. The proof follows the idea of a similar result in [7]. Multiplying (1.15) by e^{p-1} for $p \ge 2$ and integration gives the estimate

$$\frac{d}{dt}\int_{\mathbb{R}} e^p\,dx \le -4\frac{(p-1)}{p}\int_{\mathbb{R}}(\partial_x e^{p/2})^2\,dx + p\sigma_1\int_{\mathbb{R}} e^p\,dx\,. \tag{3.2}$$

We observe that, by interpolation,

$$\|e_I\|_p \le \|e_I\|_\infty^{(p-1)/p}\,\|e_I\|_1^{1/p} \le K\,.$$

Our aim is to derive a uniform-in-p and uniform-in-time estimate on $\|e(t,\cdot)\|_p$ for a sequence of p such that $p \to \infty$. We use the Nash inequality [10]

$$\|u\|_2^3 \le c\|u\|_1^2\|\partial_x u\|_2$$

in one space dimension with $u = e^{p/2}$; thus, with the notation $z_p(t) = \|e(t,\cdot)\|_p^p$,

$$\frac{dz_p}{dt} \le p\sigma_1 z_p\left(1 - \frac{\tilde{c}(p-1)}{p^2}\frac{z_p^2}{z_{p/2}^4}\right)\,, \tag{3.3}$$

where $\tilde{c} = 4/(c^2\sigma_1)$. Starting with $z_1(t) = U(t) \le K$, the above inequality can be used recursively for obtaining bounds M_k for $z_{2^k}(t)$. Suppose $z_{2^{k-1}}(t) \le M_{k-1}$; then

$$z_{2^k}(t) \le M_k = \max\left\{K^{2^k}, \frac{2^k}{\sqrt{\tilde{c}(2^k-1)}}M_{k-1}^2\right\}\,.$$

Let us now examine the sequence M_k, defined by the recursion and by $M_0 = K$. Since, obviously, $M_{k-1} \ge K^{2^{k-1}}$ and $2^k/\sqrt{2^k-1} \ge 1$,

$$K^{2^k} \le \frac{2^k}{\sqrt{2^k-1}}M_{k-1}^2$$

holds. Thus, we make the upper bound M_k larger by the new definition

$$M_k = B2^{(k+1)/2}M_{k-1}^2\,, \quad M_0 = K\,, \quad B := \max\{1, \tilde{c}^{-1/2}\}\,,$$

where we have used $2^k/\sqrt{2^k-1} \le 2^{(k+1)/2}$. This recursion can be solved explicitly:

$$M_k = (\sqrt{2}\,B)^{a_k}2^{b_k/2}K^{2^k}\,,$$

where $a_k = \sum_{n=0}^{k-1} 2^n = 2^k - 1 < 2^k$ and $b_k = \sum_{n=0}^{k-1}(k-n)2^n = 2^{k+1} - 2 - k < 2^{k+1}$. Thus, since $B \geq 1$,

$$M_k \leq (2\sqrt{2}\, BK)^{2^k} \,,$$

and hence

$$\|e(t,\cdot)\|_{2^k} \leq \sqrt{2}\, K \, \max\{2, c\sqrt{\sigma_1}\} \quad \text{for all } k\,.$$

The proof is completed by passing to the limit $k \to \infty$. $\square$

Now we prepare a decoupled solution approach and examine (1.18) as an equation for $E_\infty(t)$.

Proposition 3.2. *Let the function $e \in L_x^1(\mathbb{R}) \cap L_x^\infty(\mathbb{R})$ satisfy $\|e\|_1 \geq \gamma > 0$ and $\|e\|_\infty \leq M$. Then the function $F(E;e) := \int_{\mathbb{R}}(v(E) - v(E+e(x)))dx$ is strictly increasing on $(0,\bar{E})$ with*

$$\bar{E}(\gamma, M) = 1 - \frac{v'(1+M)\gamma}{2M^2\sigma_3} > 1\,.$$

Furthermore,

$$F(0;e) \leq -v(M)\frac{\gamma}{M}\,, \quad F(\bar{E};e) \geq \frac{3v'(1+M)^2\gamma^2}{8M^3\sigma_3}\,,$$
$$F'(E;e) \geq -\frac{v'(1+M)\gamma}{2M} \quad \text{for } 0 \leq E \leq \bar{E}\,.$$

Proof. By the convexity of v' on $(0,E_i)$ and by the fact that v' is increasing and negative on (E_i,∞), the secant between E and $E+M$ lies above the graph of v' for $E \leq 1$. Therefore

$$F'(E) \geq \int_{\mathbb{R}} \left(v'(E) - v'(E)\left(1 - \frac{e}{M}\right) - v'(E+M)\frac{e}{M}\right)dx$$
$$= \int_{\mathbb{R}} (v'(E) - v'(E+M))\frac{e}{M}dx \geq (v'(E) - v'(E+M))\frac{\gamma}{M}$$

for $0 \leq E \leq 1$. Again by the same properties of v', the right-hand side takes its minimum value for $E = 1$, so $F'(E) \geq -v'(1+M)\gamma/M$ for $0 \leq E \leq 1$.

Since

$$|F''(E)| \leq \int_{\mathbb{R}} |v''(E) - v''(E+e)|dx \leq \sigma_3 M\,,$$

the derivative of F for $E > 1$ can be estimated by

$$F'(E) \geq -v'(1+M)\frac{\gamma}{M} - (E-1)\sigma_3 M\,,$$

proving that F is increasing on $(0,\bar{E})$ and the lower bound on F' in the statement of the proposition. The lower bound for $F(\bar{E})$ is obtained by integrating the above inequality from $E = 1$ to $E = \bar{E}$ and using that $F(1) > 0$, which holds, obviously, since v has its maximum at $E = 1$.

For estimating $F(0) = -\int_{\mathbb{R}} v(e)dx$, we use the L^∞-bound on e and the fact that secants between the origin and other points on the graph of v lie below the graph by the properties of v:

$$F(0) \leq -\int_{\mathbb{R}} v(M)\frac{e}{M}dx \leq -v(M)\frac{\gamma}{M}\,,$$

where the second inequality is due to the lower bound on the L^1-norm of e. $\square$

On the other hand, we consider the problem for e with given E_∞. In this case, the integral of e will not necessarily be equal to $U(t)$, which was the basis of the proof of Proposition 3.1. As a consequence, the estimates below are not uniform in time.

PROPOSITION 3.3. *Let $E_\infty(t)$ be given. Then the problem* (1.15), (1.16) *for e has a unique positive solution satisfying*

$$\int_{\mathbb{R}} e(t,x)dx \geq U(0)e^{-t\sigma_1} \quad \text{and} \quad e(t,x) \leq K e^{t\sigma_1}\,, \quad x \in \mathbb{R},\ t > 0\,.$$

Proof. Existence and uniqueness are standard results for semilinear parabolic equations. Positivity is a consequence of the maximum principle. The first estimate follows easily from integration of (1.15). The upper bound in the second estimate is a supersolution. □

We are now ready to formulate the main existence result.

THEOREM 3.4. *Let $M = C(\sigma_1)K$ denote the bound from Proposition* 3.1 *and let*

$$-v(M)\frac{\delta}{M} < U'(t) < \frac{3v'(1+M)^2\delta^2}{8M^3\sigma_3}\,, \quad t \geq 0\,.$$

Then the problem (1.15)–(1.18) *has a unique global solution satisfying $0 < E_\infty(t) < \bar{E}(\delta, M)$ and $0 < e(t,x) \leq M$.*

Remark 3.5. It seems unsatisfactory that the bounds on $U(t)$ (in Assumption 2) and on its derivative (in the formulation of the theorem) are required. However, examples of nonexistence of a solution for data violating such bounds are easily constructed. The range of the function $F(E_\infty, e(t,\cdot))$ (the right-hand side of (1.18)) as a function of E_∞ is a subset of $(-\sigma_1 U(t), \sigma_1 U(t))$. Therefore it is a necessary condition for the existence of a solution that $U'(t)$ lies in this interval for all t. The more restrictive bounds of the theorem guarantee stable (unique) solvability. For an example of nonexistence see the following section.

Proof. The first step is the construction of a local solution by a fixed point iteration on E_∞ acting on the set $\mathcal{E} := \{E(t) \in L_t^\infty((0,T)) :\ 0 \leq E(t) \leq \bar{E}\}$ with $T > 0$. For a given $E \in \mathcal{E}$, we first solve the problem (1.15), (1.16) with E_∞ replaced by E. By Proposition 3.3, this problem has a unique solution $e[E]$ satisfying

$$Ke^{t\sigma_1} \geq \int_{\mathbb{R}} e[E](t,x)dx \geq U(0)e^{-t\sigma_1} \geq \delta e^{-T\sigma_1} =: \gamma_T$$

and

$$e[E](t,x) \leq Ke^{t\sigma_1} \leq Me^{T\sigma_1} =: M_T$$

for $0 \leq t \leq T$. With Proposition 3.2, the range of $F(\cdot\,; e[E])$ includes the interval $(-v(M_T)\frac{\gamma_T}{M_T}, \frac{3v'(1+M_T)^2\gamma_T^2}{8M_T^3\sigma_3})$. For T small enough, this in turn includes the range of $U'(t)$ as given in the formulation of the theorem. Therefore the equation $F(\hat{E}; e[E]) = U'$ has a unique solution $\hat{E} = \mathcal{F}(E) \in [0, \bar{E}]$ which completes the definition of the fixed point operator $\mathcal{F} :\ \mathcal{E} \to \mathcal{E}$.

We shall prove that, for T small enough, $\mathcal{F}$ is a contraction and start with the mild formulation of (1.15), (1.16):

$$\begin{aligned} e(t,\cdot) = G(t,\cdot) * e_I &+ \int_0^t \partial_x G(t-s,\cdot) * [V(E(s)+e(s,\cdot)) - V(E(s))]ds \\ &+ \int_0^t G(t-s,\cdot) * [v(E(s)) - v(E(s)+e(s,\cdot))]ds\,, \end{aligned}$$

where $G(t,x) = (4\pi t)^{-1/2}e^{-x^2/(4t)}$ is the fundamental solution of the one-dimensional heat equation, $*$ denotes convolution with respect to x, and V is a primitive of v. For estimating the difference between $e_1 = e[E_1]$ and $e_2 = e[E_2]$, we start with

$$\begin{aligned}
&|v(E_1) - v(E_1+e_1) - v(E_2) + v(E_2+e_2)| \\
&\le \left|\int_{E_2}^{E_1} (v'(E) - v'(E+e_1))dE\right| + |v(E_2+e_2) - v(E_2+e_1)| \\
&\le e_1\,\sigma_2|E_1 - E_2| + \sigma_1|e_1 - e_2|\,,
\end{aligned}$$

and, analogously,

$$\begin{aligned}
&|V(E_1) - V(E_1+e_1) - V(E_2) - V(E_2+e_2)| \\
&\le e_1\sigma_1|E_1 - E_2| + \sigma_0|e_1 - e_2|\,.
\end{aligned}$$

We shall also use the properties

$$\int_{\mathbb{R}} G(t,x)\,dx = 1\,, \qquad \int |\partial_x G(t,x)|\,dx = \frac{1}{\sqrt{t\pi}} \quad \text{for all } t > 0$$

of the fundamental solution as well as the convolution inequality $\|f*g\|_1 \le \|f\|_1\|g\|_1$. A combination of these ingredients leads to an estimate of the form

$$\begin{aligned}
&\sup_{0<t<T} \|e_1(t,\cdot) - e_2(t,\cdot)\|_1 \\
&\le c\sqrt{T}\left(\sup_{0<t<T} \|e_1(t,\cdot) - e_2(t,\cdot)\|_1 + \sup_{0<t<T} |E_1(t) - E_2(t)|\right)
\end{aligned}$$

for $T \le 1$. It is an obvious consequence that the map $E \mapsto e[E]$ from $\mathcal{E}$ to $L_t^\infty((0,T), L_x^1(\mathbb{R}))$ is Lipschitz continuous with an arbitrarily small Lipschitz constant for small enough T.

Denoting $\hat{E}_1 = \mathcal{F}(E_1)$ and $\hat{E}_2 = \mathcal{F}(E_2)$, then $F(\hat{E}_i; e_i) = U'(t)$ holds for $i = 1, 2$. The difference of the two equations can be written as

$$F'(\tilde{E}; e_1)(\hat{E}_1 - \hat{E}_2) + \int_{\mathbb{R}} [v(\hat{E}_2 + e_2) - v(\hat{E}_2 + e_1)]dx = 0\,,$$

with $\tilde{E}$ between $\hat{E}_1$ and $\hat{E}_2$. This implies the estimate

$$\sup_{0<t<T} |\hat{E}_1(t) - \hat{E}_2(t)| \le -\frac{2M\sigma_1}{v'(1+M)\gamma} \sup_{0<t<T} \|e_1(t,\cdot) - e_2(t,\cdot)\|_1\,,$$

proving Lipschitz continuity also for the second step of the fixed point map. This concludes the proof of existence and uniqueness of a local solution.

Since solutions satisfy the uniform-in-time bounds $0 < e \le M$ and $\int_{\mathbb{R}} e\,dx \ge \delta$ and the above construction of local solutions works for initial conditions satisfying these bounds, the solution actually exists for all times, concluding the proof. □

4. Numerical results. In this section we present numerical experiments approximating (1.1)–(1.4) by solving the initial value problem for (1.15) subject to (1.18). In the time iteration we solve alternatively (1.18) and (1.15); for a given bounded positive initial condition e_I with finite mass we find the corresponding initial value of E_∞

by solving (1.18), this value is then used in (1.15) to get e in the next time step, and so on.

We discretize the equations on a domain $(0, L)$ and impose Neumann boundary conditions for (1.15). The scheme treats the second order term implicitly (backward Euler) and the first order term explicitly (forward Euler) in time. Also, the first order term is discretized in space by first order upwinding. For a given $U'(t)$ we approximate the integral (1.18) in the interval $[0, L]$ as a Riemann integral by using the trapezoidal rule. At each time step k a unique solution of the discretized equation

$$\int_0^L \{v(E_\infty^{k+1}) - v(E_\infty^{k+1} + e^k)\}dx - U'(t_k) = 0$$

is achieved by using the MATLAB implemented routine fzero, where the starting guess is E_∞^k.

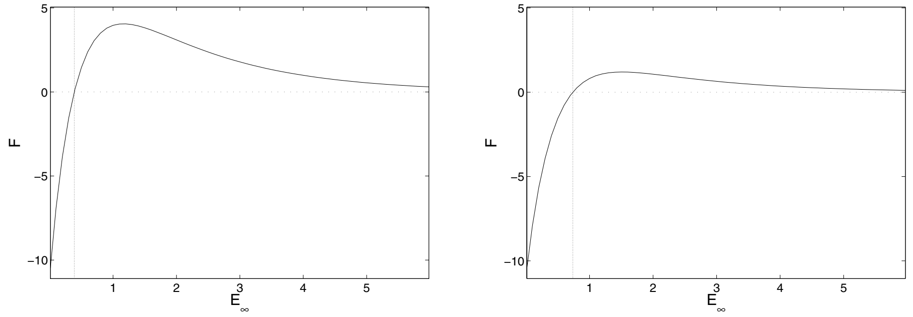
In all examples below we have taken $L = 200$, the spatial step $h = 0.1$, and the time step $\tau = 0.01$. As electron velocity function we use

$$v(E) = c\,e^{-a\,E} - d\,e^{-b\,E} + 1\,, \tag{4.1}$$

with

$$a = \ln(6)/3\,,\ b = 4\ln(6)/3\,,\ c = 2,\ \text{and } d = 3.$$

This v is normalized according to Assumption 1.



(a) $F(E_\infty, e_I)$ for e_I as in (4.2) with $l = 5$. (b) $F(E_\infty, e_I)$ for e_I as in (4.2) with $l = 1$.

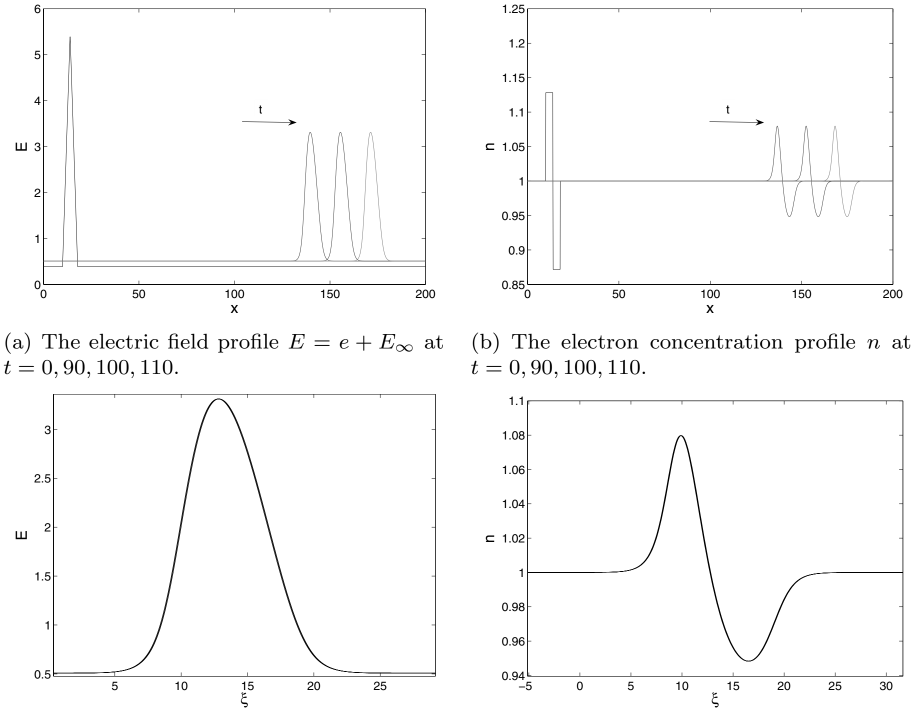
FIG. 2. *The function F computed for the initial data* (4.2).

As initial condition we take the piecewise linear function

$$e_I(x) = \begin{cases} 0 & \text{if } 0 \le x \le 10 \text{ or } x > 18, \\ \frac{l}{4}x - \frac{5}{2}l & \text{if } 10 < x \le 14, \\ -\frac{l}{4}x + \frac{9}{2}l & \text{if } 14 < x \le 18; \end{cases} \tag{4.2}$$

here l is the maximum of e_I giving the initial voltage $U(0) = 4l$. The function $E_\infty \to F(E, e_I)$ for e_I with $l = 1$ and $l = 5$, respectively, is plotted in Figure 2(a). Observe that the values at which F vanishes are, respectively, $E_\infty(0) \approx 0.77$ and $E_\infty(0) \approx 0.37$; i.e., the smaller the integral of e, the closer is E_∞ to 1, as expected for solitary waves (see Theorem 2.1). Since the speed of the solitary waves is given by $c = v(E_\infty)$, we expect the profiles to move to the right faster for smaller values of l. From now on we take $l = 5$ in (4.2); in this case $U(0) = 20$.

We start with examples for constant U. Figures 3(a) and 3(b) show, respectively, electric field and electron concentration profiles at $t = 0$ and $t = 90, 100, 110$. Figures 3(c) and 3(d) show the same solutions against the moving variable $\xi = x - ct$, where the speed $c = v(E_\infty(t))$ is evaluated at $t = 110$. The profiles at times $t = 90, 100, 110$ overlap in this frame, suggesting the stability of solitary waves.



(a) The electric field profile $E = e + E_\infty$ at $t = 0, 90, 100, 110$.

(b) The electron concentration profile n at $t = 0, 90, 100, 110$.

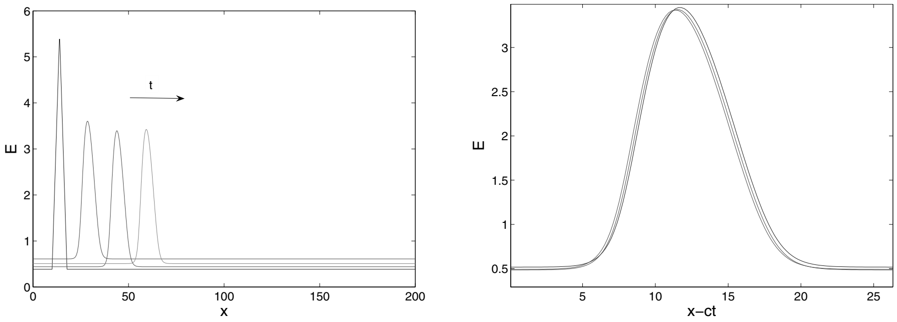
(c) The electric field profile $E = e + E_\infty$ at $t = 90, 100, 110$ against the travelling wave coordinate ξ.

(d) The electron concentration profile n at $t = 90, 100, 110$ against ξ.

FIG. 3. *Numerical solutions for constant U. Figures* 3(a) *and* 3(c) *show electric field values, and for completeness those corresponding to the electron concentration are shown to the right in Figures* 3(b) *and* 3(d)*. The wave speed used above has been computed by using the value of E_∞ at $t = 110$; here $E_\infty(110) \approx 0.5$ and $c \sim v(E_\infty) \approx 1.58$.*

For nonconstant U we first choose $U'(t) = 4\sin(4t)/(1 + t/10)$. Thus initially $U'(0) = 0$, and $U'(t)$ oscillates about this value, while the amplitude of the oscillations decays to 0 as $t \to \infty$. Figure 4(a) shows electric field profiles initially and at times $t = 10, 20, 30$. In Figure 4(b) electric field profiles are shown at times $t = 90, 80, 110$ against the variable $x - ct$. Since $U(t) \to const.$ as $t \to \infty$, we have taken $c = v(E_\infty(t))$ for $t = 110$, as before. The profiles now do not overlap precisely, but are fairly close to each other, again suggesting convergence to a solitary wave as $t \to \infty$ with wave speed $c = \lim_{t\to\infty} v(E_\infty(t))$.

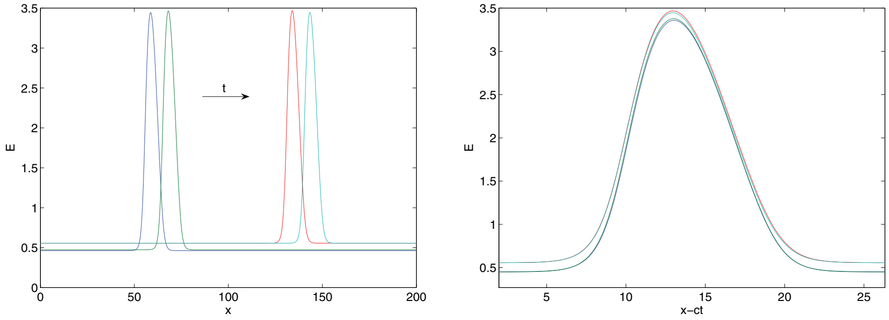
We now consider a t-periodic U, simply choosing $U'(t) = \sin(t)$. Although U does not approach a constant value as $t \to \infty$ and convergence to solitary waves is not expected, the solution profiles move to the right with an apparently constant speed. Figure 5(a) shows the solution profiles at $t = 31, 37$ (left) and at $t = 79, 85$ (right), i.e., profiles at, roughly, the beginning and the end of two time periods. The two



(a) Electric field profile at $t = 0, 10, 20, 30$.

(b) Electric field profile at $t = 90, 100, 110$ against $x - ct$ with $c = v(E_\infty(110)) \approx 1.55$.

FIG. 4. *Numerical solutions with* $U'(t) = 4\sin(4t)/(1 + t/10)$. *Only electric field profiles are shown. Figure* 4(a) *shows profiles at early time steps, where the amplitude of the oscillations of* $U'(t)$ *is appreciated. In Figure* 4(b) *late time steps are shown in the moving frame* ξ.



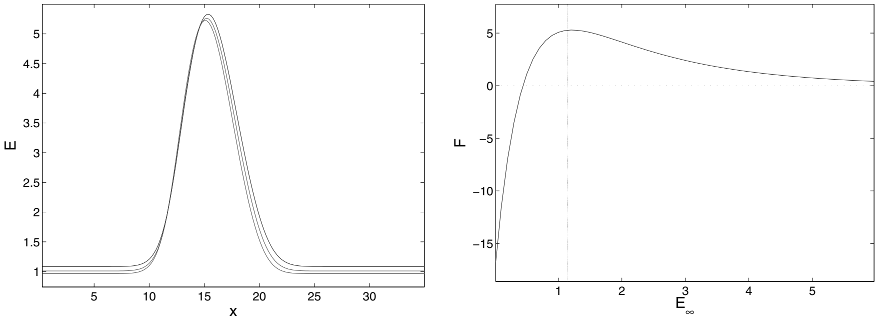
(a) Electric field profile at $t = 31, 37$ (left) and at $t = 79, 85$ (right).

(b) Electric field profile at against $x - ct$ with average speed $c \approx 1.57$.

FIG. 5. *Numerical solutions with* $U'(t) = \sin(t)$. *Only electric field profiles are shown. Figure* 5(a) *shows profiles at times* $t = 31, 37$ *(left) and at* $t = 79, 85$ *(right). Figure* 5(b) *shows the same profiles as Figure* 5(a) *against the coordinate* $x - ct$ *with the average speed* $c = \sum_{t_k=50}^{100} v(E_\infty(t_k))/5000 \approx 1.57$.

profiles to the left are almost a translation of each other, so are the two profiles on the right. This indicates that, as $t \to \infty$, a t-periodic "translating speed" is reached, presumably given by $c = v(E_\infty(t))$. To support this idea, we have computed the "averaged" speed of the solution at late time steps, including at least one period, namely $c = \sum_{t_k=50}^{100} v(E_\infty(t_k))/5000 \approx 1.57$. Figure 5(b) shows well-centered profiles against the moving coordinate with the average speed; these are at times $t = 51, 57$ and at $t = 79, 85$ (on top).

Finally, as an illustration of nonexistence we take $U'(t) = t^2 + 3.8$, so that initially U' is close to the maximum of F; see Figure 2(b). In this case the (numerical) solution ceases to exist at $t = 1.23$; i.e., $U'(1.22)$ exceeds the maximum of F. Electric field profiles for $t < 1.23$ are shown in Figure 6(a). The function F for e at $t = 1.2$ is shown in Figure 6(b). Observe that the maximum of F is approximately attained at $E_\infty = 1.1$ and that the solution (e, E_∞) has $E_\infty(1.2) \approx 1.084$.



(a) Electric field profile at $t = 1, 1.1, 1.2$.

(b) The function F for e at $t = 1.2$.

FIG. 6. *Numerical solutions for $U'(t) = t^2 + 3.8$ and the function F for e evaluated at $e(1.2, x)$. In this case the numerical solution ceases to exist at $t = 1.23$ when the value of $U'(t)$ exceeds the maximum of F.*

5. Small wave limit: Linearized stability. In this section we prove linearized stability of *small* solitary waves. We consider a small given constant voltage:

$$U = \varepsilon \ll 1\,.$$

We derive the limit $\varepsilon \to 0$ formally. From Theorem 2.1, solitary waves have $E_\infty \sim 1$ as $\varepsilon \to 0$, hence also $c \sim v(1)$ as $\varepsilon \to 0$. The amplitude of the waves is also small by (2.6). With this in mind we introduce the moving coordinate $\xi = x - v(1)t$ and the scaling

$$e = \varepsilon^2 e_1\,,\ E_\infty = 1 - \varepsilon^2 E_1\,,\ \tau = \varepsilon^2 t\,,\ \eta = \varepsilon\xi\,.$$

Then, in (1.15), after dividing by ε^4 and formally passing to the limit $\varepsilon \to 0$, we obtain

$$\partial_\tau e_1 = \partial_\eta^2 e_1 + \frac{v''(1)}{2}(2E_1 e_1 - e_1^2) \tag{5.1}$$

and, from (1.17) and (1.18),

$$\int_{\mathbb{R}} e_1\, d\eta = 1\,, \quad 2E_1 = \int_{\mathbb{R}} e_1^2\, d\eta\,. \tag{5.2}$$

As mentioned in the introduction, problem (5.1)–(5.2) is the conserved Fisher equation; see [11]. We now look at stability of stationary solutions to (5.1), since these are the limiting profiles of solitary waves as $\varepsilon \to 0$.

With the abbreviation

$$\kappa := -v''(1) > 0\,,$$

the family of stationary solutions is given explicitly by

$$\bar{e}(\eta) = \frac{\kappa}{48}\,\mathrm{sech}^2\left(\frac{\kappa}{24}(\eta + C)\right)\,, \quad \bar{E} = \frac{\kappa}{144}\,, \tag{5.3}$$

with the shift $C \in \mathbb{R}$.

We observe that rescaling with

$$\eta \to \kappa^{-1}\eta\,,\ e_1 \to \kappa e_1\,,\ E_1 \to \kappa E_1\,,\ \tau \to \kappa^{-2}\tau\,,$$

we can set $\kappa = 1$ in (5.1), with no changes in (5.2).

Denoting perturbations of e_1 and E_1 by u and A, respectively, the linearized problem (with $\kappa = 1$) reads

$$\partial_\tau u = \partial_\eta^2 u + (\bar{e} - \bar{E})u - \bar{e}A[u]\,, \tag{5.4}$$

$$\int_{\mathbb{R}} u\,d\eta = 0\,, \qquad A[u] = \int_{\mathbb{R}} \bar{e}\,u\,d\eta\,, \tag{5.5}$$

with

$$\bar{e}(\eta) = \frac{1}{48}\operatorname{sech}^2\left(\frac{\eta}{24}\right)\,, \quad \bar{E} = \frac{1}{144}\,, \tag{5.6}$$

where, without loss of generality, the shift has been set to zero. Note that there is a one-dimensional family of stationary solutions spanned by $u = \bar{e}'$, $A = 0$. This fact corresponds to the translation invariance of the nonlinear problem.

THEOREM 5.1. *The family of stationary solutions of* (5.4), (5.5) *is asymptotically stable: for an initial condition u_0 satisfying*

$$\int_{\mathbb{R}} u_0\,\bar{e}'\,d\eta = 0\,,$$

the solution of (5.4), (5.5) *subject to $u(\tau = 0) = u_0$ satisfies*

$$\|u(\tau,\cdot)\|_2 \le e^{\mu\tau}\|u_0\|_2 \quad \textit{with} \quad \mu \le -\frac{1}{192} < 0\,.$$

Proof. The linearized operator can be written as the sum of two self-adjoint operators on the space $L_0^2(\mathbb{R}) = \{u \in L^2(\mathbb{R}) :\ \int_{\mathbb{R}} u\,d\eta = 0\}$ equipped with the L^2-inner product $\langle\cdot,\cdot\rangle$:

$$\mathcal{L}u = \mathcal{L}_1 u + \mathcal{L}_2 u\,, \qquad \mathcal{L}_1 u = \partial_\eta^2 u + (\bar{e} - \bar{E})u\,,\ \mathcal{L}_2 u = -\bar{e}A[u]\,.$$

Obviously, $\mathcal{L}_2$ is nonpositive: $\langle \mathcal{L}_2 u, u\rangle = -A[u]^2 \le 0$.

The spectrum of $\mathcal{L}_1$ considered on all of $L^2(\mathbb{R})$ can be computed explicitly; see [6]: we obtain the essential spectrum $(-\infty, -\bar{E}]$ and the isolated eigenvalues

$$\lambda_1 = -\frac{3}{4}\bar{E} = -\frac{1}{192}\,, \quad \lambda_2 = 0\,, \quad \lambda_3 = \frac{5}{4}\bar{E}\,.$$

This can be obtained by using (5.6) and transforming the linear eigenvalue problem for $\mathcal{L}_1$ into a hypergeometric equation; see [3] for details. In the computation of λ_1 we also used (5.6).

The eigenfunction corresponding to λ_3 has $\int_{\mathbb{R}} u\,d\eta \ne 0$, since, according to the Sturm–Liouville theory (see, e.g., [2]), the eigenfunction corresponding to the largest eigenvalue does not change sign. This implies that in the restricted space $L_0^2(\mathbb{R})$ we actually have $\langle \mathcal{L}_1 u, u\rangle \le 0$.

Finally, $\bar{e}'$ is the eigenfunction corresponding to λ_2 and $\ker(\mathcal{L}_1) = \operatorname{span}\{\bar{e}'\}$. If P is the spectral projection onto $\ker(\mathcal{L}_1)$ then for $u \in L^2(\mathbb{R})$ satisfying (5.5), we have

$$\langle \mathcal{L}_1(I-P)u, (I-P)u\rangle \le \lambda_1 \|(I-P)u\|_2^2\,. \tag{5.7}$$

Since $\mathcal{L}_1$ is self-adjoint, P can be expressed as $Pu = \langle u, \bar{e}'\rangle\,\bar{e}'$.

By choosing the initial condition u_0 of (5.4), (5.5) such that $Pu_0 = 0$, it is easily checked that also $Pu = 0$ for all $t > 0$, which finishes the proof. □

REFERENCES

[1] L. L. Bonilla, F. J. Higuera, and S. Venakides, *The Gunn effect: Instability of the steady state and stability of the solitary wave in long extrinsic semiconductors*, SIAM J. Appl. Math., 54 (1994), pp. 1521–1541.

[2] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, Toronto, London, 1955.

[3] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J., 50 (2001), pp. 443–507.

[4] J. Gunn, *Microwave oscillations of current in* III–V *semiconductors*, Solid State Commun., 1 (1963), pp. 88–91.

[5] J. Gunn, *A topological theory of domain velocity in semiconductors*, IBM J. Res. Dev., 13 (1969), pp. 591–595.

[6] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1981.

[7] T. Hillen, K. Painter, and C. Schmeiser, *Global existence for chemotaxis with finite sampling radius*, Discrete Contin. Dyn. Syst. Ser. B, 7 (2007), pp. 125–144.

[8] J. Liang, *On a nonlinear integrodifferential drift-diffusion semiconductor model*, SIAM J. Math. Anal., 25 (1994), pp. 1375–1392.

[9] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer-Verlag, Vienna, 1990.

[10] J. Nash, *Continuity of solutions of parabolic and elliptic equations,* Amer. J. Math., 80 (1958), pp. 931–954.

[11] T. Newman, E. Kolomeisky, and J. Antonovics, *Population dynamics with global regulation: The conserved Fisher equation*, Phys. Rev. Lett., 92 (2004), 228103.

[12] D. H. Sattinger, *Stability of travelling waves of nonlinear parabolic equations*, in VII Internationale Konferenz über Nichtlineare Schwingungen (Berlin, 1975), Akademie-Verlag, Berlin, 1977, pp. 209–213.

[13] S. Sze, *Physics of Semiconductors Devices*, 2nd ed., Wiley, New York, 1981.

[14] P. Szmolyan, *A singular perturbation analysis of the transient semiconductor device equations*, SIAM J. Appl. Math., 49 (1989), pp. 1122–1135.

[15] P. Szmolyan, *Traveling waves in GaAs semiconductors*, Phys. D, 39 (1989), pp. 393–404.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1439–1463

ASYMMETRIC CHANNEL DIVIDER IN STOKES FLOW*

I. DAVID ABRAHAMS†, ANTHONY M. J. DAVIS‡, and
STEFAN G. LLEWELLYN SMITH‡

Abstract. This article examines the classic problem of Stokes flow into a divided channel with, in contrast to previous literature, the divider barrier asymmetrically placed with respect to the moving, parallel channel walls. The boundary value problem is reduced to a Wiener–Hopf equation that is of matrix form and of a class for which no exact solution is known. An explicit approximate solution, in general accurate to any specified degree, is obtained by a recent method which employs Padé approximants. Numerical results exhibit the flows due to moving walls or various combinations of downstream pressure gradients.

Key words. Stokes flow, channel flow, Wiener–Hopf technique, matrix Wiener–Hopf equations, Padé approximants

AMS subject classification. 78A45

DOI. 10.1137/070703211

1. Introduction.

A classic problem in two-dimensional creeping flow, having an analogy in plane elastostatics, is the disturbance created by the presence of a semi-infinite barrier in a channel flow (see Figure 1.1) driven by a pressure gradient and or shearing. The "parallel lines" geometry suggests the use of the Wiener–Hopf technique; however, the advantage of the constricting walls in creating unidirectional flows both upstream and downstream is offset by the appearance, in general, of a matrix Wiener–Hopf system. The exception is the case of symmetric geometry which yields Wiener–Hopf equations of standard (scalar) type [1], since then the flow components that are even and odd with respect to the centerline can be considered separately. Despite this simplification, the even problem (no flow across the line of the barrier), which is the case of greater interest, requires an intricate factorization constructed and used by Buchwald and Doran [2] and Foote and Buchwald [3]. An erroneous attempt was presented earlier by Graebel [4] with the aim of achieving better accuracy than the approximate, yet still complicated, solutions given by Koiter [5]. Richardson [6] neglected an important feature of the factorization in [2]. Jensen and Halpern [7] verified the calculations of Buchwald and coworkers in using their solution to examine the role of the stress singularity at the edge of surfactant between thin fluid layers. Without a general procedure for solving matrix Wiener–Hopf problems (see further discussion on this point is section 3.1), an alternative strategy for the biharmonic equation is to employ complex variable techniques, facilitated by the removal of one wall (that is, the receding of one channel wall to infinity). Approximations are still required using this approach, as presented by Moore, Buchwald, and Brewster [8] for a Stokesian entry problem, in which the remaining wall translates, and by Kim, Choi, and Jeong [9] for a model of the half-pitot tube, in which a shear flow is prevented from generating any flux into the channel. The following study is both an application

*Received by the editors September 19, 2007; accepted for publication (in revised form) January 14, 2008; published electronically May 2, 2008.
http://www.siam.org/journals/siap/68-5/70321.html

†Department of Mathematics, University of Manchester, Manchester M13 9PL, UK (i.d.abrahams@manchester.ac.uk).

‡Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering, UCSD, La Jolla, CA 92093-0411 (amdavis@ucsd.edu, sgls@ucsd.edu).

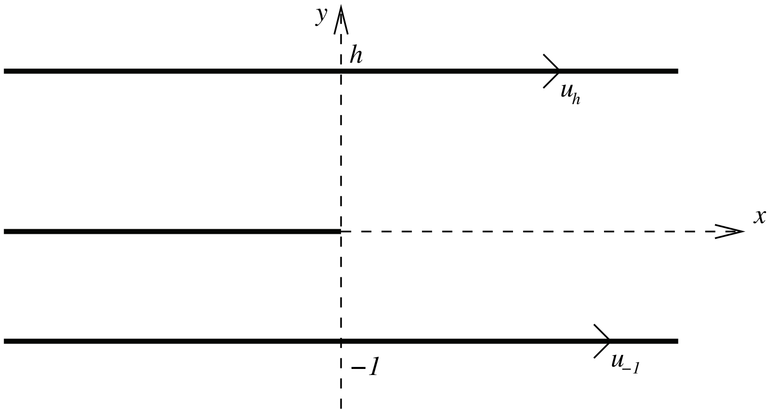


FIG. 1.1. *Geometry of problem.*

of a new Padé approximant procedure [10] and the first solution of this classic problem by the Wiener–Hopf technique.

Consider the unidirectional flow between rigid walls at $y = -1$, h, where (x, y) are Cartesian coordinates, at which the velocity u has the prescribed values u_{-1}, u_h, respectively (see Figure 1.1). The flow $u^\infty(y)\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is the unit vector in the x direction, is given by

$$u^\infty(y) = u_h\left(\frac{y+1}{h+1}\right) + u_{-1}\left(\frac{h-y}{h+1}\right) - \frac{G}{2\mu}(h-y)(y+1). \tag{1.1}$$

Here the first two terms of the velocity profiles may be identified as a shear flow with different wall speeds and the last term with a flow driven by a pressure gradient G that accounts for the prescribed flux being different from the flux generated by the shear flow. It is readily observed from (1.1) that only the weighted average of the wall velocities has a role in the study of the disturbance flow generated by the introduction of a fixed plate at $y = 0, x < 0$. With $u_h + u_{-1}h = -U(h+1)$, this occurs when the flow speed at $y = 0$, namely

$$u^\infty(0) = \frac{u_h + u_{-1}h}{h+1} - \frac{Gh}{2\mu} = -\left(U + \frac{Gh}{2\mu}\right), \tag{1.2}$$

is nonzero, or when there is a flux “mismatch” in $(-1, 0)$ between the upstream flow and that in the downstream channel.

In terms of pressure gradients G_-, G_+ at infinity (as shown in Figure 1.1), the downstream $(x \to -\infty)$ unidirectional velocity profiles are given by

$$u_-^\infty(y) = -u_{-1}y + \frac{G_-}{2\mu}y(y+1), \qquad -1 < y < 0, \tag{1.3}$$

$$u_+^\infty(y) = u_h\frac{y}{h} - \frac{G_+}{2\mu}y(h-y), \qquad 0 < y < h, \tag{1.4}$$

whose total flux must equal that in the upstream $(x \to \infty)$ unidirectional velocity profile (1.1). Thus

$$\frac{(G_+ - G)h^3 + (G_- - G)}{6\mu} = (h+1)\left(U + \frac{Gh}{2\mu}\right), \tag{1.5}$$

which is to be viewed as determining the upstream pressure gradient G in terms of G_+, G_-, U. The flux "mismatch," ΔQ, is now given by

$$\Delta Q = \int_{-1}^{0} \left[u^\infty(y) - u_-^\infty(y)\right] dy = -\frac{1}{2}\left(U + \frac{Gh}{2\mu}\right) + \frac{(G_- - G)}{12\mu}. \tag{1.6}$$

Evidently the sets of values of the wall velocities and pressure gradients in (1.1), (1.3) for which the presence of the semi-infinite barrier creates a disturbance flow form a two-parameter family described by nonzero values of the vector $[u^\infty(0), \Delta Q]$, with only its direction being significant. Thus any two flows of type (1.1), (1.3) that yield parallel values of this vector, determined by (1.2), (1.6), may be regarded as equivalent because their suitably weighted difference must be a unidirectional flow with zero velocity at $y = 0$.

For example, the two flows determined by $G_- = 0 = G$ and either $u_{-1} = 0, u_h = V^*$ or $u_{-1} = -V, u_h = 0$ both yield values of $[u^\infty(0), \Delta Q]$ that are parallel to $(2, 1)$ because, if $V^* = -Vh$, they differ by the shear flow $u = Vy$. The former is the finite version of the two-dimensional model of a half-pitot tube studied by Kim, Choi, and Jeong [9], whose motivation was the experimental work reported by Stanton, Marshall, and Bryant [11] and Taylor [12]. The latter is the finite version of a Stokesian entry problem, with no pressure gradient far down the semi-infinite channel, studied by Moore, Buchwald, and Brewster [8]. The condition of no pressure gradient upstream ensures, for any u_h, that their flow is recovered in the limit $h \to \infty$.

In the case of symmetric geometry, $h = 1$ and evidently (1.2)–(1.6) show that flows with $u_{-1} = -U = u_h$, $G_+ = G_-$ are equivalent to the even case:

$$\Delta Q = 0, \quad u_-^\infty(-y) = u_+^\infty(y), \quad 0 < y < 1, \tag{1.7}$$

which consists downstream of a shear and pressure-driven flow combination, while flows with $u_{-1} = 0 = u_h$, $G_+ = -G_-$ are equivalent to the odd case:

$$G = 0, \quad u^\infty(0) = 0, \quad \Delta Q = \frac{G_-}{12\mu}, \quad u_-^\infty(-y) = -u_+^\infty(y), \quad 0 < y < 1, \tag{1.8}$$

which is a pressure-driven flow out of one channel into the other.

In view of the above discussion, a generic study which covers *all possible* flow cases in fact need only consider forcing due solely to the moving walls and various combinations of downstream pressure gradients. The two cases are therefore the following:

1. $U \neq 0$ and $G_+ = 0 = G_-$. Therefore, $[u^\infty(0), \Delta Q]$ is parallel to $[2(h^2 - h + 1), h(h-1)]$, so its direction depends on h only.
2. $U = 0$ and various flux ratios. Hence

$$\frac{\Delta Q}{u^\infty(0)} = \frac{3h + 1 - G_-/G}{6h}, \tag{1.9}$$

which displays a two-parameter dependence.

As an illustration of the use of MATLAB, a computational solution of this channel flow, using approximate boundary conditions, was given by Fehribach and Davis [13].

2. The Wiener–Hopf problem.

The equations of steady creeping flow, the Stokes equations [14], are

$$\nabla p = \mu \nabla^2 \mathbf{v}, \qquad \nabla \cdot \mathbf{v} = 0, \tag{2.1}$$

where $\mathbf{v}$ is the velocity, p the dynamic pressure, and μ the viscosity. For two-dimensional flow referred to Cartesian coordinates (x, y), equations (2.1) allow a stream function $\psi(x, y)$ to be introduced such that

$$\mathbf{v} = \frac{\partial \psi}{\partial y}\hat{\mathbf{x}} - \frac{\partial \psi}{\partial x}\hat{\mathbf{y}}, \qquad \nabla^4 \psi = 0. \tag{2.2}$$

Consider the flow between rigid walls at $y = -1$, h and a semi-infinite fixed barrier at $y = 0$, $x < 0$ at which the stream function has distinct constant values and its y-derivative has the prescribed values $u_{-1} = -U$, $u_h = -U$, 0, respectively (see Figure 1.1). Then $\mathbf{v} \sim u^{\infty}(y)\hat{\mathbf{x}}$ as $x \to \infty$ and $\mathbf{v} \sim u^{\infty}_{\pm}(y)\hat{\mathbf{x}}$ as $x \to -\infty$, where $u^{\infty}(y)$ and $u^{\infty}_{\pm}(y)$ are given by (1.1) and (1.3), (1.4), with the upper "plus" (lower "minus") sign referring to the upper (lower) duct region. It is advantageous to choose for the disturbance field not $\mathbf{v} - u^{\infty}(y)\hat{\mathbf{x}}$ but rather $\mathbf{v} - u^{\infty}_{\pm}(y)\hat{\mathbf{x}}$. Thus, on setting, as in (2.2),

$$\mathbf{v} - u^{\infty}_{\pm}(y)\hat{\mathbf{x}} = \frac{\partial \bar{\psi}}{\partial y}\hat{\mathbf{x}} - \frac{\partial \bar{\psi}}{\partial x}\hat{\mathbf{y}}, \qquad \begin{cases} 0 < y < h, \\ -1 < y < 0, \end{cases} \tag{2.3}$$

the disturbance stream function $\bar{\psi}(x, y)$ is biharmonic, satisfies the homogeneous conditions

$$\bar{\psi} = 0 = \frac{\partial \bar{\psi}}{\partial y} \text{ at } y = -1, h, \quad -\infty < x < \infty, \quad \text{ and } y = 0, \quad x < 0, \tag{2.4}$$

is continuous along with its y-derivative on $y = 0$, $x > 0$, and, according to (1.3), (1.4), and (2.3), is generated by the discontinuities

$$\left[\frac{\partial^2 \bar{\psi}}{\partial y^2}\right]_{0-}^{0+} = U\frac{h+1}{h} + \frac{G_+ h + G_-}{2\mu}, \qquad \left[\frac{\partial^3 \bar{\psi}}{\partial y^3}\right]_{0-}^{0+} = -\frac{G_+ - G_-}{\mu} \tag{2.5}$$

on $y = 0, x > 0$. If, for convenience, these discontinuities tend to zero as $x \to \infty$, i.e., (2.5) is modified to

$$\left[\frac{\partial^2 \bar{\psi}}{\partial y^2}\right]_{0-}^{0+} = \left[U\frac{h+1}{h} + \frac{G_+ h + G_-}{2\mu}\right] e^{-\epsilon x}, \qquad \left[\frac{\partial^3 \bar{\psi}}{\partial y^3}\right]_{0-}^{0+} = -\frac{G_+ - G_-}{\mu} e^{-\epsilon x} \tag{2.6}$$

on $y = 0, x > 0$, where ϵ is a small positive real constant, then the disturbance stream function $\bar{\psi}$ and its derivatives tend to zero as $x \to \pm\infty$ as a consequence of the choice (2.3), which further implies that the unknown functions $s(x)$, $t(x)$, defined by

$$\left[\frac{\partial^2 \bar{\psi}}{\partial y^2}\right]_{0-}^{0+} = s(x), \quad \left[\frac{\partial^3 \bar{\psi}}{\partial y^3}\right]_{0-}^{0+} = t(x), \quad y = 0, \quad x < 0, \tag{2.7}$$

also decay to zero as $x \to -\infty$. On completion of the solution procedure ϵ will be set to zero.

In terms of the Fourier transform

$$\Psi(k, y) = \int_{-\infty}^{\infty} \bar{\psi}(x, y) e^{ikx} dx, \tag{2.8}$$

the boundary conditions (2.4), (2.6), (2.7) yield

$$\Psi(k, -1) = 0 = \Psi_y(k, -1), \qquad \Psi(k, h) = 0 = \Psi_y(k, h), \tag{2.9}$$

$$\Psi(k,0)=\int_0^\infty \bar\psi(x,0)e^{ikx}dx=\Psi^+(k,0), \qquad \Psi_y(k,0)=\Psi_y^+(k,0), \tag{2.10}$$

$$\begin{aligned}[\Psi_{yy}(k,y)]_{0-}^{0+}&=\int_{-\infty}^0 s(x)e^{ikx}dx+\left[U\frac{h+1}{h}+\frac{G_+h+G_-}{2\mu}\right]\int_0^\infty e^{ix(k+i\epsilon)}dx\\&=S^-(k)+\left[U\frac{h+1}{h}+\frac{G_+h+G_-}{2\mu}\right]\frac{i}{k+i\epsilon},\end{aligned} \tag{2.11}$$

$$\begin{aligned}[\Psi_{yyy}(k,y)]_{0-}^{0+}&=\int_{-\infty}^0 t(x)e^{ikx}dx-\frac{G_+-G_-}{\mu}\int_0^\infty e^{ix(k+i\epsilon)}dx\\&=T^-(k)-\frac{G_+-G_-}{\mu}\frac{i}{k+i\epsilon}.\end{aligned} \tag{2.12}$$

Convergence of the above Fourier full- and half-range transforms is ensured if k lies in an infinite strip containing the real line, here and henceforth referred to as $\mathcal{D}$, with its width limited from below by the singularity at $k=-i\epsilon$. Evidently (see [1]) the unknown pairs of (half-range transform) functions $\Psi^+(k,0)$, $\Psi_y^+(k,0)$ and $S^-(k)$, $T^-(k)$ are regular in the region above and including $\mathcal{D}$, denoted $\mathcal{D}^+$, and the region below and including $\mathcal{D}$, denoted $\mathcal{D}^-$, respectively. Thus, $\mathcal{D}^+\cap\mathcal{D}^-\equiv\mathcal{D}$.

In view of the behavior of $\bar\psi$ at $x=\pm\infty$, the Fourier transform (2.8) can be applied to the biharmonic equation, whence

$$\left(\frac{d^2}{dy^2}-k^2\right)^2\Psi=0 \tag{2.13}$$

and hence a general solution which satisfies (2.9) is

$$\begin{aligned}\Psi(k,y)&=A(k)k(1+y)\sinh[k(1+y)]\\&\quad+B(k)\{k(1+y)\cosh[k(1+y)]-\sinh[k(1+y)]\}, \qquad -1<y<0,\end{aligned} \tag{2.14}$$

$$\begin{aligned}\Psi(k,y)&=C(k)k(h-y)\sinh[k(h-y)]\\&\quad+D(k)\{k(h-y)\cosh[k(h-y)]-\sinh[k(h-y)]\}, \qquad 0<y<h.\end{aligned} \tag{2.15}$$

Application of the conditions (2.10) now yields

$$\begin{aligned}\begin{pmatrix}A(k)\\B(k)\end{pmatrix}&=\frac{1}{\sinh^2k-k^2}\\&\times\begin{pmatrix}k\sinh k & -(k\cosh k-\sinh k)\\ -(k\cosh k+\sinh k) & k\sinh k\end{pmatrix}\begin{pmatrix}\Psi^+(k,0)\\k^{-1}\Psi_y^+(k,0)\end{pmatrix},\end{aligned} \tag{2.16}$$

$$\begin{aligned}\begin{pmatrix}C(k)\\D(k)\end{pmatrix}&=\frac{1}{\sinh^2kh-k^2h^2}\\&\times\begin{pmatrix}kh\sinh kh & kh\cosh kh-\sinh kh\\ -(kh\cosh kh+\sinh kh) & -kh\sinh kh\end{pmatrix}\begin{pmatrix}\Psi^+(k,0)\\k^{-1}\Psi_y^+(k,0)\end{pmatrix}.\end{aligned} \tag{2.17}$$

As $\Psi^+(k,y)$ and $\Psi_y^+(k,y)$ are continuous across the line $y=0$, the discontinuities (2.11), (2.12) may be regarded as conditions on $\Psi_{yy}-k^2\Psi$ and its derivative, which facilitates the deduction of the following *matrix* Wiener–Hopf equation:

$$\begin{pmatrix}T^-(k)\\-S^-(k)\end{pmatrix}-\frac{i}{k+i\epsilon}\left(U\frac{h+1}{h}\begin{bmatrix}0\\1\end{bmatrix}+\frac{1}{2\mu}\begin{bmatrix}2G_+-2G_-\\G_+h+G_-\end{bmatrix}\right)=\mathbf{K}(k)\begin{pmatrix}\Psi^+(k,0)\\\Psi_y^+(k,0)\end{pmatrix}, \tag{2.18}$$

where

$$\mathbf{K}(k) = \begin{pmatrix} k^2[f(k)+g(k)] & -ke(k) \\ -ke(k) & g(k)-f(k) \end{pmatrix}, \tag{2.19}$$

$$e(k) = 2k\left(\frac{k^2}{\sinh^2 k - k^2} - \frac{k^2h^2}{\sinh^2 kh - k^2h^2}\right), \tag{2.20}$$

$$f(k) = 2k\left(\frac{k}{\sinh^2 k - k^2} + \frac{kh}{\sinh^2 kh - k^2h^2}\right), \tag{2.21}$$

$$g(k) = k\left[\frac{\sinh 2k}{\sinh^2 k - k^2} + \frac{\sinh 2kh}{\sinh^2 kh - k^2h^2}\right]. \tag{2.22}$$

It can easily be seen that $\mathbf{K}(k)$ possesses the properties

$$\mathbf{K}(k) = \mathbf{K}(-k) = [\mathbf{K}(k)]^T, \tag{2.23}$$

where T denotes the transpose, a fact that is exploited subsequently. The determinant of the kernel is

$$|\mathbf{K}(k)| = \frac{4k^4[\sinh^2 k(h+1) - k^2(h+1)^2]}{(\sinh^2 kh - k^2h^2)(\sinh^2 k - k^2)}. \tag{2.24}$$

The forcing term in (2.18) displays the two independent flows identified above. In the case of symmetric geometry, $h = 1$ implies that $e(k)$ is identically zero, and hence the Wiener–Hopf equation (2.18) separates into disjoint scalar equations of standard type. Then, in the even case, $G_+ \equiv G_-$ implies that T^- and Ψ^+ vanish, while, in the odd case, $U \equiv 0 \equiv G_+ + G_-$ implies that S^- and Ψ_y^+ vanish, as expected.

It remains to consider the pressure singularity at the barrier edge. In the neighborhood of $r = 0$, using the obvious polar coordinate representation,

$$\bar{\psi} \sim 2^{1/2} r^{3/2} \cos\frac{1}{2}\theta[\Lambda_1 \sin\theta + \Lambda_2(1+\cos\theta)] \tag{2.25}$$

$$= \Lambda_1(r+x)^{1/2}y + \Lambda_2(r+x)^{3/2}, \tag{2.26}$$

after rejecting the more singular terms of order $r^{1/2}$. Thus

$$\bar{\psi}(x,0) \sim \Lambda_2(2x)^{3/2}, \qquad \frac{\partial\bar{\psi}}{\partial y}(x,0) \sim \Lambda_1(2x)^{1/2} \text{ as } x \to 0+, \tag{2.27}$$

and by writing, for $x < 0$,

$$\bar{\psi} \sim \frac{\Lambda_1 y|y|}{(r-x)^{1/2}} + \frac{\Lambda_2|y|^3}{(r-x)^{3/2}}; \tag{2.28}$$

it follows that

$$\frac{\partial^2\bar{\psi}}{\partial y^2}(x,0) \sim \pm\Lambda_1\left(\frac{2}{-x}\right)^{1/2}, \qquad \frac{\partial^3\bar{\psi}}{\partial y^3}(x,0) \sim \pm\Lambda_2\frac{6}{(-2x)^{3/2}} \tag{2.29}$$

as $x \to 0-$ on the upper/lower side of the barrier. The latter result indicates that the pressure jump across the barrier behaves as $6\mu\Lambda_2(2/(-x))^{1/2}$ as the edge is approached. This agrees with the asymptotic form

$$\mu^{-1}\bar{p} \sim 2^{1/2} r^{-1/2} \left(-\Lambda_1 \cos \frac{1}{2}\theta + 3\Lambda_2 \sin \frac{1}{2}\theta \right), \tag{2.30}$$

obtained from (2.25) by noting that (2.1) and (2.2) ensure that $\bar{p}$ and $\mu\nabla^2\bar{\psi}$ are conjugate functions. The rejection of order $r^{1/2}$ terms in (2.25) thus minimizes the *order* of this edge singularity in the pressure, as, for example, in the calculations for the spherical cap [15] and the hollow sphere with caps removed [16]. A bounded pressure jump occurs if $\Lambda_2 = 0$, which may be achieved by a suitable choice of the direction of the forcing vector in (2.18). Such a procedure is unnecessary for the geometrically symmetric even case since then Λ_2 must be zero for $\bar{\psi}$ to be an odd function of θ in (2.25). The Wiener–Hopf calculation [2, 3, 5] generates an entire function that is identically zero, from which it is deduced that $\Lambda_1 = 2/\sqrt{\pi}$.

3. Factorization of the duct kernel.

3.1. Introduction and overview of the factorization procedure. In the previous section the matrix Wiener–Hopf equation was derived, in which the kernel, $\mathbf{K}(k)$, is written in (2.19). The aim of this section is to factorize $\mathbf{K}(k)$ into a product of two matrices

$$\mathbf{K}(k) = \mathbf{K}^{-}(k)\mathbf{K}^{+}(k), \tag{3.1}$$

one containing those singularities of $\mathbf{K}(k)$ lying in the lower half-plane, referred to as $\mathbf{K}^{+}(k)$, and $\mathbf{K}^{-}(k)$, which is analytic in the lower half-plane $\mathcal{D}^{-}$ and hence contains the singularities of $\mathbf{K}(k)$ lying above the strip $\mathcal{D}$. Note that $[\mathbf{K}^{+}(k)]^{-1}$ and $[\mathbf{K}^{-}(k)]^{-1}$ are also analytic in the regions $\mathcal{D}^{+}$ and $\mathcal{D}^{-}$, respectively. Further, it is necessary for successful completion of the Wiener–Hopf procedure that $\mathbf{K}^{\pm}(k)$ are at worst of algebraic growth (see Noble [1]). Unfortunately, although matrix kernel factorization with the requisite growth behavior has been proven to be possible for a wide class of kernels (Gohberg and Krein [17]), to which the kernel (2.19) belongs, no constructive method has been found to complete this in general. There are classes of matrices for which product factorization can be achieved explicitly, the most important of which are those amenable to Hurd's method [18] and Khrapkov–Daniele commutative matrices [19, 20]. Details of these, and an extensive bibliography on matrix kernel factorization, can be found in [21, 10, 22]. The present problem yields a kernel which, to the authors' knowledge, falls outside of the classes permitting an exact factorization, and so an approximate decomposition will be performed here. The approach follows that developed recently by one of the authors and has been successfully applied to problems in elasticity [21, 22] and acoustics [10]. Essentially, the procedure is to rearrange the kernel into an appropriate form, namely, to resemble a Khrapkov (commutative) matrix, and then to replace a scalar component of it by a function which approximates it accurately in the strip of analyticity $\mathcal{D}$. The new approximate kernel is able to be factorized exactly (into an explicit noncommutative decomposition), and, in the previous cases cited above, strong numerical evidence was offered for convergence of the resulting approximate factors to the exact ones as the scalar approximator is increased in accuracy. Further, the convergence to the solution has been validated for one particular matrix kernel [23], where an *exact* noncommutative factorization can be derived by an alternative procedure.

The kernel in (2.19) appears, on face value, significantly simpler to factorize than those in the previously mentioned articles [21, 10], because it contains only simple pole singularities rather than branch cuts. Therefore, it could perhaps be considered as more appropriately factorized by pole removal methods, such as those suggested by Idemen [24], Noble [1], Rawlins [25], Abrahams [26], and Abrahams and Wickham [27], reducing the problem down to an infinite algebraic system of equations which needs to be solved numerically. However, there are three reasons why this approach is not useful here. The first is a technical point; it can be shown that the procedure for removing singularities from the kernel, and thereby obtaining the kernel factors $\mathbf{K}^{\pm}(k)$, is not nearly as straightforward as for those kernels considered by the aforementioned authors. Second, the pole locations are complex and are found from the zeros of the determinant of the kernel $\mathbf{K}(k)$ in (2.24), i.e., the roots of the transcendental Papkovich–Fadle dispersion relation. This creates further complications in the factorization scheme. The third, and most compelling, reason for avoiding this approach is that we would like a final solution which will offer *uniformly* accurate results for all values of upper duct height h, from $h = 1$ to $h = \infty$, a range that does not imply any loss of generality. Clearly, as $h \to \infty$, more and more of the Papkovich–Fadle poles need to be included to maintain constant accuracy (more and more move down close to the strip $\mathcal{D}$), and so the corresponding algebraic system to solve has to be truncated after a greater and greater number of terms. Thus, we cannot expect to recover the $h \to \infty$ case, that is, when the upper duct top wall is removed, so we could not employ such a factorization in a solution which we would hope to compare with other results for this particular flow domain (Moore, Buchwald, and Brewster [8], Kim, Choi, and Jeong [9], etc.)

In view of the above arguments we aim to employ the Wiener–Hopf approximant matrix (WHAM) method [10] discussed previously and to do this in such a way that maintains the requisite accuracy over all values of $h \geq 1$. As mentioned in the introduction, when $h = 1$, then the kernel should reduce to two scalar functions, reflecting the symmetric and antisymmetric motions clearly evident to exist from the symmetry in duct geometry. When $h = \infty$, the upper duct wall is removed, and although others have tackled this by alternative approximate/numerical means, it can, in fact, be shown that the kernel actually reduces to a commutative (Khrapkov) form [23]. Therefore, an exact factorization is again achievable. Hence, if the approximate factorization, to be achieved here by the WHAM method, is organized so that in both limits, $h \to 1$ and $h \to \infty$, it reduces to the exact kernel decomposition, then very good accuracy can be expected for all intermediate h values. This is what will be done below. However, as a complication to this factorization, we must take account of two unfortunate features of the kernel. The first is the fact that the elements of $\mathbf{K}(k)$, and in particular $e(k)$, $f(k)$ shown in (2.20)–(2.22), differ by a factor of k near the origin. This is identical to that found for the kernel in [21] and can be handled by a suitable rearrangement of terms. The second is due to the fact that $\mathbf{K}(k)$ must be written as a product of three matrices such that the inner matrix $\mathbf{L}(k)$ (see (3.6) below) has a determinant with behavior proportional to k^{-2} as $k \to 0$. This is a removable singularity because it is pre- and postmultiplied by matrices which each have determinant k. Unfortunately, the inner matrix is the one we initially factorize, and so additional arrangement is necessary to take this small-k behavior into account. With these points of explanation in mind, the factorization procedure is now elucidated.

3.2. Conditioning of the matrix kernel.

3.2.1. Behavior for large $|k|$. The matrix $\mathbf{K}(k)$ is characterized by its elements $e(k)$, $f(k)$, $g(k)$, given in (2.20)–(2.22), and in particular by their behavior for large and small k. For large k it is easily deduced that

$$e(k) \sim 8k^3(e^{-2|k|} - h^2 e^{-2|k|h}), \quad |k| \to \infty, \quad k \in \mathcal{D}, \tag{3.2}$$

$$f(k) \sim 8k^2(e^{-2|k|} + h e^{-2|k|h}), \quad |k| \to \infty, \quad k \in \mathcal{D}, \tag{3.3}$$

$$g(k) \sim 4|k|, \quad |k| \to \infty, \quad k \in \mathcal{D}. \tag{3.4}$$

It is appropriate to arrange the kernel to be diagonally dominant as $k \to \infty$ in $\mathcal{D}$, and so simple algebra gives

$$\mathbf{K}(k) = \frac{1}{2}\begin{pmatrix} 0 & -k \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ i & i \end{pmatrix}\mathbf{L}(k)\begin{pmatrix} i & 1 \\ i & -1 \end{pmatrix}\begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.5}$$

where $\mathbf{L}(k)$ may be written in the form

$$\mathbf{L}(k) = g(k)\mathbf{I} + \begin{pmatrix} 0 & f(k) + ie(k) \\ f(k) - ie(k) & 0 \end{pmatrix}, \tag{3.6}$$

with $\mathbf{I}$ the identity.

3.2.2. Behavior of the kernel near the origin. Near the origin the scalar functions $e(k)$, $f(k)$, $g(k)$ take the form

$$e(k) \sim 6\left(\frac{h^2-1}{h^2}\right)\frac{1}{k}, \tag{3.7}$$

$$f(k) \sim 6\left(\frac{h^3+1}{h^3}\right)\frac{1}{k^2}, \tag{3.8}$$

$$g(k) \sim 6\left(\frac{h^3+1}{h^3}\right)\frac{1}{k^2}, \tag{3.9}$$

to leading order, and it is easy to show that at the next order

$$g(k) - f(k) \sim 4\left(\frac{h+1}{h}\right), \tag{3.10}$$

so that we may write

$$g(k) - f(k) \sim \beta^2 k^2 f(k), \tag{3.11}$$

in which

$$\beta^2 = \frac{2}{3}\frac{h^2}{h^2-h+1}. \tag{3.12}$$

Note that β^2 tends to its minimum value $2/3$ as $h \to 1$ or $h \to \infty$ and takes the maximum value $8/9$. This small variation in value over all h is important to ensure an eventually uniform factorization accuracy. We may also express $e(k)$ in terms of $f(k)$ near the origin:

$$e(k) \sim \delta k f(k), \tag{3.13}$$

where the parameter δ takes the value

$$\delta = \frac{h(h-1)}{h^2-h+1} \tag{3.14}$$

and is monotonic in h going from $\delta = 0$ at $h = 1$ to $\delta = 1$ at $h = \infty$. Again, this small variation will prove helpful to the factorization.

We now arrange $\mathbf{L}(k)$ to appear in Khrapkov form, namely, that the square of the second matrix term should be a scalar polynomial in k times the identity. We can do this by removing the factor $\sqrt{f^2(k) + e^2(k)}$ from this matrix in (3.6). However, as

$$f(k) \pm ie(k) \sim f(k)(1 \pm i\delta k), \qquad k \to 0, \tag{3.15}$$

it is more effective [21] to write $\mathbf{L}(k)$ as

$$\mathbf{L}(k) = g(k)\mathbf{I} + \sqrt{\frac{f^2(k) + e^2(k)}{1 + \delta^2 k^2}}\mathbf{J}(k), \tag{3.16}$$

$$\mathbf{J}(k) = \begin{pmatrix} 0 & d(k)(1 + i\delta k) \\ d^{-1}(k)(1 - i\delta k) & 0 \end{pmatrix}, \tag{3.17}$$

in which

$$d(k) = \sqrt{\left(\frac{f(k) + ie(k)}{f(k) - ie(k)}\right)\left(\frac{1 - i\delta k}{1 + i\delta k}\right)}. \tag{3.18}$$

It is a simple matter to show that we can choose a branch of $d(k)$ which is regular in $\mathcal{D}$, takes the value unity at $k = 0$, and, in view of the relative magnitudes of $e(k)$, $f(k)$ as $|k| \to \infty$ in $\mathcal{D}$, (3.2), (3.3), also tends to unity at infinity in the strip. If we had omitted the factor $(1 - i\delta k)/(1 + i\delta k)$ in $d(k)$, then $\arg(d(k))$ would not have tended to zero as $k \to \pm\infty$.

The matrix $\mathbf{L}(k)$ now appears to be in Khrapkov form, in view of the property

$$\mathbf{J}^2(k) = \Delta^2(k)\mathbf{I}, \tag{3.19}$$

where $\Delta^2(k)$ is the polynomial

$$\Delta^2(k) = 1 + \delta^2 k^2. \tag{3.20}$$

However, $\mathbf{J}(k)$ is not entire, as required for a Khrapkov factorization, but contains $d(k)$, which has infinite sequences of finite branch cuts at rotationally symmetric locations in the upper and lower half-planes. These will have to be considered once the partial Khrapkov decomposition is complete but, for the present, will be ignored.

3.3. Partial decomposition of $\mathbf{K}(k)$.

3.3.1. Limiting values of $\mathbf{L}(k)$. The first point to remark here is that (3.19) is arranged in appropriate form for the limiting values of h. As $h \to 1$, expression (2.20) reveals immediately that $e(k) = 0$ and similarly, from (3.14), $\delta = 0$. Hence $\mathbf{L}(k)$ reduces to

$$\mathbf{L}(k) = g(k)\mathbf{I} + f(k)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad h = 1. \tag{3.21}$$

By adding and subtracting rows this can be trivially reduced to two scalar decomposition problems, but this will also be decomposed *exactly* in the following Khrapkov factorization as $d(k) \equiv 1$. Similarly, as $h \to \infty$, $\delta \to 1$ and

$$e(k) = kf(k) = \frac{2k^3}{\sinh^2 k - k^2}, \qquad h \to \infty. \tag{3.22}$$

Hence in this limit $d(k) = 1$ also and

$$\mathbf{L}(k) = g(k)\mathbf{I} + f(k)\begin{pmatrix} 0 & 1+ik \\ 1-ik & 0 \end{pmatrix}, \qquad h = \infty, \tag{3.23}$$

which also permits an exact factorization [23] and justifies the particular form of $d(k)$ chosen in (3.18).

3.3.2. Introduction of resolvent matrix. Before performing the Khrapkov factorization on $\mathbf{L}(k)$, there is a problem, alluded to above, which must be resolved first. Note that as $k \to 0$, from (3.6), (3.11), and (3.13),

$$\mathbf{L}(k) \sim f(k)\begin{pmatrix} 1+\beta^2k^2 & 1+i\delta k \\ 1-i\delta k & 1+\beta^2k^2 \end{pmatrix} \tag{3.24}$$

so that

$$|\mathbf{L}(k)| \sim \left[6\frac{h^3+1}{h^3k^2}\right]^2 (2\beta^2-\delta^2)k^2 \sim \frac{12(h+1)^4}{h^4k^2} + \mathcal{O}(1). \tag{3.25}$$

This is clearly singular at the origin and therefore violates the original assumption of regularity in $\mathcal{D}$. Of course, this is because we are working with $\mathbf{L}(k)$ and not the original kernel $\mathbf{K}(k)$. To overcome this "removable singularity" in the determinant it is convenient to introduce the new matrix, $\mathbf{R}(k)$, called the resolvent, where

$$\mathbf{R}^{-1}(k) = (1+\beta^2k^2)\mathbf{I} - \mathbf{J}(k), \tag{3.26}$$

with $\mathbf{J}(k)$ as in (3.17), which commutes with $\mathbf{L}(k)$. The combined matrix

$$\mathbf{T}(k) = \mathbf{R}^{-1}(k)\mathbf{L}(k) \tag{3.27}$$

has determinant value

$$\left[\frac{2(h+1)^3}{h(h^2-h+1)}\right]^2 \tag{3.28}$$

at $k = 0$, and so $\mathbf{T}(k)$ may now be factorized instead of $\mathbf{L}(k)$. We will later have to deal with factorizing $\mathbf{R}(k)$, but this will not prove to be a problem.

3.3.3. Partial decomposition of matrix $\mathbf{T}(\boldsymbol{k})$. We have seen above that $\mathbf{R}^{-1}(k)$ and $\mathbf{L}(k)$ commute, and indeed any matrices of the form $\alpha\mathbf{I} + \beta\mathbf{J}(k)$ will commute with any other. Therefore, we may pose (see [19]) the product factors of $\mathbf{T}(k)$ in the form

$$\mathbf{T}^{\pm}(k) = r^{\pm}(k)\left(\cosh[\Delta(k)\theta^{\pm}(k)]\mathbf{I} + \frac{1}{\Delta(k)}\sinh[\Delta(k)\theta^{\pm}(k)]\mathbf{J}(k)\right), \tag{3.29}$$

where $r^{\pm}(k)$, $\theta^{\pm}(k)$ are scalar functions of k with the analyticity property indicated by their superscript. The function $\Delta(k)$, given by (3.20), generates no branch cuts because (3.29) contains only even powers of $\Delta(k)$. The scalar factors $r^{\pm}(k)$, $\theta^{\pm}(k)$ are deduced by equating

$$\mathbf{T}(k) = \mathbf{T}^{+}(k)\mathbf{T}^{-}(k), \tag{3.30}$$

which yields

$$r^+(k)r^-(k)\cosh[\Delta(k)(\theta^+(k)+\theta^-(k))] = g(1+\beta^2k^2) - \Delta\sqrt{f^2+e^2}, \tag{3.31}$$

$$r^+(k)r^-(k)\sinh[\Delta(k)(\theta^+(k)+\theta^-(k))] = \sqrt{f^2+e^2}(1+\beta^2k^2) - \Delta g. \tag{3.32}$$

These may be separated to give

$$[r^+(k)r^-(k)]^2 = (g^2-f^2-e^2)\left[(2\beta^2-\delta^2)+\beta^4k^2\right]k^2, \tag{3.33}$$

$$\tanh[\Delta(k)(\theta^+(k)+\theta^-(k))] = \frac{\sqrt{f^2+e^2}(1+\beta^2k^2)-\Delta g}{g(1+\beta^2k^2)-\Delta\sqrt{f^2+e^2}}, \tag{3.34}$$

and by the usual sum-split formula (e.g., equation (1.17) of Noble [1])

$$\begin{aligned}\theta^+(k) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{1}{\Delta(\zeta)}\tanh^{-1}\left\{\frac{\sqrt{f^2(\zeta)+e^2(\zeta)}(1+\beta^2\zeta^2)-\Delta(\zeta)g(\zeta)}{g(\zeta)(1+\beta^2\zeta^2)-\Delta(\zeta)\sqrt{f^2(\zeta)+e^2(\zeta)}}\right\}\frac{d\zeta}{\zeta-k}\\ &= \frac{k}{\pi i}\int_0^{\infty}\frac{1}{\Delta(\zeta)}\tanh^{-1}\left\{\frac{\sqrt{f^2(\zeta)+e^2(\zeta)}(1+\beta^2\zeta^2)-\Delta(\zeta)g(\zeta)}{g(\zeta)(1+\beta^2\zeta^2)-\Delta(\zeta)\sqrt{f^2(\zeta)+e^2(\zeta)}}\right\}\frac{d\zeta}{\zeta^2-k^2},\end{aligned} \tag{3.35}$$

valid for $\Im(k) > 0$. Note that the last result is true because the integrand is even in ζ, and this further implies that

$$\theta^-(k) = \theta^+(-k), \qquad k \in \mathcal{D}^-. \tag{3.36}$$

Actually, the full range integral could be taken along any path in $\mathcal{D}$ parallel to the real axis, and so if $\theta^+(k)$ is required for real k, then the first integral would be indented below (above for $\theta^-(k)$) this point. We can confirm that the integral representations in (3.35) exist by examining the integrand as $\zeta \to 0$ and $\zeta \to \infty$ (it is finite valued at all other points in $\mathcal{D}$). From (3.7)–(3.9) a little algebra reveals that the right-hand side of (3.34) is $\mathcal{O}(k^2)$, $k \to 0$, and similarly (3.2)–(3.4) suggests that (3.34) is $\mathcal{O}(k^{-1})$, $k \to \infty$. Hence, the first integrand in (3.35) is bounded in $\mathcal{D}$ and decays proportionally to $\mathcal{O}(\zeta^{-3})$ as $|\zeta| \to \infty$ in the strip. Therefore, this representation is ideal for computing $\theta^\pm(k)$ and can be directly coded for numerical evaluation.

This procedure has to be modified for $[r^+(k)r^-(k)]^2$ in (3.33) whose right-hand side tends to $[2(h+1)^3/h(h^2-h+1)]^2$ as $k \to 0$ and $\sim 16\beta^4k^6$ as $|k| \to \infty, k \in \mathcal{D}$. The latter behavior is not suitable for direct application of the product decomposition formula (Noble [1, equation (1.20)]), which requires a function that tends to the value unity at infinity. This is simply circumvented by applying a suitable divisor to (3.33), employing the standard factorization formula, and then decomposing the divisor into upper- and lower-half functions by inspection. This yields

$$\begin{aligned}r^+(k) = &\left[3^{1/2}\left(\frac{h+1}{h}\right)-2ik\right]^{1/2}\left[3^{1/4}\left(\frac{h+1}{h}\right)(1+i)-2ik\right]^{1/2}\\ &\times\left[3^{1/4}\left(\frac{h+1}{h}\right)(1-i)-2ik\right]^{1/2}\frac{h}{[3(h^2-h+1)]^{1/2}}\\ &\times\exp\left\{\frac{1}{4\pi i}\int_{-\infty}^{\infty}\log\left[\frac{[g^2(\zeta)-f^2(\zeta)-e^2(\zeta)]\zeta^2}{12\left(\frac{h+1}{h}\right)^4+16\zeta^4}\right]\frac{d\zeta}{\zeta-k}\right\}\end{aligned} \tag{3.37}$$

for $\Im(k) > 0$, and indentation of the contour below k is taken if k is real. Note that the exponential function in this expression may be reexpressed as

$$\exp\left\{\frac{k}{2\pi i}\int_0^\infty \log\left[\frac{[g^2(\zeta)-f^2(\zeta)-e^2(\zeta)]\zeta^2}{12\left(\frac{h+1}{h}\right)^4+16\zeta^4)}\right]\frac{d\zeta}{\zeta^2-k^2}\right\}, \tag{3.38}$$

where convergence of this and the above integral are now ensured. The function $r^-(k)$, analytic in the lower half-plane, is again, due to the symmetry, simply obtained from

$$r^-(-k) = r^+(k), \qquad k \in \mathcal{D}^+. \tag{3.39}$$

Hence $\mathbf{T}^\pm(k)$ have been determined (see (3.29), (3.35), (3.37)) in a form which can be evaluated directly, and these are analytic in their indicated half-planes, $\mathcal{D}^\pm$, except for the singularities occurring in $\mathbf{K}(k)$ (due to $d(k)$) which have yet to be resolved.

3.3.4. Partial decomposition of the resolvent matrix. Having introduced the inverse of $\mathbf{R}(k)$ above in order to improve the convergence of $\mathbf{L}(k)$, we now need to factorize it directly. The form of $\mathbf{R}(k)$ has been chosen to enable us to do this easily. First, $\mathbf{R}^{-1}(k)$ may, by inspection, be written in the form

$$\frac{1}{2}[(1+ik\sqrt{2\beta^2-\delta^2})\mathbf{I}-\mathbf{J}(k)][(1-ik\sqrt{2\beta^2-\delta^2})\mathbf{I}-\mathbf{J}(k)], \tag{3.40}$$

where both matrices are entire save for the finite cuts in the scalar function $d(k)$ contained within $\mathbf{J}(k)$. The first matrix has determinant

$$-2\beta^2 k(k-i\gamma), \tag{3.41}$$

where

$$\gamma = \frac{\sqrt{2\beta^2-\delta^2}}{\beta^2} = \frac{\sqrt{3}}{2}\left(\frac{h+1}{h}\right), \tag{3.42}$$

and the second has determinant

$$-2\beta^2 k(k+i\gamma). \tag{3.43}$$

Hence we may write

$$\mathbf{R}(k) = 2\mathbf{R}^+(k)\mathbf{R}^-(k), \tag{3.44}$$

where

$$\mathbf{R}^\pm(k) = \frac{1}{2\beta^2 k(k\pm i\gamma)}\left[(1\mp ik\beta^2\gamma)\mathbf{I}+\mathbf{J}(k)\right] \tag{3.45}$$

are the partial decomposition matrices; i.e., they are analytic in their indicated half-planes except for poles at $k=0$ and the finite branch cuts in $d(k)$. Note that $\mathbf{R}^\pm(k)$ commute with each other and with $\mathbf{T}^\pm(k)$, while the pole at $k=i\gamma$ lies in the upper half-plane. Hence, this completes the partial product factorization of $\mathbf{K}(k)$, and from (3.5), (3.27), (3.30), (3.44), we obtain

$$\mathbf{K}(k) = \mathbf{Q}^-(k)\mathbf{Q}^+(k), \tag{3.46}$$

where

$$\mathbf{Q}^-(k) = \begin{pmatrix} -ik & -ik \\ 1 & -1 \end{pmatrix} \mathbf{R}^-(k)\mathbf{T}^-(k), \tag{3.47}$$

$$\mathbf{Q}^+(k) = \mathbf{T}^+(k)\mathbf{R}^+(k) \begin{pmatrix} ik & 1 \\ ik & -1 \end{pmatrix}. \tag{3.48}$$

Note that $\mathbf{Q}^{\pm}(k)$ are free of a pole singularity at $k = 0$ even though $\mathbf{R}^{\pm}(k)$ contain this singularity (verified in section 3.4.3). All that remains is to remove (approximately) the residual singularities appearing in $\mathbf{J}(k)$.

3.4. Approximate factorization.

3.4.1. Padé approximation and partial decomposition of approximate kernel. There is no exact procedure known for eliminating the finite branch cuts in $d(k)$ from the upper (lower) half-planes of the matrix factor $\mathbf{Q}^+(k)$ ($\mathbf{Q}^-(k)$). To obtain an approximate factorization we replace the original matrix $\mathbf{K}(k)$ by a new one, $\mathbf{K}_N(k)$, where

$$\mathbf{K}_N(k) = \frac{1}{2}\begin{pmatrix} 0 & -k \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ i & i \end{pmatrix}\mathbf{L}_N(k)\begin{pmatrix} i & 1 \\ i & -1 \end{pmatrix}\begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.49}$$

$$\mathbf{L}_N(k) = g(k)\mathbf{I} + \sqrt{\frac{f^2(k) + e^2(k)}{1 + \delta^2 k^2}}\mathbf{J}_N(k), \tag{3.50}$$

and $\mathbf{J}_N(k)$ is as given in (3.17) but with a modified scalar $d(k) \to d_N(k)$, i.e.,

$$\mathbf{J}_N(k) = \begin{pmatrix} 0 & d_N(k)(1 + i\delta k) \\ d_N^{-1}(k)(1 - i\delta k) & 0 \end{pmatrix}. \tag{3.51}$$

We follow the procedure outlined in articles [21, 10, 23, 22] closely and so do not give the arguments here, contained in those papers, for the convergence of approximate factors to the exact ones. It will suffice to later verify the results obtained herein by numerical experiment. The scalar $d_N(k)$ is any function which approximates $d(k)$ accurately in the strip $\mathcal{D}$, and for efficacy of the following method it is most convenient to use a rational function approximation

$$d_N(k) = \frac{P_N(k)}{Q_N(k)}, \tag{3.52}$$

where $P_N(k)$, $Q_N(k)$ are polynomial functions of order N. Note that the order of each polynomial is the same, as we require that $d_N(k) \to 1$ as $|k| \to \infty$. There is a variety of ways of generating the coefficients of these polynomials, and the simplest and perhaps most justifiable (in terms of its analyticity properties) is to use Padé approximants [28]. As a note of caution, we must check that $d_N(k)$ does not introduce spurious singularities into the strip of analyticity $\mathcal{D}$; otherwise we will produce an inaccurate factorization. One-point Padé approximants, if they exist, are determined uniquely from the Taylor series expansion of the original function at any point of regularity. If we work with the origin, then the (one-point) Padé approximant of $d(k)$ is found by solving

$$\sum_{i=0}^{\infty} e_i k^i - \frac{P_N(k)}{Q_N(k)} = \mathcal{O}(k^{2N+1}), \tag{3.53}$$

where $\sum_{i=0}^{\infty} e_i k^i$ is the Maclaurin expansion of $d(k)$. This provides ample accuracy for our purposes (see section 5), due to the rapid decay at large real k that is otherwise present in the Fourier transform inversion formulas (4.7) and (4.8).

Note that the approximation of just $d(k)$ ensures that the scalar Khrapkov factors (3.35), (3.37) remain the same, etc., and so a partial decomposition of $\mathbf{K}_N(k)$ is simply

$$\mathbf{K}_N(k) = \mathbf{Q}_N^-(k)\mathbf{Q}_N^+(k), \tag{3.54}$$

in which $\mathbf{Q}_N^{\pm}(k)$ are given by (3.47), (3.48), with $\mathbf{R}^{\pm}(k)$ replaced by $\mathbf{R}_N^{\pm}(k)$ and $\mathbf{T}^{\pm}(k)$ replaced by $\mathbf{T}_N^{\pm}(k)$, for which the subscript N denotes that $\mathbf{J}_N(k)$, given by (3.51), replaces $\mathbf{J}(k)$ everywhere. Thus, the factorization of $\mathbf{K}_N(k)$ has been accomplished apart from sequences of poles, arising from the zeros and poles of $d_N(k)$ occurring in both half-planes exterior to $\mathcal{D}$. If we can remove these singularities, then an explicit exact factorization of $\mathbf{K}_N(k)$ will have been achieved, which approximates the actual factors $\mathbf{K}^{\pm}(k)$ in their regions of analyticity.

3.4.2. Removal of pole singularities. The exact factorization of $\mathbf{K}_N(k)$, given by (3.49), may be written as

$$\mathbf{K}_N(k) = \mathbf{K}_N^-(k)\mathbf{K}_N^+(k), \tag{3.55}$$

$$\mathbf{K}_N^-(k) = \mathbf{Q}_N^-(k)\mathbf{M}(k), \qquad \mathbf{K}_N^+(k) = \mathbf{M}^{-1}(k)\mathbf{Q}_N^+(k), \tag{3.56}$$

in which $\mathbf{M}(k)$ must be a meromorphic matrix which has to be chosen to eliminate the poles of $\mathbf{Q}_N^-(k)$ in the lower half-plane and the poles of $\mathbf{Q}_N^+(k)$ in $\mathcal{D}^+$. We can pose a (nonunique) ansatz for $\mathbf{M}(k)$ after noting certain symmetry properties of $\mathbf{Q}_N^{\pm}(k)$. First, from (3.18),

$$d(-k) = 1/d(k), \tag{3.57}$$

which must be reflected in the similar approximant behavior:

$$d_N(-k) = 1/d_N(k). \tag{3.58}$$

Thus,

$$\mathbf{J}_N(-k) = [\mathbf{J}_N(k)]^T, \tag{3.59}$$

where the superscript denotes the transpose, and so by inspection of (3.45),

$$\mathbf{R}_N^+(-k) = [\mathbf{R}_N^-(k)]^T. \tag{3.60}$$

Similarly, from (3.36), (3.39) and the obvious evenness of $\Delta(k)$ in (3.20), changing k to $-k$ in (3.29) reveals

$$\mathbf{T}_N^+(-k) = [\mathbf{T}_N^-(k)]^T. \tag{3.61}$$

Hence we find (see (3.48)) that

$$\mathbf{Q}_N^+(-k) = [\mathbf{T}_N^-(k)]^T[\mathbf{R}_N^-(k)]^T \begin{pmatrix} -ik & -ik \\ 1 & -1 \end{pmatrix}^T = [\mathbf{Q}_N^-(k)]^T \tag{3.62}$$

and deduce that the second equation in (3.56) gives

$$\mathbf{K}_N^+(-k) = \mathbf{M}^{-1}(-k)[\mathbf{Q}_N^-(k)]^T. \tag{3.63}$$

Symmetry properties dictate, by comparison with the first equation of (3.56), that we can construct a suitably scaled $\mathbf{M}(k)$ so that

$$\mathbf{M}^{-1}(-k) = [\mathbf{M}(k)]^T. \tag{3.64}$$

After this is achieved, it suffices to eliminate poles of $\mathbf{K}_N^-(k)$ in the lower half-plane.

Now suppose that $d_N(k)$ has N_p poles in the upper half-plane at $k = ip_n$, $n = 1, 2, \ldots, N_p$ ($ip_n \notin \mathcal{D}^-$), and N_q poles in the region below the strip at $k = -iq_n$, $n = 1, 2, \ldots, N_q$. That is, $Q_N(k)$ in (3.52) has zeros at $k = ip_n, -iq_n$. As has already been stated, there are, in total,

$$N_p + N_q = N \tag{3.65}$$

simple poles in the complex plane, and, due to the symmetry (3.58), there are N simple zeros of $P_N(k)$ at

$$k = -ip_n, \qquad n = 1, 2 \ldots, N_p; \qquad k = iq_n, \qquad n = 1, 2 \ldots, N_q, \tag{3.66}$$

in the lower and upper regions, respectively. Thus, $d_N(k)$ and its inverse may be expressed as Mittag–Leffler expansions:

$$d_N(k) = 1 + \sum_{n=1}^{N_p} \frac{\alpha_n}{p_n + ik} + \sum_{n=1}^{N_q} \frac{\beta_n}{q_n - ik}, \tag{3.67}$$

$$\frac{1}{d_N(k)} = 1 + \sum_{n=1}^{N_p} \frac{\alpha_n}{p_n - ik} + \sum_{n=1}^{N_q} \frac{\beta_n}{q_n + ik}, \tag{3.68}$$

where both tend to unity at infinity by virtue of $d_N(k)$ being a one-point Padé approximant of $d(k)$ in (3.18). The coefficients α_n, β_n are easily determined from the coefficients of the polynomials $P_N(k)$, $Q_N(k)$, the numerator and denominator, respectively, of $d_N(k)$. By inspection of the location of $d_N(k)$ in $\mathbf{Q}_N^-(k)$, the ansatz for $\mathbf{M}(k)$ is now posed (cf. those offered in [10, 23]) as

$$\begin{aligned} &\mathbf{M}(k) \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{A_n}{p_n+ik} + \sum_{n=1}^{N_q} \frac{B_n}{q_n-ik} & -\left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{C_n}{p_n+ik} + \sum_{n=1}^{N_q} \frac{D_n}{q_n-ik}\right) \\ \left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{C_n}{p_n-ik} + \sum_{n=1}^{N_q} \frac{D_n}{q_n+ik}\right) & \frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{A_n}{p_n-ik} + \sum_{n=1}^{N_q} \frac{B_n}{q_n+ik} \end{pmatrix}, \end{aligned} \tag{3.69}$$

where A_n, B_n, C_n, D_n are as yet undetermined constants. This form encapsulates the zeros and singularities of $d_N(k)$ and is chosen to satisfy the symmetry relation (3.64). However, the latter holds only if $|\mathbf{M}(k)| \equiv 1$, whereas (3.69) gives

$$\begin{aligned} |\mathbf{M}(k)| = {} & \left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{A_n}{p_n + ik} + \sum_{n=1}^{N_q} \frac{\bar{B}_n}{q_n - ik}\right) \left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{A_n}{p_n - ik} + \sum_{n=1}^{N_q} \frac{B_n}{q_n + ik}\right) \\ & + \left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{C_n}{p_n - ik} + \sum_{n=1}^{N_q} \frac{D_n}{q_n + ik}\right) \left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p} \frac{C_n}{p_n + ik} + \sum_{n=1}^{N_q} \frac{D_n}{q_n - ik}\right). \end{aligned} \tag{3.70}$$

The four sets of poles can be eliminated by setting the coefficients to satisfy the two systems of equations

$$\begin{aligned} &A_m\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{A_n}{p_n+p_m}+\sum_{n=1}^{N_q}\frac{B_n}{q_n-p_m}\right)\\ &+C_m\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{C_n}{p_n+p_m}+\sum_{n=1}^{N_q}\frac{D_n}{q_n-p_m}\right)=0 \quad (1\le m\le N_p), \end{aligned}\tag{3.71}$$

$$\begin{aligned} &B_m\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{A_n}{p_n-q_m}+\sum_{n=1}^{N_q}\frac{B_n}{q_n+q_m}\right)\\ &+D_m\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{C_n}{p_n-q_m}+\sum_{n=1}^{N_q}\frac{D_n}{q_n+q_m}\right)=0 \quad (1\le m\le N_q). \end{aligned}\tag{3.72}$$

Then $|\mathbf{M}(k)|$ is entire and takes the value unity at infinity. Hence Liouville's theorem implies that the determinant is indeed $|\mathbf{M}(k)|=1$ everywhere, as required.

By premultiplying $\mathbf{M}(k)$ by $\mathbf{R}^-(k)\mathbf{T}^-(k)$ and eliminating poles in the lower half-plane, conditions relating these coefficients can be found. From (3.29) and (3.45) we know that

$$\begin{aligned} &\left[(1+ik\beta^2\gamma)\mathbf{I}+\mathbf{J}_N(k)\right]\\ &\times\left[\cosh[\Delta(k)\theta^-(k)]\mathbf{I}+\frac{1}{\Delta(k)}\sinh[\Delta(k)\theta^-(k)]\mathbf{J}_N(k)\right]\mathbf{M}(k) \end{aligned}\tag{3.73}$$

must be analytic in $\mathcal{D}^-$, and so from (3.51) we wish to remove poles in the lower half-plane from

$$\begin{pmatrix} a^-(k) & b^-(k)(1+i\delta k)d_N(k)\\ b^-(k)(1-i\delta k)/d_N(k) & a^-(k)\end{pmatrix}\mathbf{M}(k),\tag{3.74}$$

where

$$a^\pm(k)=(1\mp ik\beta^2\gamma)\cosh[\Delta(k)\theta^\pm(k)]+\Delta(k)\sinh[\Delta(k)\theta^\pm(k)],\tag{3.75}$$

$$b^\pm(k)=\cosh[\Delta(k)\theta^\pm(k)]+\frac{(1\mp ik\beta^2\gamma)}{\Delta(k)}\sinh[\Delta(k)\theta^\pm(k)]\tag{3.76}$$

are scalar functions analytic in the indicated regions. Note that

$$a^-(k)=a^+(-k),\qquad b^-(k)=b^+(-k).\tag{3.77}$$

The top-left element of the matrix in (3.74) is, by employing (3.67),

$$\begin{aligned} &a^-(k)\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{A_n}{p_n+ik}+\sum_{n=1}^{N_q}\frac{B_n}{q_n-ik}\right)+b^-(k)(1+i\delta k)\\ &\times\left(1+\sum_{n=1}^{N_p}\frac{\alpha_n}{p_n+ik}+\sum_{n=1}^{N_q}\frac{\beta_n}{q_n-ik}\right)\left(\frac{1}{\sqrt{2}}+\sum_{n=1}^{N_p}\frac{C_n}{p_n-ik}+\sum_{n=1}^{N_q}\frac{D_n}{q_n+ik}\right), \end{aligned}\tag{3.78}$$

which appears to contain simple poles at $k=-ip_n$, $n=1,\dots,N_p$, $k=-iq_n$, $n=1,\dots,N_q$, in the lower half-plane unless they are suppressed. However, there are in fact no poles at $k = -ip_n$ because the sum of these terms multiplies $d_N(k)$, which we know is zero at these points. Thus, setting the expression in (3.78) to remain finite at the remaining singularity locations $k = -iq_n$ gives the relation, after use of (3.77),

(3.79)

$$a^+(iq_m)B_m + b^+(iq_m)\beta_m(1+\delta q_m)\left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p}\frac{C_n}{p_n - q_m} + \sum_{n=1}^{N_q}\frac{D_n}{q_n + q_m}\right) = 0, \qquad 1 \le m \le N_q.$$

Similarly the bottom-left element of (3.74) contains no poles in the lower half-plane if and only if

(3.80)

$$a^+(ip_m)C_m + b^+(ip_m)\alpha_m(1-\delta p_m)\left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p}\frac{A_n}{p_n + p_m} + \sum_{n=1}^{N_q}\frac{B_n}{q_n - p_m}\right) = 0, \qquad 1 \le m \le N_p.$$

Likewise, suppression of the lower half-plane poles in the second column of (3.74) yields

(3.81)

$$a^+(iq_m)D_m - b^+(iq_m)\beta_m(1+\delta q_m)\left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p}\frac{A_n}{p_n - q_m} + \sum_{n=1}^{N_q}\frac{B_n}{q_n + q_m}\right) = 0, \qquad 1 \le m \le N_q,$$

(3.82)

$$a^+(ip_m)A_m - b^+(ip_m)\alpha_m(1-\delta p_m)\left(\frac{1}{\sqrt{2}} + \sum_{n=1}^{N_p}\frac{C_n}{p_n + p_m} + \sum_{n=1}^{N_q}\frac{D_n}{q_n - p_m}\right) = 0, \qquad 1 \le m \le N_p.$$

By inspection, (3.79), (3.81) imply (3.72) and similarly (3.80), (3.82) imply (3.71). Therefore, not only do (3.79)–(3.82) enforce $\mathbf{K}_N^-(k)$ to be analytic in $\mathcal{D}^-$ as required, but relations (3.56), (3.62)–(3.64) reveal that they are also sufficient to ensure that $\mathbf{K}_N^+(k)$ is free of singularities in the half-plane $\mathcal{D}^+$, as are the inverses $[\mathbf{K}_N^{\pm}(k)]^{-1}$ in their indicated half-planes $\mathcal{D}^{\pm}$.

Thus (3.79)–(3.82) constitute a linear system of $2N$ equations for the $2N$ unknowns A_m, B_m, C_n, D_n and are easily solved to determine their values. Note that it may transpire that $1+\delta q_m$ or $1-\delta p_m$ is zero for particular choices of m, h, N, etc., in which case (B_m, D_m) or (A_m, C_m) would vanish. However, this does not present any difficulty (cf. equation (80) in [10]), and no cases have been encountered in which the system for A_m–D_m is singular.

3.4.3. Approximate noncommutative factorization. The explicit approximate factorization of $\mathbf{K}(k)$ is complete, having obtained an exact noncommutative matrix product decomposition of $\mathbf{K}_N(k)$. The factors $\mathbf{K}_N^{\pm}(k)$ are constructed from

(3.56), with $\mathbf{Q}_N^{\pm}(k)$ given from (3.51), (3.47), (3.48), (3.45), and (3.29). The meromorphic matrix $\mathbf{M}(k)$ takes the explicit form (3.69), in which the coefficients satisfy algebraic equations (3.79)–(3.82). As N increases it is expected that $\mathbf{K}_N^{\pm}(k)$ will converge rapidly to the exact factors $\mathbf{K}^{\pm}(k)$, and this will be borne out by numerical results given in section 5. All that remains here is to verify that the apparent pole at $k = 0$ in $\mathbf{R}^{\pm}(k)$ is removed and to give the behavior of $\mathbf{K}_N^{\pm}(k)$ for large $|k|$ in $\mathcal{D}^{\pm}$.

As $k \to 0$, we know that $d_N(k) \to 1$ by virtue of the function $d(k)$ in (3.18), and hence $\mathbf{R}_N^{\pm}(k)$ behaves as, from (3.45), (3.51),

$$\mathbf{R}_N^{\pm}(k) = \frac{\mp i}{2\beta^2 k\gamma}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \mathcal{O}(1). \tag{3.83}$$

Therefore,

$$\begin{pmatrix} -ik & -ik \\ 1 & -1 \end{pmatrix}\mathbf{R}_N^{-}(k) \sim \frac{1}{\beta^2\gamma}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} + \mathcal{O}(1), \qquad k \to 0. \tag{3.84}$$

Now, $\mathbf{T}_N^{\pm}(k)$, from their definitions, are bounded at the origin, and, by inspection, so is $\mathbf{M}(k)$ in (3.69). Hence, from (3.56) and (3.84) we can deduce that

$$\mathbf{K}_N^{-}(k) = \mathcal{O}(1), \qquad k \to 0. \tag{3.85}$$

Similarly, from above,

$$\mathbf{R}_N^{+}(k)\begin{pmatrix} ik & 1 \\ ik & -1 \end{pmatrix} = \mathcal{O}(1), \qquad k \to 0, \tag{3.86}$$

and so $\mathbf{K}_N^{+}(0)$ is bounded too.

As $|k|$ tends to infinity it is a straightforward matter to deduce the asymptotic behavior of the product factors. First, by inspection of (3.45),

$$\mathbf{R}_N^{\pm}(k) \sim \frac{i}{2\beta^2 k}\begin{pmatrix} \mp\beta^2\gamma & +\delta \\ -\delta & \mp\beta^2\gamma \end{pmatrix}, \tag{3.87}$$

in view of the fact that we *defined* $d_N(k)$ in $\mathbf{J}_N(k)$ to behave as

$$d_N(k) \to 1, \qquad |k| \to \infty. \tag{3.88}$$

Second, the asymptotic form of the Krapkhov decomposition elements $r^{\pm}(k)$, $\theta^{\pm}(k)$ can be deduced from their integral definitions written in (3.37), (3.39), and (3.35), (3.36), respectively. The latter identities are easily shown to give

$$\theta^{\pm}(k) = \pm\epsilon/k + \mathcal{O}(k^{-2}), \qquad |k| \to \infty, k \in \mathcal{D}^{\pm}, \tag{3.89}$$

where

$$\epsilon = \frac{i}{\pi}\int_0^{\infty} \frac{1}{\Delta(\zeta)} \tanh^{-1}\left\{\frac{\sqrt{f^2(\zeta) + e^2(\zeta)}(1 + \beta^2\zeta^2) - \Delta(\zeta)g(\zeta)}{g(\zeta)(1 + \beta^2\zeta^2) - \Delta(\zeta)\sqrt{f^2(\zeta) + e^2(\zeta)}}\right\} d\zeta, \tag{3.90}$$

and for $r^{\pm}(k)$ the integral in the exponent of (3.37) is also $\mathcal{O}(k^{-1})$ for large $|k|$. Hence by inspection we find that

$$r^{\pm}(k) = 2\beta(\mp ik)^{3/2} + \mathcal{O}(k^{1/2}), \qquad |k| \to \infty, k \in \mathcal{D}^{\pm}, \tag{3.91}$$

and so the asymptotic form of $\mathbf{T}_N^{\pm}(k)$, (3.29), is

$$\mathbf{T}_N^{\pm}(k) \sim 2\beta(\mp ik)^{3/2} \begin{pmatrix} \cosh(\epsilon\delta) & \pm i\sinh(\epsilon\delta) \\ \mp i\sinh(\epsilon\delta) & \cosh(\epsilon\delta) \end{pmatrix}. \tag{3.92}$$

Therefore, $\mathbf{Q}_N^{\pm}(k)$ in (3.47), (3.48) can be estimated. Finally, the meromorphic matrix (3.69) has the large $|k|$ form

$$\mathbf{M}(k) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. \tag{3.93}$$

Thus the asymptotic growth of $\mathbf{K}_N^{\pm}(k)$ in (3.56) is found to be

$$\begin{aligned} \mathbf{K}_N^{-}(k) \sim -\frac{(ik)^{1/2}}{\sqrt{2}\beta} & \begin{pmatrix} -ik & -ik \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \beta^2\gamma & \delta \\ -\delta & \beta^2\gamma \end{pmatrix} \\ & \times \begin{pmatrix} \cosh(\epsilon\delta) & -i\sinh(\epsilon\delta) \\ +i\sinh(\epsilon\delta) & \cosh(\epsilon\delta) \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \end{aligned} \tag{3.94}$$

$$\begin{aligned} \mathbf{K}_N^{+}(k) \sim -\frac{(-ik)^{1/2}}{\sqrt{2}\beta} & \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \cosh(\epsilon\delta) & i\sinh(\epsilon\delta) \\ -i\sinh(\epsilon\delta) & \cosh(\epsilon\delta) \end{pmatrix} \\ & \times \begin{pmatrix} \beta^2\gamma & -\delta \\ \delta & \beta^2\gamma \end{pmatrix} \begin{pmatrix} ik & 1 \\ ik & -1 \end{pmatrix}. \end{aligned} \tag{3.95}$$

The kernel decomposition is now complete.

4. Solution of the Wiener–Hopf equation.

Having obtained an approximate factorization of $\mathbf{K}(k)$, it is now a straightforward matter to complete the solution of the Wiener–Hopf equation (2.18). Dropping the subscript N in the factorization (3.55) for brevity, the Wiener–Hopf equation can be recast into the form

$$\begin{aligned} & [\mathbf{K}^-(k)]^{-1} \begin{pmatrix} T^-(k) \\ -S^-(k) \end{pmatrix} - \frac{i}{k+i\epsilon} \left\{ [\mathbf{K}^-(k)]^{-1} - [\mathbf{K}^-(-i\epsilon)]^{-1} \right\} \\ & \times \left(U\frac{h+1}{h} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{2\mu} \begin{bmatrix} 2G_+ - 2G_- \\ G_+h + G_- \end{bmatrix} \right) = \mathbf{E}(k) = \mathbf{K}^+(k) \begin{pmatrix} \Psi^+(k,0) \\ \Psi_y^+(k,0) \end{pmatrix} \\ & + \frac{i}{k+i\epsilon} [\mathbf{K}^-(-i\epsilon)]^{-1} \left(U\frac{h+1}{h} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{2\mu} \begin{bmatrix} 2G_+ - 2G_- \\ G_+h + G_- \end{bmatrix} \right), \end{aligned} \tag{4.1}$$

where $k \in \mathcal{D}$. The left-hand side is analytic in $\mathcal{D}^-$, whereas the right-hand side is regular in $\mathcal{D}^+$. Thus the equation has been arranged so that the two sides offer analytic continuation into the whole complex k-plane which must therefore be equal to an entire function, denoted $\mathbf{E}(k)$, say. To determine $\mathbf{E}(k)$ we must examine the growth at infinity of both sides of (4.1) in their respective half-planes of analyticity. To do this we require the large k behavior of $T^-(k)$, $S^-(k)$, $\Psi^+(k)$, $\Psi_y^+(k)$, which relate directly to the values of the untransformed physical variables near the tip of the splitter plate. For example, a function which behaves like x^n, $x \to 0+$, has a half-range (0 to ∞) Fourier transform which decays like $\mathcal{O}(k^{-n-1})$, $k \to \infty$, in the upper half-plane (see equation (1.74) of [1]). Hence from (2.27) and (2.29) we deduce, respectively, that

$$\Psi^{+}(k,0)=\mathcal{O}(k^{-5/2}),\qquad \Psi_y^{+}(k,0)=\mathcal{O}(k^{-3/2}) \tag{4.2}$$

as $|k|\to\infty$, $k\in\mathcal{D}^{+}$ and hence

$$\left(\begin{array}{c} T^{-}(k)\\ -S^{-}(k)\end{array}\right)-\frac{i}{k}\left(U\frac{h+1}{h}\left[\begin{array}{c}0\\1\end{array}\right]+\frac{1}{2\mu}\left[\begin{array}{c}2G_{+}-2G_{-}\\ G_{+}h+G_{-}\end{array}\right]\right)=\left(\begin{array}{c}\mathcal{O}(k^{1/2})\\ \mathcal{O}(k^{-1/2})\end{array}\right) \tag{4.3}$$

as $|k|\to\infty, k\in\mathcal{D}^{-}$. These are used, together with the asymptotic forms (3.94), (3.95), to reveal that both elements of the left-hand side of (4.1) decay as $\mathcal{O}(k^{-1})$ in the lower half-plane, and similarly the right-hand side has the form $\mathcal{O}(k^{-1})$ as $|k|\to\infty$ in the upper half-plane. Hence, $\mathbf{E}(k)$ is an entire function which decays to zero at infinity and so, by Liouville's theorem, is identically zero. Thus, the solution of the Wiener–Hopf equation is

$$\begin{aligned}\left(\begin{array}{c}\Psi^{+}(k,0)\\ \Psi_y^{+}(k,0)\end{array}\right)&=-\frac{i}{k+i\epsilon}[\mathbf{K}^{+}(k)]^{-1}[\mathbf{K}^{-}(-i\epsilon)]^{-1}\\ &\quad\times\left(U\frac{h+1}{h}\left[\begin{array}{c}0\\1\end{array}\right]+\frac{1}{2\mu}\left[\begin{array}{c}2G_{+}-2G_{-}\\ \\ G_{+}h+G_{-}\end{array}\right]\right)\end{aligned} \tag{4.4}$$

or, equivalently,

$$\begin{aligned}\left(\begin{array}{c}T^{-}(k)\\ -S^{-}(k)\end{array}\right)&=\frac{i}{k+i\epsilon}\left\{\mathbf{I}-\mathbf{K}^{-}(k)[\mathbf{K}^{-}(-i\epsilon)]^{-1}\right\}\\ &\quad\times\left(U\frac{h+1}{h}\left[\begin{array}{c}0\\1\end{array}\right]+\frac{1}{2\mu}\left[\begin{array}{c}2G_{+}-2G_{-}\\ G_{+}h+G_{-}\end{array}\right]\right).\end{aligned} \tag{4.5}$$

From this we can directly deduce the coefficients $A(k)$–$D(k)$, via (2.16), (2.17), and hence establish $\Psi(k,y)$ in $-1<y<h$, from (2.14), (2.15). Finally, on setting the convergence factor ϵ to zero in (4.4), the disturbance stream function is

$$\bar{\psi}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Psi(k,y)e^{-ikx}dk, \tag{4.6}$$

where the integral path runs along the real line indented above the origin, and

$$\begin{aligned}\Psi(k,y)&=\frac{-i}{\sinh^2 k-k^2}\left(\begin{array}{c}(1+y)\sinh[k(1+y)]\\ (1+y)\cosh[k(1+y)]-k^{-1}\sinh[k(1+y)]\end{array}\right)^{T}\\ &\quad\times\left(\begin{array}{cc}k\sinh k & -(\cosh k-k^{-1}\sinh k)\\ -(k\cosh k+\sinh k) & \sinh k\end{array}\right)[\mathbf{K}^{+}(k)]^{-1}[\mathbf{K}^{-}(0)]^{-1}\\ &\quad\times\left(U\frac{h+1}{h}\left[\begin{array}{c}0\\1\end{array}\right]+\frac{1}{2\mu}\left[\begin{array}{c}2G_{+}-2G_{-}\\ G_{+}h+G_{-}\end{array}\right]\right)\end{aligned} \tag{4.7}$$

in $-1<y<0$, whereas in $0<y<h$

$$\begin{aligned}\Psi(k,y)&=\frac{-i}{\sinh^2 kh-k^2h^2}\left(\begin{array}{c}(h-y)\sinh[k(h-y)]\\ (h-y)\cosh[k(h-y)]-k^{-1}\sinh[k(h-y)]\end{array}\right)^{T}\\ &\quad\times\left(\begin{array}{cc}kh\sinh kh & h\cosh kh-k^{-1}\sinh kh\\ -(kh\cosh kh+\sinh kh) & -h\sinh kh\end{array}\right)[\mathbf{K}^{+}(k)]^{-1}[\mathbf{K}^{-}(0)]^{-1}\\ &\quad\times\left(U\frac{h+1}{h}\left[\begin{array}{c}0\\1\end{array}\right]+\frac{1}{2\mu}\left[\begin{array}{c}2G_{+}-2G_{-}\\ G_{+}h+G_{-}\end{array}\right]\right).\end{aligned} \tag{4.8}$$

It is a straightforward matter to verify that this solution satisfies the biharmonic equation and the boundary and jump conditions (2.4), (2.5), (2.7). From this the downstream velocity field is determined from (2.3).

Formulas (4.7), (4.8) are suitable for evaluating $\bar{\psi}$, given by (4.6), in the $x < 0$ channels by completing the contour in the upper half-plane. Consistent with (2.3), there is no contribution from the pole at $k = 0$ to (4.6), which consists of Papkovich–Fadle strip eigenfunctions, generated by residues at the zeros of $\sinh^2 k - k^2$ or $\sinh^2 kh - k^2h^2$. These infinite sums describe how, in each channel, the flow differs from its far downstream profile $u^{\infty}_{\mp}$.

However, the evaluation of $\bar{\psi}$ in the upstream ($x > 0$) channel is achieved by completing the contour in the lower half-plane. When $[\mathbf{K}^{+}(k)]^{-1}$ is replaced, according to (3.1), by $\operatorname{adj}\mathbf{K}(k)\mathbf{K}^{-}(k)/|\mathbf{K}(k)|$, the substitution of (2.19), (2.24) into (4.7) yields

$$\begin{aligned}(4.9)\qquad &\Psi(k,y)\\ &= \frac{i}{k^2[\sinh^2 k(h+1) - k^2(h+1)^2]}\begin{pmatrix} (1+y)\sinh[k(1+y)] \\ (1+y)\cosh[k(1+y)] - k^{-1}\sinh[k(1+y)] \end{pmatrix}^{T}\\ &\times \mathbf{V}(k)\begin{pmatrix} 1 & 0 \\ 0 & -k \end{pmatrix}\mathbf{K}^{-}(k)[\mathbf{K}^{-}(0)]^{-1}\left(U\frac{h+1}{h}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{2\mu}\begin{bmatrix} 2G_{+} - 2G_{-} \\ G_{+}h + G_{-} \end{bmatrix}\right),\end{aligned}$$

in which the elements of $\mathbf{V}(k)$ are concisely defined by

$$V_{11} + V_{22} = kh(h+1)\sinh k, \qquad V_{11} - V_{22} = -\sinh kh \sinh k(h+1),$$

$$V_{21} + V_{12} = k^{-1}\sinh kh \sinh k(h+1) - kh(h+1)\cosh k,$$

$$V_{21} - V_{12} = \sinh kh \cosh k(h+1) - h\sinh k.$$

The residue at the pole $k = 0$ yields the upstream behavior ($x \to \infty$)

$$\begin{aligned}(4.10)\qquad \bar{\psi} \sim h&\begin{pmatrix} \left(\frac{1+y}{h+1}\right)^2 & \left(\frac{1+y}{h+1}\right)^3 \end{pmatrix}\begin{pmatrix} -\frac{h}{2} & 1-\frac{h}{2} \\ \frac{h}{6}(h+3) & -1 \end{pmatrix}\\ &\times\left(U\frac{h+1}{h}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{2\mu}\begin{bmatrix} 2G_{+} - 2G_{-} \\ G_{+}h + G_{-} \end{bmatrix}\right)\end{aligned}$$

in $-1 < y < 0$, which, when substituted into (2.3), gives the net upstream flow form. That is, the y-derivative of (4.10) is verified, with use of (1.5), to equal $u^{\infty}(y) - u^{\infty}_{-}(y)$, given by (1.1), (1.3). A similar calculation, based on (4.8), verifies that $u^{\infty}(y) - u^{\infty}_{+}(y)$ is obtained in $(0, h)$. Identical series of Papkovich–Fadle eigenfunctions arise in (4.6) from residues associated with the zeros of $[\sinh^2 k(h+1) - k^2(h+1)^2]$. This infinite sum describes how the flow differs from its far upstream profile u^{∞} in $-1 < y < h$.

5. Numerical computation. The calculations are performed with MATLAB, except for the evaluation of the Padé approximants d_{2m} by means of Maple. The resulting fractions are then converted to floating points (16 digit accuracy) and returned to MATLAB. Accuracy is low unless $N = 2M$ with M even to take account of symmetries about both axes. Maximum accuracy occurs at about $N = 2$, $M = 16$ and could be increased by means of variable precision arithmetic. The poles $ip_m, -iq_m$

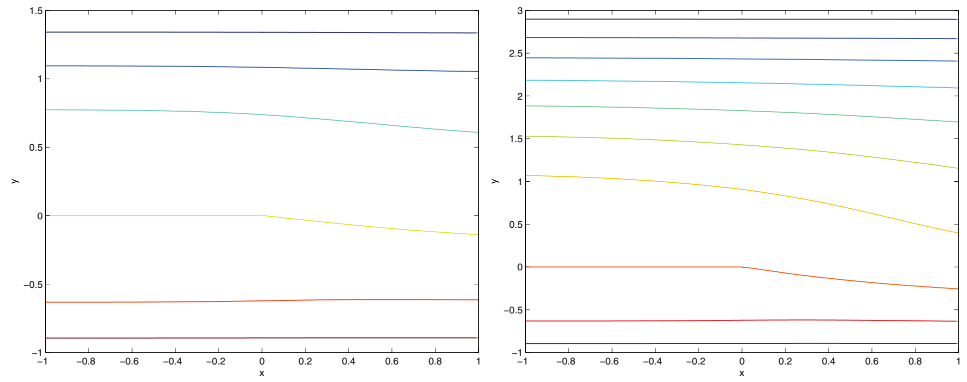


FIG. 5.1. *Stream function plots for the downstream shear case with $U = 1$ for (left) $h = 1.5$ and (right) $h = 3$ (note the different y-scale). The difference in stream function values (volume flux) between contours is* 0.2.

and coefficients α_m, β_m in (3.67) are then readily determined, followed by $\mathbf{J}_N(k)$ and $\mathbf{R}_N^+(k)$. Evaluation of $r_N^+(k), \theta_N^+(k)$ yields $\mathbf{T}_N^+(k)$ and hence $\mathbf{Q}_N^+(k)$, given by (3.48). The scalar Wiener–Hopf decomposition of r and θ is achieved by using standard MATLAB numerical integration. The default relative accuracy of 10^{-6} easily suffices because higher accuracy is actually obtained, as in many contour integrals of analytic functions. Finally, $\mathbf{K}_N^+(k)$ is constructed. The stream function, $\psi(x, y)$, is evaluated in either of the $x < 0$ channels as a sum of the residues of (4.7) or (4.8) at the respective first 50 poles in the upper half-plane. Note that knowledge of the two sets of residues allows the inverse Fourier transform (4.6) to be computed for any $x(< 0)$ at essentially zero marginal cost. The companion matrix function $\mathbf{K}_N^-(k)$, needed in (4.9) for $x > 0$, is constructed similarly.

The numerical evaluation of the approximation $[\mathbf{K}_N^+(k)]^{-1}$ to $[\mathbf{K}^+(k)]^{-1}$ in (4.7), (4.8) depends on the accurate determination of the coefficients $a^+(ip_m), \ldots$ in (3.79)–(3.82). These are given by (3.75), (3.76), in which $\Delta(k)$ appears analytically, but branch cuts may arise from the presence $\theta^\pm(k)$ in the sinh functions. By factoring $\Delta(\zeta)$ from the numerator of the fraction in (3.35), it is evident that the branch cuts created by the approximate factorization arise solely from the square root in the definition (3.16) of $\mathbf{L}(k)$.

Very high accuracy would require variable precision arithmetic and a large number of terms in the residue sum, especially when computing the stream function values near the entrance to the downstream channels.

Figure 5.1 displays streamlines for $h = 1.5$ and $h = 3$ in the downstream shear case ($G_+ = 0 = G_-$). For the same values of h, Figure 5.2 shows streamlines when the walls are stationary ($U = 0$) with respective flux ratios $\Lambda = 1, -0.5, -2$, which typify the physically distinct ranges, $\Lambda > 0$, $-1 < \Lambda < 0$, $\Lambda < -1$. The curves have an imperceptible defect at $x = 0$; the values of $\psi(0, y)$ computed using (4.7) and (4.9) are not exactly the same in the Padé approximant technique but would be identical for the exact matrix $\mathbf{K}$. This discrepancy provides an estimate of the error, which is found to decrease with N until at least $N = 12$, which is the value used in the figures. While the qualitative behavior in the pressure-driven case depends only on Λ, plots require a normalization of G_- and G_+: for convenience, $G_- = 12\mu$ was taken.

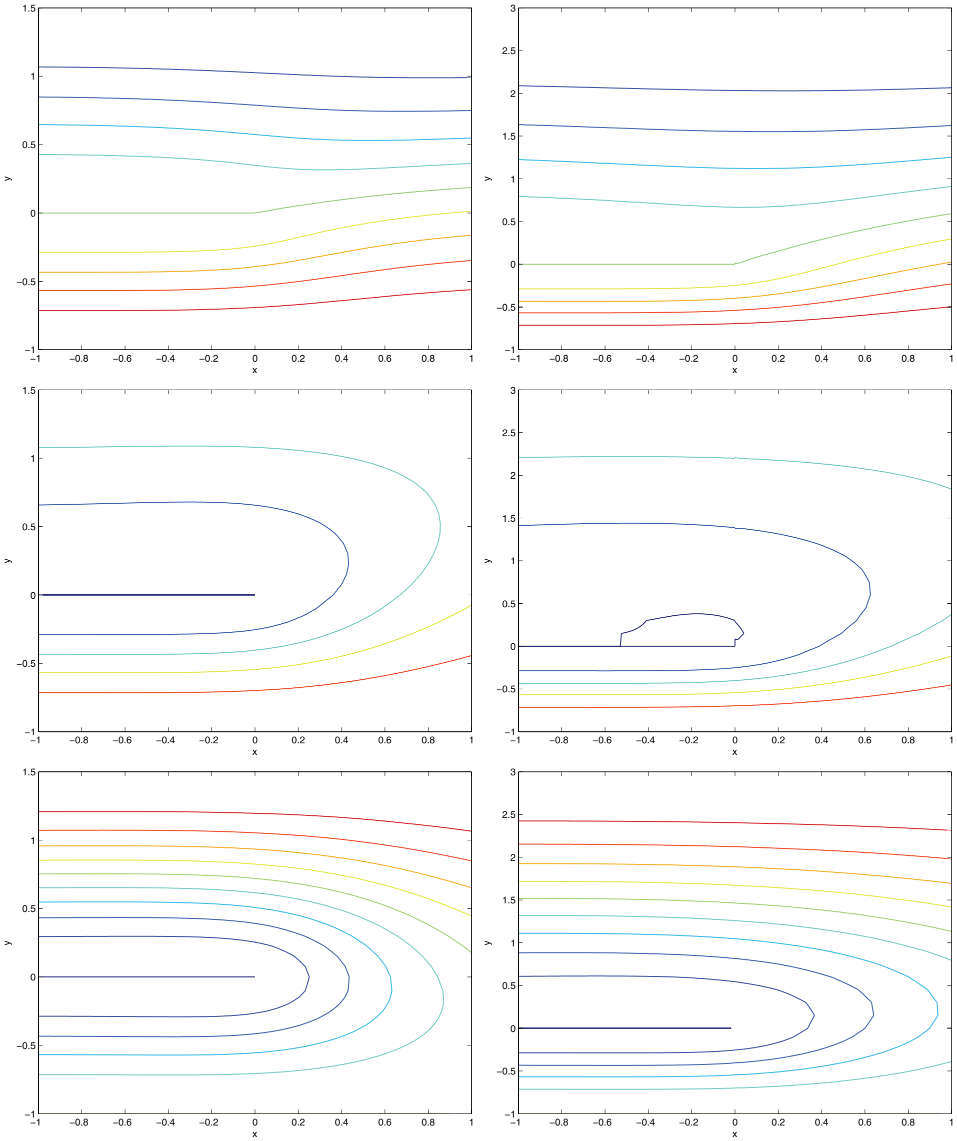


FIG. 5.2. *Stream function plots for the pressure-driven case. Left-hand panels:* $h = 1.5$*; right-hand panels:* $h = 3$*. Top row:* $\Lambda = 1$*, middle row:* $\Lambda = -0.5$*, bottom row:* $\Lambda = -2$*. The difference in stream function values (volume flux) between contours is* 0.2*. The jagged contour for* $h = 3$*,* $\Lambda = -0.5$ *is a contouring artifact.*

6. Conclusion. The Padé approximant technique for matrix Wiener–Hopf equations yields accurate numerical results for a classic Stokes flow problem for all channel width ratios. The theory is complicated by the need for successive modifications $\mathbf{L}$, $\mathbf{T}$ of the kernel and $\mathbf{M}$, $\mathbf{M}^{-1}$ of the matrix factors $\mathbf{Q}^{-}$, $\mathbf{Q}^{+}$ in order to establish the required analyticity of $\mathbf{K}^{-}$ and $\mathbf{K}^{+}$. The numerical implementation is not difficult conceptually but demands the usual careful attention to the analyticity properties of the functions involved. The technique provides a constructive scheme to obtain the physical solution without the major difficulties encountered in matching the three sets of biorthogonal Papkovich–Fadle eigenfunctions.

REFERENCES

[1] B. Noble, *Methods Based on the Wiener–Hopf Technique*, 2nd ed., Chelsea Press, New York, 1988.

[2] V. T. Buchwald and H. E. Doran, *Eigenfunctions of plane elastostatics.* II. *A mixed boundary value problem of the strip*, Proc. Roy. Soc. Ser. A, 284 (1965), pp. 69–82.

[3] R. M. L. Foote and V. T. Buchwald, *An exact solution for the stress intensity factor for a double cantilever beam*, Int. J. Fracture, 29 (1985), pp. 125–134.

[4] W. P. Graebel, *Slow viscous shear flow past a plate in a channel*, Phys. Fluids, 8 (1965), pp. 1929–1935.

[5] W. T. Koiter, *Approximate solution of Wiener–Hopf type integral equations with applications.* I. *General theory*, Proc. Acad. Sci. Amst. B, 57 (1954), pp. 558–579.

[6] S. Richardson, *A "stick-slip" problem related to the motion of a free jet at low Reynolds numbers*, Proc. Camb. Phil. Soc., 67 (1970), pp. 477–489.

[7] O. E. Jensen and D. Halpern, *The stress singularity in surfactant-driven thin-film flows. Part* 1. *Viscous effects*, J. Fluid Mech., 372 (1998) pp. 273–300.

[8] A. M. Moore, V. T. Buchwald, and M. E. Brewster, *The Stokesian entry flow*, Quart. J. Mech. Appl. Math., 43 (1990), pp. 107–133.

[9] M.-U. Kim, D. H. Choi, and J.-T. Jeong, *A two-dimensional model of a half-pitot tube*, Fluid Dyn. Res., 5 (1989), pp. 135–145.

[10] I. D. Abrahams, *On the solution of Wiener–Hopf problems involving noncommutative matrix kernel decompositions*, SIAM J. Appl. Math., 57 (1997), pp. 541–567.

[11] T. E. Stanton, D. Marshall, and C. N. Bryant, *On the conditions at the boundary of a fluid in turbulent motion*, Proc. Roy. Soc. Lond. A, 97 (1920), pp. 413–434.

[12] G. I. Taylor, *Measurements with a half-pitot tube*, Proc. Roy. Soc. Lond. A, 166 (1938), pp. 476–481.

[13] J. D. Fehribach and A. M. J. Davis, *Stokes flow around an asymmetric channel divider; a computational approach using MATLAB*, J. Engrg. Math., 39 (2001), pp. 207–220.

[14] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1993.

[15] W. D. Collins, *A note on the axisymmetric Stokes flow of viscous fluid past a spherical cap*, Mathematika, 10 (1963), pp. 72–79.

[16] A. M. J. Davis, *Axisymmetric Stokes flow past a spherical hollow boundary and concentric sphere*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 537–559.

[17] I. C. Gohberg and M. G. Krein, *Systems of integral equations on a half-line with kernels depending on the difference of arguments*, Amer. Math. Soc. Transl. (2), 14 (1960), pp. 217–287.

[18] R. A. Hurd, *The Wiener–Hopf Hilbert method for diffraction problems*, Canad. J. Phys., 54 (1976), pp. 775–780.

[19] A. A. Khrapkov, *Certain cases of the elastic equilibrium of an infinite wedge with a nonsymmetric notch at the vertex, subjected to concentrated forces*, Appl. Math. Mech. (PMM), 35 (1971), pp. 625–637.

[20] V. G. Daniele, *On the factorization of Wiener–Hopf matrices in problems solvable with Hurd's method*, IEEE Trans. Antennas and Propagation, 26 (1978), pp. 614–616.

[21] I. D. Abrahams, *Radiation and scattering of waves on an elastic half-space; a noncommutative matrix Wiener–Hopf problem*, J. Mech. Phys. Solids, 44 (1996), pp. 2125–2154.

[22] I. D. Abrahams, *On the application of the Wiener–Hopf technique to problems in dynamic elasticity*, Wave Motion, 36 (2002), pp. 311–333.

[23] I. D. Abrahams, *On the non-commutative factorization of Wiener–Hopf kernels of Khrapkov type*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 1719–1743.

[24] M. Idemen, *A new method to obtain exact solutions of vector Wiener–Hopf equations*, Z. Angew. Math. Mech., 59 (1976), pp. 656–658.

[25] A. D. Rawlins, *Simultaneous Wiener–Hopf equations*, Canad. J. Phys., 58 (1980), pp. 420–428.

[26] I. D. Abrahams, *Scattering of sound by two parallel semi-infinite screens*, Wave Motion, 9 (1987), pp. 289–300.

[27] I. D. Abrahams and G. R. Wickham, *The scattering of water waves by two semi-infinite opposed vertical walls*, Wave Motion, 14 (1991), pp. 145–168.

[28] G. A. Baker, Jr., and P. Graves-Morris, *Padé Approximants*, 2nd ed., Cambridge University Press, Cambridge, UK, 1996.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1464–1476

GLOBAL ASYMPTOTIC STABILITY FOR A CLASS OF NONLINEAR CHEMICAL EQUATIONS[*]

DAVID F. ANDERSON[†]

Abstract. We consider a class of nonlinear differential equations that arises in the study of chemical reaction systems known to be locally asymptotically stable and prove that they are in fact globally asymptotically stable. More specifically, we will consider chemical reaction systems that are weakly reversible, have a deficiency of zero, and are equipped with mass action kinetics. We show that if for each $c \in \mathbb{R}^m_{>0}$ the intersection of the stoichiometric compatibility class $c + S$ with the subsets on the boundary that could potentially contain equilibria, L_W, are at most discrete, then global asymptotic stability follows. Previous global stability results for the systems considered in this paper required $(c + S) \cap L_W = \emptyset$ for each $c \in \mathbb{R}^m_{>0}$.

Key words. chemical systems, deficiency, global stability, persistence, Petri nets

AMS subject classifications. 37C10, 80A30, 92C40, 92D25, 92E10, 93D20

DOI. 10.1137/070698282

1. Introduction.

This paper is motivated by the consideration of a class of nonlinear systems that arises in the study of chemistry and biochemistry. Suppose there are m chemical species, $\{X_1, \dots, X_m\}$, undergoing a series of chemical reactions. For a given reaction, denote by $y, y' \in \mathbb{Z}^m_{\geq 0}$ the vectors representing the number of molecules of each species consumed and created in one instance of that reaction, respectively. Using a slight abuse of notation, we associate each such y (and y') with a linear combination of the species in which the coefficient of X_i is y_i. For example, if $y = [1, 2, 3]^T$ for a system consisting of three species, we associate with y the linear combination $X_1 + 2X_2 + 3X_3$. Under this association, each y (and y') is termed a *complex* of the system. We may now denote any reaction by the notation $y \to y'$, where y is the source, or reactant, complex and y' is the product complex. We note that each complex may appear as both a source complex and a product complex in the system. Let $\mathcal{S} = \{X_i\}$, $\mathcal{C} = \{y\}$, and $\mathcal{R} = \{y \to y'\}$ denote the sets of species, complexes, and reactions, respectively. Denote the concentration vector of the species as $x \in \mathbb{R}^m$. In order to know how the state of the system is changing, we need to know the rate at which each reaction is taking place. Therefore, for each reaction $y \to y'$, there is a C^1 function $R_{y\to y'}(\cdot)$ satisfying the following:

1. $R_{y\to y'}(\cdot)$ is a function of the concentrations of those species contained in the source complex, y.
2. $R_{y\to y'}(\cdot)$ is monotone increasing in each of its inputs, and $R_{y\to y'}(x) = 0$ if *any* of its inputs are zero.

The dynamics of the system are then given by the coupled set of ordinary differential equations

$$\dot{x}(t) = \sum_{y\to y'\in\mathcal{R}} R_{y\to y'}(x(t))(y' - y) \doteq f(x(t)), \tag{1}$$

[*] Received by the editors July 25, 2007; accepted for publication (in revised form) January 30, 2008; published electronically May 16, 2008. This work was done under the support of grants NSF-DMS-0109872, NSF-DMS-0616710, and NSF-DMS-0553687.
http://www.siam.org/journals/siap/68-5/69828.html

[†] Department of Mathematics, University of Wisconsin, Madison, WI 53706 (anderson@math.wisc.edu).

where the last equality is a definition. The functions $R_{y\to y'}$ are typically referred to as the *kinetics* of the system. This notation closely matches that of Feinberg, Horn, and Jackson, and it is their work that the main results in this paper extend [12, 7, 9, 10, 8].

Integrating (1) gives

$$x(t) = x(0) + \sum_{y\to y'\in\mathcal{R}} \left(\int_0^t R_{y\to y'}(x(s))ds \right) (y' - y).$$

Therefore, $x(t) - x(0)$ remains in the linear space $S = \text{Span}\{y' - y\}_{y\to y'\in\mathcal{R}}$ for all time. We shall refer to the space S as the stoichiometric subspace of the system and refer to the sets $c + S$, for $c \in \mathbb{R}^m$, as stoichiometric compatibility classes, or just compatibility classes. Later we will demonstrate that trajectories with positive initial conditions remain in $\mathbb{R}^m_{>0}$ for all time. The sets $(c+S)\cap\mathbb{R}^m_{>0}$ will therefore be referred to as the *positive stoichiometric compatibility classes.* Given that trajectories remain in their positive stoichiometric compatibility classes for all time, we see that the types of questions that one should ask about these systems differ from the questions one normally asks about nonlinear systems. For example, instead of asking whether there is a unique equilibrium value to the system (1) and then asking about its stability properties, it is clearly more appropriate to ask whether there is a unique equilibrium *within each positive stoichiometric compatibility class* and, if so, what are its stability properties *relative to its compatibility class.*

The most common kinetics chosen is that of *mass action kinetics.* A chemical reaction system is said to have mass action kinetics if

$$R_{y\to y'}(x) = k_{y\to y'} x_1^{y_1} x_2^{y_2} \cdots x_m^{y_m} \tag{2}$$

for some constant $k_{y\to y'}$. It has been shown that, for many systems of the form (1) with mass action kinetics, there is within each stoichiometric compatibility class precisely one equilibrium with strictly positive components, and that equilibrium is locally asymptotically stable relative to its class [12, 9, 8]. In order to show that the equilibrium values are locally stable, the following Lyapunov function is used (one for each compatibility class):

$$V(x,\bar{x}) = V(x) = \sum_{i=1}^m \left[x_i(\ln(x_i) - \ln(\bar{x}_i) - 1) + \bar{x}_i \right], \tag{3}$$

where $\bar{x}$ is the unique equilibrium of a given positive stoichiometric compatibility class. It turns out that the function V "almost" acts as a global Lyapunov function. That is, V is nonnegative for $x \in (\bar{x}+S)\cap\mathbb{R}^m_{>0}$, zero only at $\bar{x}$, and strictly decreasing along trajectories. However, V does not tend to infinity as trajectories near the boundary of $(\bar{x}+S)\cap\mathbb{R}^m_{>0}$, and without such unboundedness one cannot, in general, conclude global stability. It has been shown in numerous papers, however, that global stability of $\bar{x}$ does hold if there are no equilibria on the boundary of $(\bar{x}+S)\cap\mathbb{R}^m_{>0}$ [15, 2, 13, 14]. Therefore, work has been done giving sufficient conditions for the nonexistence of boundary equilibria in order to conclude that the equilibrium value within each compatibility class is globally stable relative to its class [2, 13, 14].

To each subset W of the set of species, the set of points L_W is defined to be

$$L_W = \{x \in \mathbb{R}^m : x_i = 0 \Leftrightarrow X_i \in W\}. \tag{4}$$

We will show that there are no boundary equilibria if and only if

$$[(c+S)\cap\mathbb{R}^m_{\geq 0}] \cap L_W = \emptyset \tag{5}$$

for all $c \in \mathbb{R}^m_{>0}$ and for certain subsets of the species, W. We will then prove that global stability holds if the intersection given in (5) is either empty or discrete for each $c \in \mathbb{R}^m_{>0}$ and those same subsets, W. This will imply that global stability holds even if there are boundary equilibria, so long as the boundary equilibria are extreme points of the positive stoichiometric compatibility classes. To the best of our knowledge there is only one other result concerning the global stability of mass action systems with boundary equilibria, and this is contained within the Ph.D. thesis of Chavez [6]. In order to guarantee global stability even if there exist boundary equilibria, Chavez requires that each boundary equilibrium be hyperbolic with respect to its stoichiometric compatibility class, and she requires another (more technical) condition on the stable subspaces of each boundary equilibrium (see [6, pg. 106] for details). As our results are applicable to systems with boundary equilibria that are nonhyperbolic, our results can be viewed as an extension of those in [6].

The layout of the paper is as follows. In section 2 we will introduce the systems we consider in this paper: weakly reversible, deficiency zero systems with mass action kinetics. We will then present some preliminary results and conclude with a proof that global stability follows if there are no equilibria on the boundary of the positive stoichiometric compatibility classes. No originality is claimed for this result as it is known. Also in section 2 we demonstrate how the "no boundary equilibria" assumption is equivalent to (5) holding for all $c \in \mathbb{R}^m_{>0}$ and certain subsets of the species, W. In section 3 we extend the previous theorems to prove that global stability still holds if the intersection given in (5) is always either empty *or* discrete for those same subsets, W. We also show in section 3 how the hypothesis that the intersection in (5) is always empty or discrete is equivalent to a condition on the extreme points of the nonnegative stoichiometric compatibility classes. In section 4, we demonstrate our results on a number of examples. Finally, in section 5 we sketch how to extend our results to systems with non-mass action kinetics.

2. Preliminary results.

We start with definitions taken from [12, 8, 9].

DEFINITION 2.1. *A chemical reaction network, $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$, is called* weakly reversible *if for any reaction $y \to y'$, there is a sequence of directed reactions beginning with y' and ending with y. That is, there exist complexes $y_1, \dots, y_k$ such that the following reactions are in $\mathcal{R}$: $y' \to y_1$, $y_1 \to y_2, \dots, y_k \to y$.*

To each reaction network, $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$, there is a unique, directed graph constructed in the following manner. The nodes of the graph are the complexes, $\mathcal{C}$. A directed edge is then placed from complex y to complex y' if and only if $y \to y' \in \mathcal{R}$. Each connected component of the resulting graph is termed a *linkage class* of the graph. We denote the number of linkage classes by l.

As shown in the introduction, each trajectory remains in its stoichiometric compatibility class for all time. There is another restriction on the trajectories of solutions to (1) that is given in the following lemma. The proof can be found in both [15] and [1].

LEMMA 2.2. *Let $x(t)$ be a solution to* (1) *with initial condition $x(0) \in \mathbb{R}^m_{>0}$. Then, $x(t) \in \mathbb{R}^m_{>0}$ for all $t > 0$.*

2.1. Persistence and ω-limit points.

By Lemma 2.2, each trajectory must remain within $\mathbb{R}^m_{>0}$ if its initial condition is in $\mathbb{R}^m_{>0}$; therefore the linear subsets of interest are the intersections of the stoichiometric compatibility classes and $\mathbb{R}^m_{>0}$. Recall that in the introduction these sets were termed the positive stoichiometric compatibility classes. This paper will mainly be concerned with showing that trajectories to systems given by (1) remain away from the boundaries of the positive stoichiometric compatibility classes. That is, we will show that the systems are *persistent.* To be precise, let $\phi(t,\xi)$ be the solution to (1) with initial condition $\xi \in \mathbb{R}^m_{>0}$. The set of ω-limit points of the trajectory is

$$\omega(\xi) \doteq \{x \in \mathbb{R}^m_{\geq 0} : \phi(t_n,\xi) \to x, \text{ for some } t_n \to \infty\}. \tag{6}$$

Definition 2.3. *A system is* persistent *if* $\omega(\xi) \cap \partial\mathbb{R}^m_{>0} = \emptyset$ *for each* $\xi \in \mathbb{R}^m_{>0}$.

We refer the reader to [4, 3, 5, 16] for some of the history and usage of the notion of persistence in the study of dynamical systems. In order to show that a chemical system is persistent, it is critical to understand which points on the boundary are capable of being ω-limit points. With that in mind, we introduce the following definition.

Definition 2.4. *A nonempty subset W of the set of species is called a* semilocking set *if for each reaction in which there is an element of W in the product complex, there is an element of W in the reactant complex. W is called a* locking set *if every reactant complex contains an element of W.*

Locking and semilocking sets are easily understood. First suppose that $W \subset \{X_1,\dots,X_m\}$ is a locking set. Then, because every reactant complex contains an element of W, if the concentration of each element of W is zero, each kinetic function, $R_{y\to y'}$, must equal zero. Therefore, all of the fluxes are zero, and $\dot{x}(t) = 0$. We therefore see that the system is "locked" in place. Now suppose W is a semilocking set. If the concentration of each element of W is zero, then any flux which affects the species of W is turned off, and the elements of W are "locked" at zero. Semilocking sets have another, important, interpretation in terms of the linkage classes and weak reversibility. If the concentrations of the elements of a semilocking set are equal to zero and the system is weakly reversible, then all of the fluxes of any linkage class with a complex containing an element of W are equal to zero (and so these linkage classes are "locked"), while the fluxes of the other linkage classes are not necessarily equal to zero. Therefore, the concept of a semilocking set and a locking set are equivalent for systems that are weakly reversible and have only one linkage class. We note that our notion of a semilocking set is analogous to the concept of a *siphon* in the theory of Petri nets. See [2] for a full discussion, including historical references, of the role of Petri nets in the study of chemical reaction networks.

The following theorem now characterizes the boundary points that have the capability of being ω-limit points of the system. This result was first proved in [2]; however, the proof given here is completely different and straightforward.

Theorem 2.5. *Let W be a nonempty subset of the species. If there exists a $\xi \in \mathbb{R}^m_{>0}$ such that $\omega(\xi) \cap L_W \neq \emptyset$, then W is a semilocking set.*

Proof. Suppose, in order to find a contradiction, that there is a $\xi \in \mathbb{R}^m_{>0}$ and a subset of the species, W, such that $\omega(\xi) \cap L_W \neq \emptyset$ and W is not a semilocking set. Let $y \in \omega(\xi) \cap L_W$. We note that there exists a species X_j, with $X_j \in W$, such that at least one input to X_j (term in f_j of (1) with a positive coefficient) is nonzero if the concentrations are given by y, for otherwise W would be a semilocking set. Therefore, because all outputs from species X_j (terms in f_j with a negative coefficient) are zero at y, there exist $\epsilon > 0$ and $k > 0$ such that if $x(t) \in \mathbb{R}^m_{>0} \cap B_\epsilon(y)$, then

$$f_j(x(t)) = x_j'(t) > k, \tag{7}$$

where $B_\epsilon(y) = \{x : |x-y| < \epsilon\}$.

Because $f(\cdot)$ is C^1, we have $\|f\|_{\infty,loc} < M$ for some $M > 0$, and this bound is valid in $\mathbb{R}^m_{>0} \cap B_\epsilon(y)$. Therefore, for any $0 < a < b$, if $x(t) \in \mathbb{R}^m_{>0} \cap B_\epsilon(y)$ for $t \in (a,b)$, we have that

$$|x(b) - x(a)| = \left| \int_a^b f(x(s))ds \right| \leq (b-a)M. \tag{8}$$

Now consider a partial trajectory starting on the boundary of $\mathbb{R}^m_{>0} \cap B_\epsilon(y)$ at time t_ϵ, ending on the boundary of $\mathbb{R}^m_{>0} \cap B_{\epsilon/2}(y)$ at time $t_{\epsilon/2}$, and remaining within that annulus for all time in $(t_\epsilon, t_{\epsilon/2})$. Note that one such partial trajectory must exist every time we enter $\mathbb{R}^m_{>0} \cap B_{\epsilon/2}(y)$, and this happens at least once by our assumption that $y \in \omega(\xi) \cap L_W$. By (8), $t_{\epsilon/2} - t_\epsilon \geq \epsilon/(2M)$. On the other hand, by (7), $x_j'(t) > k$ for $t \in (t_\epsilon, t_{\epsilon/2})$. Therefore,

$$\begin{aligned} x_j(t_{\epsilon/2}) &= x_j(t_\epsilon) + \int_{t_\epsilon}^{t_{\epsilon/2}} x_j'(s)ds \\ &\geq x_j(t_\epsilon) + \epsilon k/(2M) \\ &\geq \epsilon k/(2M). \end{aligned}$$

Combining the above with the fact that we still have $x_j'(t) > k$ on $\mathbb{R}^m_{>0} \cap B_{\epsilon/2}(y)$, we see that there cannot exist times t_n such that $x(t_n) \to y$, as $n \to \infty$. This is a contradiction and completes the proof. □

Remark. Theorem 2.5 is a powerful tool for understanding the dynamics of chemical reaction systems. We see that in order to prove that a chemical system is persistent, it is sufficient to show that $[(c+S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W = \emptyset$ for all $c \in \mathbb{R}^m_{>0}$ and all semilocking sets W. We will show in Lemma 2.8 that for many reaction systems such a condition is equivalent to having no equilibria on the boundaries of the positive stoichiometric compatibility classes.

2.2. Deficiency and the deficiency zero theorem. We require one more definition before we can state precisely the types of systems we consider in this paper.

Definition 2.6. *The* deficiency, *δ, of a reaction network $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ is given by $\delta = n - l - s$, where n is the number of complexes of the system, l is the number of linkage classes, and $s = dim S$, the dimension of the stoichiometric subspace.*

Remark. It has been shown that the deficiency of a reaction network is a nonnegative number. In fact, the deficiency is the dimension of a certain subspace associated with the system. See [9, 10, 8] for details.

The main types of systems considered in this paper are those with mass action kinetics that are weakly reversible and have a deficiency of zero. The following theorem by Feinberg [8, 10] is the catalyst for studying such systems. The proof can be found in [8] or [10].

Theorem 2.7 (the deficiency zero theorem). *Consider a system of the form* (1) *with mass action kinetics that is weakly reversible and has a deficiency of zero. Then, within each positive stoichiometric compatibility class there is precisely one equilibrium value, and it is locally asymptotically stable relative to its compatibility class.*

In order to prove that the systems considered in the deficiency zero theorem have equilibria that are locally asymptotically stable relative to their compatibility classes, the Lyapunov function (3) is used. It is shown that for $x \in (\bar{x} + S) \cap \mathbb{R}^m_{>0}$ (where $\bar{x}$ is the equilibrium guaranteed to exist by Theorem 2.7), $V(x) \geq 0$ with equality if and only if $x = \bar{x}$, and that $dV(x(t))/dt < 0$ for all trajectories with initial condition in $(\bar{x}+S) \cap \mathbb{R}^m_{>0}$. We will make use of these facts throughout the paper without reference; however, we point the interested reader to the original works [8, 10] for details.

2.3. Boundary equilibria. The following lemma shows that for weakly reversible, mass action systems with a deficiency of zero, having no equilibria on the boundaries of the positive stoichiometric compatibility classes is equivalent to $[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W=\emptyset$ for all $c\in\mathbb{R}^m_{>0}$ and semilocking sets W. Following Lemma 2.8 we present a theorem pertaining to any system equipped with a globally defined Lyapunov function that does not necessarily go to infinity as x goes to the boundary of the domain. We then use these results in combination with Theorem 2.5 to conclude that for weakly reversible, deficiency zero systems with mass action kinetics, having no equilibria on the boundaries of the positive stoichiometric compatibility classes implies global stability of the equilibrium values given by Theorem 2.7. We again note that it is already known that global asymptotic stability follows from a lack of boundary equilibria. For example, in [15] Sontag showed that all trajectories must converge to the set of equilibria, and so a lack of boundary equilibria implies convergence to the unique equilibrium in the interior of the positive stoichiometric compatibility class. We rederive this result here because our methods put it in a larger context in which global stability is understood through the intersections given in (5) and because it makes clear how our results in section 3 are truly a generalization of this fact.

LEMMA 2.8. *For any chemical reaction system, the set of equilibria on the boundaries of the positive stoichiometric compatibility classes is contained in* $\bigcup_c\bigcup_W[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W$, *where the first union is over* $c\in\mathbb{R}^m_{>0}$ *and the second union is over the semilocking sets. Further, if there are no equilibria on the boundaries of the positive stoichiometric compatibility classes for a weakly reversible, deficiency zero system with mass action kinetics, then* $[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W=\emptyset$ *for all* $c\in\mathbb{R}^m_{>0}$ *and semilocking sets* W.

Proof. Let y be an equilibrium on the boundary of a positive stoichiometric compatibility class. Let W be the set of species with a concentration of zero at y. Because each complex that contains an element of W is providing zero flux, in order for y to be an equilibrium value each reaction in which there is an element of W in the product complex must have an element of W in the reactant complex. Thus, W is a semilocking set and $y\in[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W$ for some $c\in\mathbb{R}^m_{>0}$.

In order to prove the second part of the lemma, we suppose W is a semilocking set for the system and suppose $y\in[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W$ for some $c\in\mathbb{R}^m_{>0}$. We will now produce an equilibrium value on the boundary of $(c+S)\cap\mathbb{R}^m_{>0}$. If $W=\mathcal{S}$, $y=\vec{0}$, and, because W is a semilocking set, y is an equilibrium. Otherwise, consider the system consisting only of those species not in the semilocking set W. By the arguments in [10], the linkage classes not "locked" by W form their own weakly reversible, deficiency zero system. Therefore, there is an equilibrium value with strictly positive components for that reduced system, $\bar{z}$. Let $\bar{y}=(\bar{z},\vec{0})$ (where we have potentially rearranged the ordering of the species so that those not in the semilocking set came first). $\bar{y}$ is a boundary equilibrium value to our original system. Therefore, the result is shown. □

THEOREM 2.9. *Let* $x(t)=x(t,x(0))$ *be the solution to* $\dot{x}=g(x)$ *with initial condition* $x(0)$, *where* g *is* C^1 *and the domain of definition of the system is the open set* $C\subset\mathbb{R}^m$. *Let* $\bar{x}\in C$ *be the unique equilibrium value to the system. Finally, suppose that there is a globally defined Lyapunov function* V *that satisfies the following:*

1. $V(x)\geq 0$ *with equality if and only if* $x=\bar{x}$.
2. $dV(x(t))/dt\leq 0$ *with equality if and only if* $x(t)=\bar{x}$.
3. $V(x)\to\infty$, *as* $|x|\to\infty$.

Then either $x(t)\to\bar{x}$ *or* $x(t)\to\partial C$.

Proof. Suppose that $x(t) \nrightarrow \bar{x}$. Because $V(\cdot)$ decreases along trajectories, the value $V(x(t))$ is bounded above by $V(x(0))$ for all $t > 0$. Therefore, because $V(x) \to \infty$ as $|x| \to \infty$, $x(t)$ remains bounded for all $t > 0$. Also, the local asymptotic stability of $\bar{x}$ combined with the fact that $x(t) \nrightarrow \bar{x}$ implies there is a $\rho > 0$ such that $|x(t) - \bar{x}| > \rho$ for all $t > 0$.

Let $\epsilon > 0$ and for $x \in C$ let $d(x, \partial C)$ represent the distance from x to the boundary of C. Let $C_\epsilon = \{x \in C \mid d(x, \partial C) \geq \epsilon \text{ and } |x - \bar{x}| \geq \rho\}$. Using that trajectories remain bounded for all time, we may use the continuity of the functions ∇V and g to conclude that there is a positive number $\eta = \eta(\epsilon)$ such that $\nabla V(x) \cdot g(x) < -\eta$ for all $x \in C_\epsilon$. Therefore, the amount of time that any trajectory spends in the set C_ϵ is bounded above by $V(x(0))/\eta$ (for, otherwise, $x(t) \to \bar{x}$). Because $\epsilon > 0$ was arbitrary we see that $x(t) \to \partial C$. □

COROLLARY 2.10. *If there are no equilibria on the boundaries of the positive stoichiometric compatibility classes for a weakly reversible deficiency zero system with mass action kinetics, then the unique positive equilibrium value within each positive stoichiometric compatibility class is globally asymptotically stable relative to its compatibility class.*

Proof. Trajectories are bounded because of the existence of the Lyapunov function (3). Combining this fact with Theorems 2.5 and 2.9 and Lemma 2.8 completes the proof. □

3. Main results. By Lemma 2.8, we see that the no boundary equilibria assumption for weakly reversible deficiency zero systems with mass action kinetics is equivalent to the assumption that $[(c + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W = \emptyset$ for all $c \in \mathbb{R}^m_{>0}$ and all semilocking sets W. This then implies global stability by Corollary 2.10. We will extend these results by proving that global stability holds if $[(c + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$ is empty *or* discrete for each $c \in \mathbb{R}^m_{>0}$ and each semilocking set W. The following definition is necessary.

DEFINITION 3.1. *For a vector $x \in \mathbb{R}^m$, the* support *of x, denoted* $\text{supp}(x)$*, is the subset of the species such that $X_i \in \text{supp}(x)$ if and only if $x_i \neq 0$.*

PROPOSITION 3.2. *Let $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ be a weakly reversible, deficiency zero, mass action chemical reaction system with dynamics given by* (1). *Suppose that $y \in \omega(x(0))$ for some $x(0) \in \mathbb{R}^m_{>0}$. Then there must exist a nonzero $z_0 \in S$ with* $\text{supp}(z_0) \subset \text{supp}(y)$.

Proof. If $y \in \mathbb{R}^m_{>0}$, there is nothing to show. Therefore, assume that y is on the boundary of the positive stoichiometric compatibility class. By Theorem 2.5, there is a semilocking set W such that $y \in [(x(0) + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$.

Let $V(x) : \mathbb{R}^m_{>0} \to \mathbb{R}$ be given by (3), and let

$$V_i(x_i) = x_i(\ln(x_i) - \ln(\bar{x}_i) - 1) + \bar{x}_i,$$

so that $V(x) = \sum_{i=1}^m V_i(x_i)$. Reordering the species if necessary, we suppose $W = \{X_1, \dots, X_d\}$. Choose $\rho > 0$ so small that for each $i \leq d$, $x_i < \rho \Longrightarrow \ln(x_i) - \ln(\bar{x}_i) < 0$. Let $\epsilon > 0$ satisfy $\epsilon < \rho$. Let t_ϵ be a time such that $x_i(t_\epsilon) \leq \epsilon$ for all $i \leq d$ and $|x_j(t_\epsilon) - y_j| < \epsilon$ for all $j \geq d + 1$. Let $T_\epsilon = \min\{t > t_\epsilon : |x_i(t) - y_i| \leq x_i(t_\epsilon)/2 \text{ for all } i \leq m\}$. We know such t_ϵ and T_ϵ exist because y is an ω-limit point of the system. Note that $T_\epsilon > t_\epsilon$ and that for each $i \leq d$, $x_i(T_\epsilon) < x_i(t_\epsilon)$. We consider how $V(x(t))$ changes from time t_ϵ to time T_ϵ. Applying the mean value theorem to each $V_i(\cdot)$ term gives

$$V(x(T_\epsilon)) - V(x(t_\epsilon)) = \sum_{i=1}^{m} V_i(x_i(T_\epsilon)) - V_i(x_i(t_\epsilon)) \tag{9}$$

$$\begin{aligned} &= \sum_{i=1}^{d} (\ln(\tilde{x}_i) - \ln(\bar{x}_i))(x_i(T_\epsilon) - x_i(t_\epsilon)) \\ &\quad + \sum_{i=d+1}^{m} (\ln(\tilde{x}_i) - \ln(\bar{x}_i))(x_i(T_\epsilon) - x_i(t_\epsilon)) \end{aligned} \tag{10}$$

for some $\tilde{x}_i \in [x_i(T_\epsilon), x_i(t_\epsilon)]$. Recalling that V decreases along trajectories of $x(t)$ by Theorem 2.7, we have $V(x(T_\epsilon)) - V(x(t_\epsilon)) < 0$. Note that because for $j \geq d+1$ we have $|\tilde{x}_j - y_j| < \epsilon$, there are positive constants c_j such that $c_j > |\ln(\tilde{x}_j) - \ln \bar{x}_j|$, and that bound is valid *for any* $\epsilon < \rho$. Let $C = \sum_{j=d+1}^{m} c_j$.

By our choices above, we know that for each $i \in \{1, \ldots, d\}$ the following inequalities hold:

1. $\ln(\tilde{x}_i) - \ln(\bar{x}_i) < 0$.
2. $x_i(T_\epsilon) - x_i(t_\epsilon) < 0$.

Therefore, each piece of the first sum in (10) is strictly positive. Thus, to ensure that V is decreasing along this trajectory, the second sum in (10) must be negative and, letting $\Delta x_i = x_i(T_\epsilon) - x_i(t_\epsilon)$ for each i, we have

$$\begin{aligned} \sum_{i=1}^{d} (\ln(\tilde{x}_i) - \ln(\bar{x}_i))\, \Delta x_i &< \left| \sum_{j=d+1}^{m} (\ln(\tilde{x}_j) - \ln(\bar{x}_j))\, \Delta x_j \right| \\ &\leq \sum_{j=d+1}^{m} c_j |\Delta x_j|. \end{aligned} \tag{11}$$

In fact, because each term on the left-hand side of (11) is positive, a similar inequality must hold for each $i = 1, \ldots, d$. That is, for $i \leq d$

$$(\ln(\tilde{x}_i) - \ln(\bar{x}_i))\Delta x_i \leq \sum_{j=d+1}^{m} c_j |\Delta x_j|.$$

For each $i \leq d$, $\tilde{x}_i \in [x_i(T_\epsilon), x_i(t_\epsilon)]$ and $x_i(T_\epsilon), x_i(t_\epsilon) < \epsilon$. Hence, letting $|\ln \bar{x}_i| = k_i$, we have that for each $i \leq d$

$$|\ln(\tilde{x}_i) - \ln(\bar{x}_i)| \geq |\ln \epsilon| - k_i.$$

Thus, for each $i = 1, \ldots, d$,

$$|\Delta x_i| \leq \frac{1}{|\ln \epsilon| - k_i} \sum_{j=d+1}^{m} c_j |\Delta x_j|.$$

Let $\Delta_{max} = \sup_{j \in \{d+1,\ldots,m\}} \{|\Delta x_j|\}$ and $\delta(\epsilon) = \sup_{i \in \{1,\ldots,d\}} (|\ln \epsilon| - k_i)^{-1}$. We know $\Delta_{max} \neq 0$ because if it were equal to zero, then the right-hand side of (11) would be zero, which it cannot be as it is strictly larger than the left-hand side. Combining the above shows that for each $i = 1, \ldots, d$,

$$|\Delta x_i| \leq \delta(\epsilon) C \Delta_{max}.$$

Now we consider the vector $\Delta x = x(T_\epsilon) - x(t_\epsilon) \in S$. Normalizing the vector Δx by dividing each entry by Δ_{max} then produces a vector $v(\epsilon) \doteq \frac{1}{\Delta_{max}} \Delta x$ with the following properties:

1. $v(\epsilon) \in S$.
2. For $i = 1, \dots, d$, $|v_i(\epsilon)| \leq \delta(\epsilon)C$.
3. There is at least one entry in $v(\epsilon)$ with norm 1 (the one for which the maximum in the definition of Δ_{max} was achieved), and none has a higher norm.
4. $1 \leq |v(\epsilon)| \leq m$.

Property 4 follows from property 3. $\epsilon > 0$ was arbitrary, so we may consider a sequence $\{\epsilon_n\}$ such that $\epsilon_n > \epsilon_{n+1}$ and $\epsilon_n \to 0$. For each ϵ_n we may redo the work above. This leads to a sequence of vectors $\{v(\epsilon_n)\}$ and a sequence of numbers $\{\delta(\epsilon_n)\}$ such that $\delta(\epsilon_n) \to 0$ and for each n all four properties above hold. Because each vector from the sequence $\{v(\epsilon_n)\}$ is contained in the compact space $\{v : 1 \leq |v| \leq m\} \cap S$, there is a convergent subsequence $\{v(\epsilon_{n_k})\}$ and a vector z_0 such that $v(\epsilon_{n_k}) \to z_0 \in \{v : 1 \leq |v| \leq m\} \cap S \subset S$, as $k \to \infty$. Note that z_0 cannot be the zero vector because $|z_0| > 1$. However, $\delta(\epsilon_{n_k}) \to 0$, and so the first d components of z_0 are equal to zero. Hence, $\text{supp}(z_0) \subset \text{supp}(y)$. □

We now present our main result.

Theorem 3.3. *Let $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ be a weakly reversible, deficiency zero, mass action chemical reaction system with dynamics given by* (1). *Suppose that for each $c \in \mathbb{R}^m_{>0}$ and each semilocking set W, the set $[(c+S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$ is either empty or discrete. Then the unique positive equilibrium of each stoichiometric compatibility class guaranteed to exist by the deficiency zero theorem is globally asymptotically stable relative to its compatibility class.*

Proof. We suppose, in order to find a contradiction, that there is a positive equilibrium, $\bar{x}$, that is not globally asymptotically stable relative to its compatibility class. By Theorems 2.5, 2.7, and 2.9, there is a semilocking set W, a $\xi \in \mathbb{R}^m_{>0}$, and a vector y such that $y \in [(\bar{x} + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$ and $y \in \omega(\xi)$. By Proposition 3.2, there exists a nonzero $z_0 \in S$ such that $\text{supp}(z_0) \subset \text{supp}(y)$. Because $y \in \bar{x} + S$ and $z_0 \in S$, for any $\eta > 0$ we have $y + \eta z_0 \in \bar{x} + S$. Further, because $\text{supp}(z_0) \subset \text{supp}(y)$, if η is small enough, we have that $y + \eta z_0 \in \mathbb{R}^m_{\geq 0} \cap L_W$. But this is valid for all η small enough, and so $[(\bar{x} + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$ is not discrete. This is a contradiction, and so the result is shown. □

Remark. In the chemistry literature there is a notion of an equilibrium value being *complex balanced*. Briefly, an equilibrium value is complex balanced if, at equilibrium, the total flux out of any complex is equal to the total flux into that complex. The conclusion of the deficiency zero theorem, including the existence of the Lyapunov function (3), holds so long as there is at least one equilibrium in the positive orthant that is complex balanced [8, 11]. The deficiency zero theorem gives a simple and checkable condition ($\delta = 0$) on the network structure alone that guarantees that a system has a complex balanced equilibrium for any choice of rate constants. We therefore note that the conclusion of Theorem 3.3 holds for any mass action system that has a complex balanced equilibrium in the positive orthant and for which $[(c + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$ is either empty or discrete for each $c \in \mathbb{R}^m_{>0}$ and semilocking set W.

Corollary 3.4. *Suppose that for a weakly reversible, deficiency zero, chemical reaction system with mass action kinetics, each semilocking set is a locking set. Suppose further that within each stoichiometric compatibility class, the set of equilibria on the boundary is discrete. Then the unique positive equilibrium of each stoichiometric compatibility class guaranteed to exist by the deficiency zero theorem is globally asymptotically stable relative to its compatibility class.*

Proof. Because each semilocking set is a locking set, the set of boundary equilibria for a given compatibility class is precisely given by $\bigcup_W [(c + S) \cap \mathbb{R}^m_{\geq 0}] \cap L_W$, where the union is over the set of semilocking sets. Therefore, each $[(c+S)\cap\mathbb{R}^m_{\geq 0}]\cap L_W$ is discrete, and invoking Theorem 3.3 completes the proof. □

Corollary 3.5. *Suppose that a weakly reversible, deficiency zero, chemical reaction system with mass action kinetics has only one linkage class. Suppose further that within each stoichiometric compatibility class, the set of equilibria on the boundary is discrete. Then the unique positive equilibrium of each stoichiometric compatibility class guaranteed to exist by the deficiency zero theorem is globally asymptotically stable relative to its compatibility class.*

Proof. For single linkage class systems that are weakly reversible, each semilocking set is a locking set. Using Corollary 3.4 completes the proof. □

3.1. Connection with extreme points.

We connect our results to a condition on the extreme points of the positive stoichiometric compatibility classes.

Proposition 3.6. *For $y\in\mathbb{R}^m_{\geq 0}$, let $W=\{X_i : y_i=0\}=\operatorname{supp}(y)^C$. Then the following are equivalent:*

(i) *y is an extreme point of $(y+S)\cap\mathbb{R}^m_{\geq 0}$.*

(ii) *$\left[(y+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W=\{y\}$.*

Proof. (i) $\Longrightarrow$ (ii). Suppose that (i) is true, but (ii) is not. Then, because $\left[(y+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W$ is not discrete, there exists a $v\in S\bigcap L_W$ such that for sufficiently small ϵ, $y\pm\epsilon v\in\left[(y+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W$. Noting that $y=(1/2)(y+\epsilon v)+(1/2)(y-\epsilon v)$ then shows that y is not an extreme point of $(y+S)\cap\mathbb{R}^m_{\geq 0}$, which is a contradiction.

(ii) $\Longrightarrow$ (i). Suppose that (ii) is true, but (i) is not. Because y is not an extreme point of $(y+S)\cap\mathbb{R}^m_{\geq 0}$, there exist nonzero vectors $v_1\neq y$, $v_2\neq y$ in $(y+S)\bigcap\mathbb{R}^m_{\geq 0}$ and $0<\lambda<1$ such that

$$y=\lambda v_1+(1-\lambda)v_2. \tag{12}$$

Because $v_1,v_2\in(y+S)\bigcap\mathbb{R}^m_{\geq 0}$, there exist $u,w\in S$ such that $v_1=y+u$ and $v_2=y+w$, and $u_i,w_i\geq 0$ if $X_i\in W$. However, because $\lambda, 1-\lambda>0$, we see by (12) that $u_i,w_i=0$ for all $X_i\in W$. Therefore, $v_1,v_2\in\left[(y+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W$, contradicting (ii). □

Theorem 3.3 can now be reformulated in the following way.

Theorem 3.7. *Let $\{\mathcal{S},\mathcal{C},\mathcal{R}\}$ be a weakly reversible, deficiency zero, mass action chemical reaction system with dynamics given by* (1). *For any boundary point $y\in[(c+S)\cap\mathbb{R}^m_{\geq 0}]$ of a positive stoichiometric compatibility class, let $W_y=\{X_i : y_i=0\}=\operatorname{supp}(y)^C$. Finally, suppose that W_y is a semilocking set only if y is an extreme point. Then the unique positive equilibrium of each stoichiometric compatibility class guaranteed to exist by the deficiency zero theorem is globally asymptotically stable relative to its compatibility class.*

Proof. Suppose W is a semilocking set. Let $y\in L_W$ be such that there exists a $c\in\mathbb{R}^m_{>0}$ with $y\in c+S$. If no such y and c exist, then $[(c+S)\bigcap\mathbb{R}^m_{\geq 0}]\bigcap L_W=\emptyset$ for all $c\in\mathbb{R}^m_{>0}$. If such a y and c do exist, then $W=W_y$ and, by assumption, y is an extreme point. Thus, by Proposition 3.6, $\left[(y+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W=\left[(c+S)\bigcap\mathbb{R}^m_{\geq 0}\right]\bigcap L_W$ is discrete. Invoking Theorem 3.3 completes the proof. □

4. Examples.

We begin with an example found in [6] for a receptor-ligand model. See [6] for full details.

Example 4.1. Consider the following system, which we assume has mass action kinetics:

$$\begin{tikzcd} 2A+C \arrow[r, shift left] \arrow[d, shift left] & A+D \arrow[l, shift left] \arrow[d, shift left] \\ B+C \arrow[u, shift left] \arrow[r, shift right] & E \arrow[l, shift right] \arrow[u, shift left] \end{tikzcd} . \tag{13}$$

For this example there are four complexes and one linkage class, and the dimension of the stoichiometric compatibility class is easily verified to be three. Therefore, the system has a deficiency of zero, and our results apply. The minimal semilocking sets (that is, those that must be contained in all others) are given by $W_1 = \{A, B, E\}$, $W_2 = \{A, C, E\}$, and $W_3 = \{C, D, E\}$. Therefore, showing that the set $\bigcup_{i=1}^{3} [(c + S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_i}$ is discrete for any $c \in \mathbb{R}^5_{>0}$ would also show that the sum over all semilocking sets is discrete. For this example, it is easily verified that

$$S = \text{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{bmatrix} \right\}. \tag{14}$$

One method to show that for a given i the set $[(c+S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_i}$ is at most discrete is to demonstrate that there are no nonzero vectors contained in S with support given by W_i^C. This method bypasses the need to check whether the intersection $[(c + S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_i}$ is nonempty. It is easily verified that there are no nonzero vectors contained in S with support given by W_1^C or W_2^C. In [6] it is shown that for each $c \in \mathbb{R}^5_{>0}$, $(c + S) \cap \mathbb{R}^5_{\geq 0}$ does intersect one of L_{W_1} or L_{W_2}. Therefore, there are always equilibria on the boundary; however, by our results or those found in [6], they will not affect the global stability of the interior equilibria.

Let $U_3 = \{x \in \mathbb{R}^5 \mid \text{supp}(x) \in W_3^C\}$. It is easy to show that $U_3 \bigcap S = \text{span}\{[2, -1, 0, 0, 0]^T\}$. Thus, the method used in the previous paragraph does not work. Therefore, for our results to apply, we need to verify that $[(c+S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_3} = \emptyset$ for any $c \in \mathbb{R}^5_{>0}$. Because L_{W_3} is characterized by having the last three entries equal to zero, in order to prove that $[(c+S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_3} = \emptyset$ for any $c \in \mathbb{R}^5_{>0}$, it is sufficient to show that the space spanned by the last three components of the vectors in (14) does not contain a vector with strictly negative components. We have

$$\text{Span}\left\{ \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \right\} = \text{Span}\left\{ \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}, \tag{15}$$

which does not include a strictly negative vector. Thus, $[(c+S) \cap \mathbb{R}^5_{\geq 0}] \cap L_{W_3} = \emptyset$ for any $c \in \mathbb{R}^5_{>0}$. Combining all of the above with Theorem 3.3 shows that for any choice of rate constants and initial condition, the system (13) has a globally asymptotically stable equilibrium value.

Example 4.2. Consider the system

$$2A \quad \rightleftarrows \quad A + B \quad \rightleftarrows \quad B + C. \tag{16}$$

There are three complexes and one linkage class, and the dimension of the stoichiometric compatibility class is two. Therefore, the system (16) has a deficiency of zero. The minimal semilocking sets are $W_1 = \{A, B\}$ and $W_2 = \{A, C\}$. The stoichiometric subspace is of dimension two and the quantity $A+B+C$ is conserved. Thus, each stoichiometric compatibility class is a plane that intersects each of $L_{W_1} = \{v : v_1 = v_2 = 0, v_3 \neq 0\}$ and $L_{W_2} = \{v : v_1 = v_3 = 0, v_2 \neq 0\}$ in precisely one point. See Figure 4.1. Therefore, by Theorem 3.3, for any choice of rate constants and initial condition, the system (16) has a globally asymptotically stable equilibrium value. We note that it is easily verified that the eigenvalues of the linearized problem around the equilibria

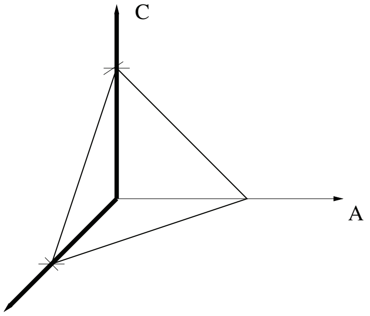


FIG. 4.1. *The stoichiometric compatibility class for the system* (16) *is a plane defined by the conservation law* $A+B+C=M$ *for some* $M>0$. *This plane intersects the sets* L_{W_1} *and* L_{W_2} *(bolded axes) each in precisely one point.*

associated with the semilocking set W_1 are all zero, and so the results of [6] do not apply here.

Example 4.3. Consider the system

$$2A \rightleftarrows A+B \qquad 2B \rightleftarrows A+C. \tag{17}$$

There are four complexes and two linkage classes, and the dimension of the stoichiometric compatibility class is two. Therefore, the system (17) has a deficiency of zero. The only minimal semilocking set is $W=\{A,B\}$, and this is also a locking set. The stoichiometric subspace is of dimension two, and the quantity $A+B+C$ is conserved. Thus, each stoichiometric compatibility class is a plane that intersects $L_W=\{v : v_1=v_2=0,\, v_3\neq 0\}$ in precisely one point. Therefore by Theorem 3.3 or Corollary 3.4, for any choice of rate constants and initial condition, the system (17) has a globally asymptotically stable equilibrium value. It is easily verified that the boundary equilibria are not hyperbolic with respect to their compatibility classes, and so the results of [6] do not apply in this case.

5. Non-mass action kinetics. In [15], Sontag extended the deficiency zero theorem to systems with non-mass action kinetics. He considered weakly reversible, deficiency zero system whose kinetic functions are given by

$$R_{y\to y'}(x)=k_{y\to y'}\theta(x_1)^{y_1}\cdots\theta(x_m)^{y_m}, \tag{18}$$

where the functions $\theta_i:\mathbb{R}\to[0,\infty)$ satisfy the following:

1. Each θ_i is locally Lipschitz.
2. $\theta_i(0)=0$.
3. $\int_0^1|\ln(\theta_i(y))|dy<\infty$.
4. The restriction of θ_i to $\mathbb{R}_{\geq 0}$ is strictly increasing and onto $\mathbb{R}_{\geq 0}$.

To prove the local asymptotic stability of the unique equilibrium within each stoichiometric compatibility class, the following Lyapunov function was used:

$$V(x)=\sum_{i=1}^{m}\int_{\bar{x}_i}^{x_i}\left(\rho_i(s)-\rho_i(\bar{x}_i)\right)ds, \tag{19}$$

where $\rho_i(s)=\ln\theta_i(s)$ and $\bar{x}$ is the unique equilibrium within the positive stoichiometric compatibility class. Note that $\theta(x)=|x|$ gives mass action kinetics, in which case the Lyapunov function given in (19) is the same as that in (3). The only dynamical property of the deficiency zero theorem used in this paper is that $\nabla V(x) \to -\infty$ as $x_i \to 0$. We note that for the Lyapunov function (19), we still have that property because

$$\nabla V(x) = \sum_{i=1}^{m} \rho_i(x_i) - \rho_i(\bar{x}_i),$$

and $\rho_i(x_i) = \ln \theta_i(x_i) \to -\infty$ as $x_i \to 0$ by the properties of $\theta_i(\cdot)$ given above. Therefore, our results in this paper, and in particular Theorem 3.3, Corollary 3.4, Corollary 3.5, and Theorem 3.7, are valid in the setting (18).

Acknowledgments. I would like to thank Gheorghe Craciun for several illuminating conversations pertaining to chemical reaction systems. I would also like to thank Martin Feinberg for pointing out the connection with extreme points.

REFERENCES

[1] D. F. Anderson, *Stochastic Perturbations of Biochemical Reaction Systems*, Ph.D. thesis, Duke University, Durham, NC, 2005.

[2] D. Angeli, P. De Leenheer, and E. D. Sontag, *A Petri net approach to the study of persistence in chemical reaction networks*, Math. Biosci., 210 (2007), pp. 598–618.

[3] G. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.

[4] G. Butler and P. Waltman, *Persistence in dynamical systems*, J. Differential Equations, 63 (1986), pp. 255–263.

[5] T. C. Card, *Persistence in food webs with general interactions*, Math. Biosci., 51 (1980), pp. 165–174.

[6] M. Chavez, *Observer Design for a Class of Nonlinear Systems, with Applications to Biochemical Networks*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 2003.

[7] M. Feinberg, *Complex balancing in general kinetic systems*, Arch. Rational Mech. Anal., 49 (1972), pp. 187–194.

[8] M. Feinberg, *Lectures on Chemical Reaction Networks*, delivered at the Mathematics Research Center, University Wisconsin-Madison, 1979; available at http://www.che.eng.ohio-state.edu/~feinberg/LecturesOnReactionNetworks.

[9] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors.* I. *The deficiency zero and deficiency one theorems*, Chem. Eng. Sci., 42 (1987), pp. 2229–2268 (review article 25).

[10] M. Feinberg, *Existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Rational Mech. Anal., 132 (1995), pp. 311–370.

[11] J. Gunawardena, *Chemical Reaction Network Theory for In-Silico Biologists*, http://vcp.med.harvard.edu/papers/crnt.pdf (20 June 2003).

[12] F. J. M. Horn and R. Jackson, *General mass action kinetics*, Arch. Rational Mech. Anal., 47 (1972), pp. 81–116.

[13] D. Siegel and Y. F. Chen, *Global stability of deficiency zero chemical networks*, Canad. Appl. Math. Quart., 2 (1994), pp. 413–434.

[14] D. Siegel and D. MacLean, *Global stability of complex balanced mechanisms*, J. Math. Chem., 27 (2000), pp. 89–110.

[15] E. D. Sontag, *Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction*, IEEE Trans. Automat. Control, 46 (2001), pp. 1028–1047.

[16] H. R. Thieme, *Uniform persistence and permanence for non-autonomous semiflows in population biology*, Math. Biosci., 166 (2000), pp. 173–201.

SIAM J. Appl. Math.
Vol. 68, No. 5, pp. 1477–1494

VOLTAGE AND CURRENT EXCITATION FOR TIME-HARMONIC EDDY-CURRENT PROBLEMS[*]

ANA ALONSO RODRÍGUEZ[†] and ALBERTO VALLI[†]

Abstract. We give a systematic presentation of voltage or current intensity excitation for time-harmonic eddy-current problems. The key point of our approach resides in a suitable power law that permits us to understand the role of voltage excitation. We also shed light on the influence of the boundary conditions on the proposed formulations.

Key words. eddy-current problems, voltage excitation, current excitation, variational formulation, finite elements

AMS subject classifications. 35Q60, 65N30, 35F15, 35R05, 78A25

DOI. 10.1137/070697677

1. Introduction and basic results.

In many electromagnetic phenomena it is useful to couple formulations in terms of electrical circuits with more general formulations based on Maxwell equations (or else on some reduced model like the eddy-current system).

This coupling is often performed by transforming some data like voltage or current intensity into suitable boundary conditions for the electromagnetic fields. In particular, it is interesting to devise efficient formulations of the eddy-current problem when the only excitation present is either an assigned voltage (typically, at contacts) or a given current intensity in the eddy-current region.

On the other hand, it is well known that the topological properties of the conductor and the type of boundary conditions imposed on the boundary of the computational domain have a strong influence on the general setting of the problem and on the structure of the solution.

Several possible approaches have been proposed in recent years, especially by engineers interested in practical computations. Let us mention only the contributions in [13], [22], [20], [21], [28], [29], [26], [19], [12], [23], [10], [8], and the references therein, though this list is far from complete.

In this paper we propose a systematic approach to eddy-current problems driven by voltage or current intensity. Our aim is to give a general mathematical formulation for these problems and to analyze their well-posedness. These theoretical results are then the basis for devising stable and convergent finite element approximation schemes.

A typical difficulty is that, in many situations, eddy-current problems are well-posed *even if* no additional condition like voltage or current intensity is imposed. As a consequence, to overcome this apparent contradiction it is necessary to focus on the modeling of the problem so that it becomes possible to impose the voltage or current intensity equation, but without giving up Maxwell equations (a flaw that was present in previous papers on this subject).

[*]Received by the editors July 18, 2007; accepted for publication (in revised form) February 12, 2008; published electronically May 16, 2008.
http://www.siam.org/journals/siap/68-5/69767.html

[†]Dipartimento di Matematica, Università di Trento, via Sommarive 14, I-38050 Povo (Trento), Italy (alonso@science.unitn.it, valli@science.unitn.it).

In the rest of this section we introduce notation and describe the problems we shall consider, and we present two basic results concerning well-posedness. In section 2 we discuss modeling, basing our argument on a global power law that relates voltage to current intensity. In the third section our proposal for treating voltage and current excitation problems is described. In section 4 we systematically present the variational formulations of these problems. Finally, the last section is devoted to giving a short description of some numerical approximation schemes based on finite elements.

In the following, for the sake of simplicity we assume that the domain $\Omega \subset \mathbb{R}^3$ is a simply connected bounded open set, with a connected boundary $\partial\Omega$. It is composed of two parts, a conductor Ω_C and an insulator Ω_I. The interface between Ω_C and Ω_I will be denoted by Γ. The unit outward normal vector on $\partial\Omega$ will be indicated by $\mathbf{n}$, while the unit normal vector on Γ, directed toward Ω_I, will be denoted by $\mathbf{n}_C = -\mathbf{n}_I$.

It is well known that (see, e.g., [15]) the time-harmonic eddy-current problem is given by Ampère and Faraday equations

$$\begin{aligned} \operatorname{curl} \mathbf{H} - \boldsymbol{\sigma}\mathbf{E} &= \mathbf{J}_e \quad \text{in } \Omega, \\ \operatorname{curl} \mathbf{E} + i\omega\boldsymbol{\mu}\mathbf{H} &= \mathbf{0} \quad \text{in } \Omega, \end{aligned}$$

where $\mathbf{H}$ and $\mathbf{E}$ are the magnetic and electric fields, respectively, $\mathbf{J}_e$ is the given electric current density, $\boldsymbol{\sigma}$ is the electric conductivity, $\boldsymbol{\mu}$ is the magnetic permeability, and $\omega \neq 0$ is a given angular frequency. Moreover, suitable boundary conditions have to be added (and also some conditions for the unique determination of the electric field in Ω_I). For a mathematical justification of the complete eddy-current model we refer the reader to [7] (see also [1], [18]).

Let us also note that the magnetic permeability $\boldsymbol{\mu}$ is assumed to be a symmetric tensor, uniformly positive definite in Ω, with entries in $L^\infty(\Omega)$. The same assumption holds for the dielectric coefficient $\boldsymbol{\epsilon}$ in Ω_I (this coefficient will come into play when imposing uniqueness conditions for $\mathbf{E}_I$), and for the electric conductivity $\boldsymbol{\sigma}$ in Ω_C; on the other hand, $\boldsymbol{\sigma}$ vanishes outside Ω_C.

We will distinguish between two different geometric situations and three different types of boundary conditions.

First geometric case: Electric ports. The conductor Ω_C is not strictly contained in Ω, namely, $\partial\Omega_C \cap \partial\Omega \neq \emptyset$. More precisely, for the sake of simplicity we assume that Ω_C is a simply connected domain with $\partial\Omega_C \cap \partial\Omega = \Gamma_E \cup \Gamma_J$, where Γ_E and Γ_J are connected and disjoint surfaces on $\partial\Omega$ ("electric ports"). Therefore, we have $\partial\Omega_C = \Gamma_E \cup \Gamma_J \cup \Gamma$. The boundary of the insulator Ω_I, which is connected, is given by $\partial\Omega_I = \Gamma_D \cup \Gamma$, where $\Gamma_D \subset \partial\Omega$. As a consequence, we have $\partial\Gamma = \partial\Gamma_D = \partial\Gamma_E \cup \partial\Gamma_J$ (see Figure 1, left).

Second geometric case: Internal conductor. The conductor Ω_C is strictly contained in Ω, namely, $\partial\Omega_C \cap \partial\Omega = \emptyset$. Moreover, for the sake of simplicity we assume that Ω_C is a torus-like domain. In this case, we simply have $\partial\Omega_C = \Gamma$ and $\partial\Omega_I = \partial\Omega \cup \Gamma$ (see Figure 1, right).

First set of boundary conditions. These are given by $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$ for both the geometric cases.

Second set of boundary conditions. These are given by $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\Gamma_E \cup \Gamma_J$ and $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on Γ_D for the electric port case, and by $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on $\partial\Omega$ for the internal conductor case.

Third set of boundary conditions. These are given by $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\Gamma_E \cup \Gamma_J$, $\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on Γ_D for the electric port case, and by $\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on $\partial\Omega$ for the internal conductor case.

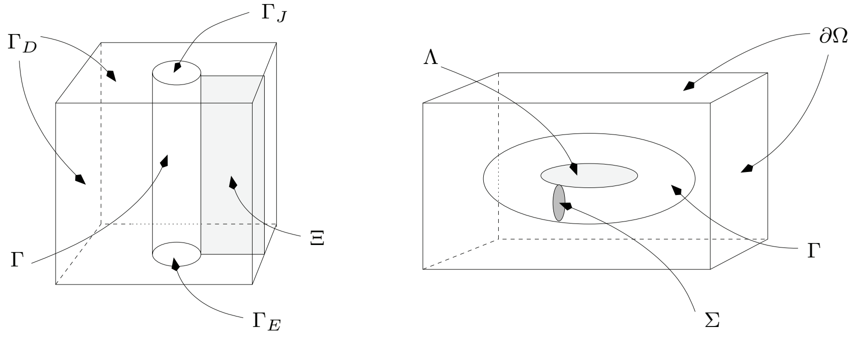


FIG. 1. *The geometry of the domain for the electric port case (left) and for the internal conductor case (right).*

Summing up, we are going to consider six different problems:

Case A. Electric ports, $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$.

Case B. Electric ports, $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\Gamma_E \cup \Gamma_J$, $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on Γ_D.

Case C. Electric ports, $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\Gamma_E \cup \Gamma_J$, $\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on Γ_D.

Case D. Internal conductor, $\mathbf{E} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$.

Case E. Internal conductor, $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on $\partial\Omega$.

Case F. Internal conductor, $\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$ and $\boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0$ on $\partial\Omega$.

Among the six boundary value problems described here above, Case C has some specific features. In fact, it is the only one for which the solution of the eddy-current problem is not unique.

Let us start by proving this result.

THEOREM 1.1. *Let us consider the solutions* $\mathbf{H}$ *and* $\mathbf{E}$ *of the eddy-current problem*

$$\begin{aligned} \operatorname{curl} \mathbf{H} - \boldsymbol{\sigma}\mathbf{E} &= \mathbf{J}_e \quad && \text{in } \Omega, \\ \operatorname{curl} \mathbf{E} + i\omega\boldsymbol{\mu}\mathbf{H} &= \mathbf{0} && \text{in } \Omega. \end{aligned}$$

The magnetic field $\mathbf{H}$ *in* Ω *and the electric field* $\mathbf{E}_C$ *in* Ω_C *are uniquely determined for each one of the set of boundary conditions described in Cases* A, B, D, E, *and* F.

Proof. Assume that $\mathbf{J}_e = \mathbf{0}$ in Ω. Multiply the Faraday equation by $\overline{\mathbf{H}}$ (the complex conjugate of $\mathbf{H}$) and integrate in Ω. Integration by parts gives

$$0 = \int_\Omega \operatorname{curl} \mathbf{E} \cdot \overline{\mathbf{H}} + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} = \int_\Omega \mathbf{E} \cdot \operatorname{curl} \overline{\mathbf{H}} + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} + \int_{\partial\Omega} \mathbf{n} \times \mathbf{E} \cdot \overline{\mathbf{H}}.$$

Replacing $\mathbf{E}_C$ by $\boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C$, and recalling that $\operatorname{curl} \mathbf{H}_I = \mathbf{0}$ in Ω_I, we have

$$0 = \int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} + \int_{\partial\Omega} \mathbf{n} \times \mathbf{E} \cdot \overline{\mathbf{H}}.$$

Thus the uniqueness result follows at once if we prove that the boundary integral vanishes.

This is clearly the case if we are considering Cases A, B, D, and E. For Case F, we have $\operatorname{div}_\tau(\mathbf{E} \times \mathbf{n}) = \operatorname{curl} \mathbf{E} \cdot \mathbf{n} = -i\omega\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$ on $\partial\Omega$; hence there exists a scalar function W such that $\mathbf{E} \times \mathbf{n} = \operatorname{grad} W \times \mathbf{n}$ on $\partial\Omega$. Therefore,

$$\begin{aligned} \textstyle\int_{\partial\Omega} \mathbf{n} \times \mathbf{E} \cdot \overline{\mathbf{H}} &= \textstyle\int_{\partial\Omega} \overline{\mathbf{H}} \times \mathbf{n} \cdot \operatorname{grad} W = -\int_{\partial\Omega} \operatorname{div}_\tau(\overline{\mathbf{H}} \times \mathbf{n})\, W \\ &= \textstyle -\int_{\partial\Omega} \operatorname{curl} \overline{\mathbf{H}} \cdot \mathbf{n}\, W = 0 \end{aligned}$$

as $\operatorname{curl} \mathbf{H}_I = \mathbf{0}$ in Ω_I. □

Remark 1.2. It is worthwhile to note that, for Case C, assuming $\mathbf{J}_e = \mathbf{0}$, one has

$$\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + \int_{\Omega} i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} = W_{|\Gamma_J} \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n}.$$

In fact, we know that on the ports Γ_E and Γ_J we still have $\operatorname{div}_\tau(\mathbf{E} \times \mathbf{n}) = 0$, and also on Γ_D we have $\operatorname{div}_\tau(\mathbf{E} \times \mathbf{n}) = \operatorname{curl} \mathbf{E} \cdot \mathbf{n} = -i\omega\boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0$; thus we conclude that $\operatorname{div}_\tau(\mathbf{E} \times \mathbf{n}) = 0$ on $\partial\Omega$. Therefore, as before, we can write $\mathbf{E} \times \mathbf{n} = \operatorname{grad} W \times \mathbf{n}$ on $\partial\Omega$, and we see that W is constant on Γ_E and on Γ_J, say, $W = 0$ on Γ_E and $W = W_J$ on Γ_J. Thus

$$\int_{\partial\Omega} \mathbf{n} \times \mathbf{E} \cdot \overline{\mathbf{H}} = -\int_{\partial\Omega} \operatorname{curl} \overline{\mathbf{H}} \cdot \mathbf{n}\, W = -W_J \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n},$$

as $\operatorname{curl} \mathbf{H}_I \cdot \mathbf{n} = 0$ on Γ_D and $W = 0$ on Γ_E.

In particular, we see that there is still a free degree of freedom: it can be either the constant value of W on Γ_J (the voltage between the two ports of Ω_C) or the value $\int_{\Gamma_J} \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}$ (the current intensity in Ω_C). Therefore, in the present case uniqueness requires that one of these conditions also be imposed.

Case C has been proposed in [16] as a valid approximation of a realistic electric port problem. Thus it is a useful starting point for developing our considerations.

In [6] (see also [10]) the following has been proved.

THEOREM 1.3. *For each $\mathbf{J}_e \in (L^2(\Omega))^3$, there exists a unique solution $\mathbf{H}$ and $\mathbf{E}$ of the eddy-current problem (Case* C*)*

$$(1)\qquad \begin{array}{ll} \operatorname{curl} \mathbf{H} - \boldsymbol{\sigma}\mathbf{E} = \mathbf{J}_e & \text{in } \Omega, \\ \operatorname{curl} \mathbf{E} + i\omega\boldsymbol{\mu}\mathbf{H} = \mathbf{0} & \text{in } \Omega, \\ \operatorname{div}(\boldsymbol{\epsilon}_I \mathbf{E}_I) = 0 & \text{in } \Omega_I, \\ \mathbf{E} \times \mathbf{n} = \mathbf{0} & \text{on } \Gamma_E \cup \Gamma_J, \\ \boldsymbol{\epsilon}\mathbf{E} \cdot \mathbf{n} = 0 & \text{on } \Gamma_D, \\ \boldsymbol{\mu}\mathbf{H} \cdot \mathbf{n} = 0 & \text{on } \Gamma_D, \end{array}$$

with one of the following additional conditions:

$$(2)\qquad \textit{either}\quad W_J = V \quad \textit{or} \quad \int_{\Gamma_J} \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n} = I,$$

where the voltage V and the current intensity I are given complex numbers, and W_J denotes the constant value on Γ_J of the function W such that $\mathbf{E} \times \mathbf{n} = \operatorname{grad} W \times \mathbf{n}$ on $\partial\Omega$, having set $W = 0$ on Γ_E.

In [10] and [6] the convergence of a finite element approximation scheme is also proved (in the former the considered unknowns are $\mathbf{H}_C$ and a scalar magnetic potential; in the latter they are the same magnetic potential and $\mathbf{E}_C$).

Other finite element schemes can be found in [26] and [12], where the problem is described through the so-called **T**-Φ formulation, namely, in terms of a current vector potential and a scalar magnetic potential.

2. The power law. In [23], a paper that has deeply inspired our work, Hiptmair and Sterz propose using a suitable power law to relate the voltage and the current intensity. They define

$$(3)\qquad \widehat{P} := \int_{\Omega_C} \boldsymbol{\sigma}\mathbf{E}_C \cdot \overline{\mathbf{E}}_C + i\omega \int_{\Omega} \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}}$$

and assume that, for the problem at hand, the power law $\widehat{P} = V\,\overline{I}$ holds.

In this paper we propose to modify the definition of the power in the following way:

$$P := \int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu} \mathbf{H} \cdot \overline{\mathbf{H}}. \tag{4}$$

Since $\operatorname{curl} \mathbf{H} = \boldsymbol{\sigma}\mathbf{E} + \mathbf{J}_e$, when $\mathbf{J}_{e,C} = \mathbf{0}$ the two definitions are clearly the same: but we will see in the following that the presence of the current density has important consequences. In particular, the relation among power, voltage, and current intensity takes a more general form.

First we note that, when $\mathbf{J}_{e,C} \neq \mathbf{0}$ and $\sigma^{-1}\mathbf{J}_{e,C}$ is a gradient of a suitable scalar function, the solution of the eddy-current problem can take the form $\mathbf{H} = \mathbf{0}$ and $\mathbf{E}_C = -\sigma^{-1}\mathbf{J}_{e,C} \neq \mathbf{0}$; therefore, in that case one has $I = 0$ and $\widehat{P} \neq 0$, and the power law $\widehat{P} = V\,\overline{I}$ does not hold.

To motivate in a more precise way our definition of the power in (4), let us look at this example in further detail. Consider the eddy-current problem (1) (Case C) with a given assigned voltage V and with $\mathbf{J}_{e,I} = \mathbf{0}$, $\mathbf{J}_{e,C} = -V\boldsymbol{\sigma} \operatorname{grad} \phi_C$, where ϕ_C is the unique solution to

$$\begin{cases} \operatorname{div}(\boldsymbol{\sigma} \operatorname{grad} \phi_C) = 0 & \text{in } \Omega_C, \\ \phi_C = 1 & \text{on } \Gamma_J, \\ \phi_C = 0 & \text{on } \Gamma_E, \\ \boldsymbol{\sigma} \operatorname{grad} \phi_C \cdot \mathbf{n} = 0 & \text{on } \Gamma. \end{cases} \tag{5}$$

It is easily seen that the solution is given by $\mathbf{E} = V \operatorname{grad} \phi$ and $\mathbf{H} = \mathbf{0}$, where ϕ is equal to ϕ_C in Ω_C and to ϕ_I in Ω_I, ϕ_I being the unique solution to

$$\begin{cases} \operatorname{div}(\boldsymbol{\epsilon}_I \operatorname{grad} \phi_I) = 0 & \text{in } \Omega_I, \\ \phi_I = \phi_C & \text{on } \Gamma, \\ \boldsymbol{\epsilon}_I \operatorname{grad} \phi_I \cdot \mathbf{n} = 0 & \text{on } \Gamma_D. \end{cases} \tag{6}$$

Therefore, as we noted before, the current intensity I is equal to 0 and $\widehat{P} \neq 0 = V\,\overline{I}$; moreover, this example is also giving us some other useful information. In fact, for each complex number $q \in \mathbb{C}$, take now $\mathbf{J}_{e,C} = q\boldsymbol{\sigma} \operatorname{grad} \phi_C$, $\mathbf{J}_{e,I} = \mathbf{0}$. Computing the power P for the corresponding solution we find, by proceeding as in Theorem 1.1 and Remark 1.2,

$$\begin{aligned} P &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} \\ &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \mathbf{J}_{e,C} \cdot \operatorname{curl} \overline{\mathbf{H}}_C + V \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n} \\ &= q \textstyle\int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + V \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n}. \end{aligned}$$

On the other hand,

$$\begin{aligned} &\textstyle\int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C \\ &\qquad = -\textstyle\int_{\Omega_C} \phi_C \operatorname{div} \operatorname{curl} \overline{\mathbf{H}}_C + \int_{\Gamma_E \cup \Gamma_J \cup \Gamma} \phi_C \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n}_C \\ &\qquad = \textstyle\int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n}. \end{aligned} \tag{7}$$

as $\phi_C = 0$ on Γ_E, $\phi_C = 1$ on Γ_J, and $\operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}_C = \operatorname{curl} \mathbf{H}_I \cdot \mathbf{n}_C = 0$ on Γ.

In conclusion, P is still proportional to $\overline{I}$, as

$$P = (q + V) \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n} = (q + V)\,\overline{I};$$

moreover, this is telling us that, when considering Case C, assigning a voltage V is in some sense equivalent to imposing a current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$ in Ω_C.

More precisely, the solution $(\widetilde{\mathbf{H}}, \widetilde{\mathbf{E}})$ with voltage V and $\mathbf{J}_{e,C} = \mathbf{0}$ and the solution $(\widehat{\mathbf{H}}, \widehat{\mathbf{E}})$ with voltage 0 and $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$ satisfy $\widetilde{\mathbf{H}} = \widehat{\mathbf{H}}$; in fact, the difference $(\widetilde{\mathbf{H}} - \widehat{\mathbf{H}}, \widetilde{\mathbf{E}} - \widehat{\mathbf{E}})$ is a solution of the problem with voltage V and $\mathbf{J}_{e,C} = -V\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$. Therefore, as we have seen above, $\widetilde{\mathbf{E}} - \widehat{\mathbf{E}} = V\,\mathrm{grad}\,\phi$ and $\widetilde{\mathbf{H}} - \widehat{\mathbf{H}} = \mathbf{0}$.

This will lead us to propose a suitable formulation for the eddy-current problem with one of the boundary conditions described in Cases A, B, D, E, and F and, moreover, subjected to a given voltage or current intensity excitation; the key point will be that these excitations, unlike Case C, have to be interpreted as a particular applied current density.

Remark 2.1. Proceeding as in the proof of Theorem 1.1, when $\mathbf{J}_e \neq \mathbf{0}$, we have

$$\begin{aligned} P &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C \cdot \mathrm{curl}\,\overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} \\ &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1}\mathbf{J}_{e,C} \cdot \mathrm{curl}\,\overline{\mathbf{H}}_C - \int_{\Omega_I} \overline{\mathbf{J}}_{e,I} \cdot \mathbf{E}_I + \int_{\partial\Omega} \mathbf{E} \times \mathbf{n} \cdot \overline{\mathbf{H}}. \end{aligned}$$

The term $\int_{\partial\Omega} \mathbf{E} \times \mathbf{n} \cdot \overline{\mathbf{H}}$ is vanishing for Cases A, B, D, and E. Instead, for Case F we have $\int_{\partial\Omega} \mathbf{E} \times \mathbf{n} \cdot \overline{\mathbf{H}} = \int_{\partial\Omega} \overline{\mathbf{J}}_{e,I} \cdot \mathbf{n}\, W$, where W is the scalar function such that $\mathrm{grad}\, W \times \mathbf{n} = \mathbf{E} \times \mathbf{n}$ on $\partial\Omega$. Finally, for Case C we have (see Remark 1.2)

$$\int_{\partial\Omega} \mathbf{E} \times \mathbf{n} \cdot \overline{\mathbf{H}} = \int_{\Gamma_D} \overline{\mathbf{J}}_{e,I} \cdot \mathbf{n}\, W + W_J \int_{\Gamma_J} \mathrm{curl}\,\overline{\mathbf{H}}_C \cdot \mathbf{n},$$

where W_J is a constant and $W_{|\Gamma_J} = W_J$, $W_{|\Gamma_E} = 0$.

When $\mathbf{J}_{e,I} = \mathbf{0}$ and $\mathbf{J}_{e,C} = q\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$, we have seen that for Case C the power law holds in the generalized form $P = (q + V)\,\overline{I}$. This is showing us that we have to consider two voltages, say, an "electric" voltage V (the value $V = W_{|\Gamma_J} - W_{|\Gamma_E}$) and a "source" voltage q, associated to the current density $q\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$. Their sum $q + V$ is the total voltage.

When considering the other cases A, B, D, E, and F, only the "source" voltage has meaning.

3. Voltage and current excitation. Since the eddy-current problem has a unique solution for each of the sets of boundary conditions described in Cases A, B, D, E, and F (see Theorem 1.1), it is not possible to impose an additional condition, say, voltage or current intensity, if we do not relax some of the other equations.

Before starting, let us mention the formulations proposed in some preceding papers. In [23], where the voltage/current excitation problem has been considered in the most systematic way, it was proposed to slightly modify the formulation for Case A, requiring $\mathbf{E} \times \mathbf{n} = \mathrm{grad}\,\varphi \times \mathbf{n}$ on $\partial\Omega$, where $\varphi \in H^{1/2}(\partial\Omega)$, $\varphi = V$ on Γ_J, $\varphi = 0$ on Γ_E (and, therefore, $\varphi \neq \mathrm{const}$ in a transition region $\Theta \subset \Gamma_D$). In other words, $\mathbf{E} \times \mathbf{n} \neq \mathbf{0}$ in Θ. This formulation, which is proved to be well-posed, depends, however, on the choice of the region Θ and of the function φ in Θ. An alternative approach, also proposed in [23], valid for all the cases considered here and for which $\Theta = \emptyset$, ends up with the violation of the Faraday law on a specific surface (either the surface that "cuts" the basic nonbounding cycle in Ω_I, or else any surface crossing the interface Γ).

In [20] and [28] the internal conductor case is considered, having assigned a given voltage V. Also in this case the Faraday law is violated on the cutting surface Λ. Instead, the approach proposed in [21] gives a solution that does not satisfy the Faraday law across the interface Γ.

In [10] a formulation for the electric port case with assigned current intensity is given, leading to the solution also obtained in [6] for Case C; however, for Case A it can be checked that the Faraday law is violated on the cutting surface Ξ (instead, the violation of the Faraday law across the interface Γ occurs in [8], where the internal conductor case is considered).

Finally, in [26] and [12] the finite element approximation of Case C is considered for an assigned voltage, by means of a formulation based on a current vector potential and a magnetic scalar potential.

Let us come now to our point of view: clearly, on one side we do not want to give up Maxwell equations, namely, Faraday and Ampère equations; on the other side, we would like to formulate a problem for which only the physical quantities and the physical domains Ω_C and Ω_I play a role (and not artificial regions like, e.g., the transition zone Θ introduced in [23]).

The main point is to recall what we have proved for Case C, where a voltage V was "equivalent" (at least, for the determination of $\mathbf{H}$ and in the power law) to the current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$ in Ω_C. Note that the function $\mathrm{grad}\,\phi_C$ is the basis function of the space of harmonic fields

$$\begin{aligned}\widehat{\mathcal{H}}(\Omega_C) \ := \{\widehat{\boldsymbol{\eta}} \in (L^2(\Omega_C))^3 \mid \mathrm{curl}\,\widehat{\boldsymbol{\eta}} = \mathbf{0}, \mathrm{div}(\boldsymbol{\sigma}\widehat{\boldsymbol{\eta}}) = 0,\\ \boldsymbol{\sigma}\widehat{\boldsymbol{\eta}}\cdot\mathbf{n} = 0 \text{ on } \Gamma, \widehat{\boldsymbol{\eta}}\times\mathbf{n} = \mathbf{0} \text{ on } \Gamma_E\cup\Gamma_J\},\end{aligned}$$

normalized with the condition

$$\int_{\widehat{\gamma}} \widehat{\boldsymbol{\eta}}\cdot d\boldsymbol{\tau} = 1,$$

where $\widehat{\gamma}$ is a path joining Γ_E to Γ_J. Thus, for the internal conductor case we are led to introduce the space of harmonic fields

$$\mathcal{H}(\Omega_C) := \{\boldsymbol{\eta} \in (L^2(\Omega_C))^3 \mid \mathrm{curl}\,\boldsymbol{\eta} = \mathbf{0}, \mathrm{div}(\boldsymbol{\sigma}\boldsymbol{\eta}) = 0, \boldsymbol{\sigma}\boldsymbol{\eta}\cdot\mathbf{n} = 0 \text{ on } \Gamma\},$$

defining as $\boldsymbol{\rho}_C$ its basis function normalized with the condition

$$\int_{\gamma} \boldsymbol{\rho}_C\cdot d\boldsymbol{\tau} = 1,$$

where the (closed) cycle γ is internal to Ω_C (and we have freely chosen an orientation of γ).

The voltage or current excitation problem is therefore formulated as follows.

Voltage rule. *When the voltage V is imposed, modify Ohm's law in Ω_C by adding to the current density $\boldsymbol{\sigma}\mathbf{E}_C$ the "applied" current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\mathbf{Q}_C$, where $\mathbf{Q}_C = \mathrm{grad}\,\phi_C$ for the electric port case, and $\mathbf{Q}_C = \boldsymbol{\rho}_C$ for the internal conductor case. Thus the Ampère law reads*

$$\mathrm{curl}\,\mathbf{H}_C - \boldsymbol{\sigma}\mathbf{E}_C = V\boldsymbol{\sigma}\mathbf{Q}_C.$$

In the former case, we intend that the voltage passes from 0 on Γ_E to V on Γ_J; in the latter case, the voltage passes from 0 to V along the basic cycle γ.

Current intensity rule. *When the current intensity I is imposed, modify Ohm's law in Ω_C by adding to the current density $\boldsymbol{\sigma}\mathbf{E}_C$ the "applied" current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\mathbf{Q}_C$, where $\mathbf{Q}_C$ is as in the "voltage rule" and V has to be determined. Thus the Ampère law reads*

$$\mathrm{curl}\,\mathbf{H}_C - \boldsymbol{\sigma}\mathbf{E}_C - V\boldsymbol{\sigma}\mathbf{Q}_C = 0.$$

Then determine the field quantities and the voltage V such that the additional constraint

$$\int_S \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n} = I$$

is also satisfied, where $S = \Gamma_J$ for the electric port case, and $S = \Sigma$, a section of Ω_C, for the internal conductor case. In the former case, the unit vector $\mathbf{n}$ is the outward normal on Γ_J; in the latter case, the unit vector $\mathbf{n}$ on Σ is oriented the same as the basic cycle γ.

Let us show that, when adopting these two rules, we are respecting the power law. Assume that we have $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\mathbf{Q}_C$ and $\mathbf{J}_{e,I} = \mathbf{0}$. Then, by proceeding as in Theorem 1.1, and taking into account the boundary conditions of Cases A, B, D, E, and F, we have that

$$\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} = \int_{\Omega_C} \boldsymbol{\sigma}^{-1}\mathbf{J}_{e,C} \cdot \operatorname{curl} \overline{\mathbf{H}}_C,$$

and hence

$$P = \int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} = V \int_{\Omega_C} \mathbf{Q}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C.$$

On the other hand, from (7) we have

$$\int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C = \int_{\Gamma_J} \operatorname{curl} \overline{\mathbf{H}}_C \cdot \mathbf{n} = \overline{I};$$

thus if $\mathbf{Q}_C = \operatorname{grad} \phi_C$, we conclude with

$$\begin{aligned} P &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} \\ &= V \textstyle\int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C = V\,\overline{I}, \end{aligned}$$

the power law for the electric port case.

The internal conductor case needs some additional information in order to express the current intensity in a suitable way. Let us denote by Σ a section of Ω_C, namely, a surface in Ω_C cutting the basic nonbounding cycle γ. We know that the basis function $\boldsymbol{\rho}_C$ is the $L^2(\Omega_C)$-extension of the gradient of a suitable scalar function q, defined in $\Omega_C \setminus \Sigma$ and having a jump equal to 1 across Σ. Hence,

$$\begin{aligned} \textstyle\int_{\Omega_C} \operatorname{curl} \mathbf{H}_C \cdot \boldsymbol{\rho}_C &= \textstyle\int_{\Omega_C \setminus \Sigma} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{grad} q \\ &= -\textstyle\int_{\Omega_C \setminus \Sigma} q \operatorname{div} \operatorname{curl} \mathbf{H}_C + \int_\Gamma q \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}_C \\ &\qquad + \textstyle\int_\Sigma \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n} \\ &= \textstyle\int_\Sigma \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}, \end{aligned} \tag{8}$$

as $\operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}_C = \operatorname{curl} \mathbf{H}_I \cdot \mathbf{n}_C = 0$ on Γ and the jump of q on Σ is equal to 1. Hence we end up with

$$\begin{aligned} P &= \textstyle\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C + i\omega \int_\Omega \boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{H}} \\ &= V \textstyle\int_{\Omega_C} \boldsymbol{\rho}_C \cdot \operatorname{curl} \overline{\mathbf{H}}_C = V\,\overline{I}, \end{aligned}$$

the power law for the internal conductor case.

Remark 3.1. As is clear from our procedure, in the electric port case we could obtain a suitable formulation (namely, satisfying the power law) for any current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\,\mathrm{grad}\,\Phi_C$ such that $\Phi_C = 1$ on Γ_J and $\Phi_C = 0$ on Γ_E. Hence, how is the choice of ϕ_C introduced in (5) motivated?

In this respect, it should be noted that, from the Ampère equation $\mathrm{curl}\,\mathbf{H}_C = \boldsymbol{\sigma}\mathbf{E}_C + \mathbf{J}_{e,C}$, the electric field satisfies the (physically consistent) conditions $\mathrm{div}(\boldsymbol{\sigma}\mathbf{E}_C) = 0$ in Ω_C and $\boldsymbol{\sigma}\mathbf{E}_C \cdot \mathbf{n} = 0$ on Γ only if $\mathrm{div}\,\mathbf{J}_{e,C} = 0$ in Ω_C and $\mathbf{J}_{e,C} \cdot \mathbf{n} = 0$ on Γ, and therefore, only if $\Phi_C = \phi_C$, the solution to (5).

The same remark applies for the internal conductor case: in that situation, the integral

$$\int_{\Omega_C} \mathbf{u}_C \cdot \mathrm{curl}\,\overline{\mathbf{H}}_C$$

has the same value for any vector field $\mathbf{u}_C$ such that $\mathrm{curl}\,\mathbf{u}_C = \mathbf{0}$ and $\int_\gamma \mathbf{u}_C \cdot d\boldsymbol{\tau} = 1$. But if we also require that $\mathrm{div}(\boldsymbol{\sigma}\mathbf{u}_C) = 0$ in Ω_C and $\boldsymbol{\sigma}\mathbf{u}_C \cdot \mathbf{n} = 0$ on Γ, then we conclude $\mathbf{u}_C = \boldsymbol{\rho}_C$.

4. Variational formulations. We can consider **H**-based formulations, or **E**-based formulations. In our opinion, the simplest approach is in terms of **H**. We will focus first on the electric port case; however, we do not present here Case C, which, for a "hybrid" formulation which is related to the **H**-formulation, has been studied in [6]. Then we will consider the internal conductor case, whose formulation is quite similar, focusing in particular on Case F.

Electric ports: Voltage excitation, ***H****-formulation.* For Case A, the problem is as follows: for each given $V \in \mathbb{C}$ find the unique solution $\mathbf{H} \in X$ to

$$\int_{\Omega_C} \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{w}} = V \int_{\Omega_C} \mathrm{grad}\,\phi_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C \tag{9}$$

for each $\mathbf{w} \in X$, where

$$X := \{\mathbf{w} \in H(\mathrm{curl};\Omega) \mid \mathrm{curl}\,\mathbf{w}_I = \mathbf{0} \text{ in } \Omega_I\}.$$

Then set $\mathbf{E}_C := \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C - V\,\mathrm{grad}\,\phi_C$ in Ω_C, and in Ω_I define $\mathbf{E}_I$ to be the solution to

$$\begin{cases} \mathrm{curl}\,\mathbf{E}_I = -i\omega\boldsymbol{\mu}_I\mathbf{H}_I & \text{in } \Omega_I, \\ \mathrm{div}(\boldsymbol{\epsilon}_I\mathbf{E}_I) = 0 & \text{in } \Omega_I, \\ \mathbf{E}_I \times \mathbf{n}_I = -\mathbf{E}_C \times \mathbf{n}_C & \text{on } \Gamma, \\ \mathbf{E}_I \times \mathbf{n} = \mathbf{0} & \text{on } \Gamma_D. \end{cases} \tag{10}$$

Let us remark that the voltage excitation problem for Case B is trivial: in fact, from (7) and the Stokes theorem we have

$$\int_{\Omega_C} \mathrm{grad}\,\phi_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C = \int_{\Gamma_J} \mathrm{curl}\,\overline{\mathbf{w}}_C \cdot \mathbf{n} = \int_{\partial\Gamma_J} \overline{\mathbf{w}} \cdot d\boldsymbol{\tau} = 0,$$

as $\mathbf{w} \times \mathbf{n} = \mathbf{0}$ on Γ_D. Therefore, for any $V \in \mathbb{C}$ we find $\mathbf{H} = \mathbf{0}$; hence, we can assume that $V = 0$ and set $\mathbf{E} = \mathbf{0}$.

The well-posedness of problem (9) comes from the coerciveness in X of the sesquilinear form $\int_{\Omega_C} \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{w}}$.

Instead, a delicate point here is the unique solvability of problem (10). In fact, as is well known, boundary-value problems for the curl-div system in general need some compatibility conditions to be satisfied in order to ensure the existence of a solution and need suitable additional conditions to be imposed to guarantee its uniqueness; some of these conditions are related to the nontrivial topology of Ω_I.

More precisely, first one has to verify the conditions $\operatorname{div}(\boldsymbol{\mu}_I \mathbf{H}_I) = 0$ in Ω_I, $\operatorname{div}_\tau(\mathbf{E}_C \times \mathbf{n}_C) = i\omega \boldsymbol{\mu}_I \mathbf{H}_I \cdot \mathbf{n}_I$ on Γ, and $\boldsymbol{\mu}_I \mathbf{H}_I \cdot \mathbf{n}_I = 0$ on Γ_D.

It is possible to check that these conditions are satisfied by means of a suitable choice of test functions in (9) (for a similar procedure, see, for instance, [3]). In fact, the first follows from (9) taking as test function $\mathbf{w} = \operatorname{grad}\psi$, ψ a smooth function with a compact support in Ω (and in this way one also obtains $\boldsymbol{\mu}_C \mathbf{H}_C \cdot \mathbf{n}_C + \boldsymbol{\mu}_I \mathbf{H}_I \cdot \mathbf{n}_I = 0$ on Γ, as, indeed, $\operatorname{div}(\boldsymbol{\mu}\mathbf{H}) = 0$ in Ω). The second comes from the Faraday equation in Ω_C, which is obtained by integration by parts, and the relation $\operatorname{div}_\tau(\mathbf{E}_C \times \mathbf{n}_C) = \operatorname{curl}\mathbf{E}_C \cdot \mathbf{n}_C$. The last is obtained by taking as test function $\mathbf{w} = \operatorname{grad}\psi$, with $\psi \in H^1(\Omega)$.

Then one has to consider some spaces of harmonic fields. Concerning uniqueness, it is clear that we are interested in requiring that the solution $\mathbf{E}_I$ is orthogonal (with weight $\boldsymbol{\epsilon}_I$) to the space

$$\begin{aligned} \mathcal{H}_A^{\mathrm{un}}(\Omega_I) \ &:= \{\widehat{\boldsymbol{\eta}} \in (L^2(\Omega_I))^3 \mid \operatorname{curl}\widehat{\boldsymbol{\eta}} = \mathbf{0}, \operatorname{div}(\boldsymbol{\epsilon}_I \widehat{\boldsymbol{\eta}}) = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad \widehat{\boldsymbol{\eta}} \times \mathbf{n} = \mathbf{0} \text{ on } \Gamma_D \cup \Gamma\}. \end{aligned}$$

However, this space is trivial (namely, it contains only $\widehat{\boldsymbol{\eta}} = \mathbf{0}$). In fact, cutting Ω_I with a surface Ξ transversal to Γ_D and Γ, an element $\widehat{\boldsymbol{\eta}}$ of $\mathcal{H}_A^{\mathrm{un}}(\Omega_I)$ in the set $\Omega_I \setminus \Xi$ is the gradient of a function p having a constant jump through Ξ. But, due to the fact that $\operatorname{grad} p \times \mathbf{n} = \mathbf{0}$ on $\Gamma_D \cup \Gamma$, a connected surface, p is constant on $\Gamma_D \cup \Gamma$, and therefore its jump through Ξ is equal to 0. Thus $\widehat{\boldsymbol{\eta}}$ is the gradient of a harmonic function p with constant boundary value: hence p is constant in Ω_I and $\widehat{\boldsymbol{\eta}} = \mathbf{0}$ in Ω_I.

Existence is instead associated to the space

$$\begin{aligned} \mathcal{H}_A^{\mathrm{ex}}(\Omega_I) \ &:= \{\widehat{\boldsymbol{\eta}} \in (L^2(\Omega_I))^3 \mid \operatorname{curl}\widehat{\boldsymbol{\eta}} = \mathbf{0}, \operatorname{div}\widehat{\boldsymbol{\eta}} = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad \widehat{\boldsymbol{\eta}} \cdot \mathbf{n} = 0 \text{ on } \Gamma_D \cup \Gamma\}, \end{aligned}$$

which, proceeding as before, is easily shown to be one-dimensional; let us denote by $\widehat{\boldsymbol{\rho}}_I$ its basis vector. For the solvability of problem (10) one has to satisfy the compatibility condition

$$\int_{\Omega_I} i\omega \boldsymbol{\mu}_I \mathbf{H}_I \cdot \widehat{\boldsymbol{\rho}}_I + \int_\Gamma \mathbf{E}_C \times \mathbf{n}_C \cdot \widehat{\boldsymbol{\rho}}_I = 0, \tag{11}$$

which comes from $(10)_1$ and $(10)_3$ by integration by parts. This relation indeed follows from (9) by choosing the test function $\mathbf{w}_I = \widehat{\boldsymbol{\rho}}_I$ and $\mathbf{w}_C = \widehat{\boldsymbol{\rho}}^*$, where $\widehat{\boldsymbol{\rho}}^* \in H(\mathrm{curl}; \Omega_C)$ satisfies $\widehat{\boldsymbol{\rho}}^* \times \mathbf{n}_I = \widehat{\boldsymbol{\rho}}_I \times \mathbf{n}_I$ on Γ, integrating by parts and using the Faraday equation in Ω_C.

In conclusion, (10) is uniquely solvable. It is important to remark that this is not the case if one defines, as in [23], where the same formulation (9) has been proposed, the electric field $\mathbf{E}_C = \boldsymbol{\sigma}^{-1} \operatorname{curl}\mathbf{H}_C$ in Ω_C: in that case, in fact, (11) is not satisfied, and therefore it is not possible to determine $\mathbf{E}_I$.

It is worthwhile to note that (11) is indeed equivalent to the Faraday equation on the surface Ξ that cuts the basic nonbounding cycle in Ω_I. Hence, setting $\mathbf{E}_C = \boldsymbol{\sigma}^{-1} \operatorname{curl}\mathbf{H}_C$ leads to the violation of the Faraday equation on that surface.

Electric ports: Current intensity excitation, **H***-formulation.* Let us start noting that this problem does not have a meaning for Case B. In fact, one has

$$I = \int_{\Gamma_J} \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n} = \int_{\partial\Gamma_J} \mathbf{H} \cdot d\boldsymbol{\tau} = 0,$$

as $\mathbf{H} \times \mathbf{n} = \mathbf{0}$ on Γ_D.

Therefore, we consider only Case A. The problem can be expressed in this way: for each given $I \in \mathbb{C}$ find the unique solution $(\mathbf{H}, V) \in X \times \mathbb{C}$ to

$$(12) \qquad \begin{cases} \int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{w}}_C & + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{w}} \\ & - V \int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \overline{\mathbf{w}}_C = 0, \\ \int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \mathbf{H}_C = I & \end{cases}$$

for each $\mathbf{w} \in X$, where X is as in (9). Then $\mathbf{E}_C$ and $\mathbf{E}_I$ are determined in the same way as before. Let us also recall that, from (7), we have

$$\int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \mathbf{H}_C = \int_{\Gamma_J} \operatorname{curl} \mathbf{H}_C \cdot \mathbf{n}.$$

The well-posedness of problem (12) comes from the theory of saddle-point problems. In fact, the sesquilinear form $\int_{\Omega_C} \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C \cdot \operatorname{curl} \overline{\mathbf{w}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{w}}$ is coercive in X; moreover, since the unknown $V \in \mathbb{C}$ is a number, to show that the inf-sup condition is satisfied it is enough to find $\mathbf{w}^* \in X$ such that

$$\left| \int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \mathbf{w}^*_C \right| > 0.$$

This can be done by taking the solution $\mathbf{w}^*_C$ to

$$\begin{cases} \operatorname{curl} \mathbf{w}^*_C = \boldsymbol{\sigma} \operatorname{grad} \phi_C & \text{in } \Omega_C, \\ \operatorname{div} \mathbf{w}^*_C = 0 & \text{in } \Omega_C, \\ \mathbf{w}^*_C \cdot \mathbf{n} = 0 & \text{on } \Gamma_E \cup \Gamma_J \cup \Gamma \end{cases}$$

and the solution $\mathbf{w}^*_I$ to

$$\begin{cases} \operatorname{curl} \mathbf{w}^*_I = \mathbf{0} & \text{in } \Omega_I, \\ \operatorname{div} \mathbf{w}^*_I = 0 & \text{in } \Omega_I, \\ \mathbf{w}^*_I \times \mathbf{n}_I = -\mathbf{w}^*_C \times \mathbf{n}_C & \text{on } \Gamma, \\ \mathbf{w}^*_I \cdot \mathbf{n} = 0 & \text{on } \Gamma_D. \end{cases}$$

Formulation (12) has been proposed also in [10] (for both Cases A and C). However, there $\mathbf{E}_C = \boldsymbol{\sigma}^{-1} \operatorname{curl} \mathbf{H}_C$, thus violating, for Case A, the Faraday equation on the surface Ξ.

Electric ports: Voltage excitation, **E***-formulation.* Having clarified that voltage excitation is equivalent to a source $V\boldsymbol{\sigma} \operatorname{grad} \phi_C$, the electric field formulation for Case A is easily devised: for each given $V \in \mathbb{C}$ find $\mathbf{E} \in Y$ such that

$$(13) \qquad \int_\Omega \boldsymbol{\mu}^{-1} \operatorname{curl} \mathbf{E} \cdot \operatorname{curl} \overline{\mathbf{z}} + \int_{\Omega_C} i\omega\boldsymbol{\sigma}\mathbf{E}_C \cdot \overline{\mathbf{z}}_C = -i\omega V \int_{\Omega_C} \boldsymbol{\sigma} \operatorname{grad} \phi_C \cdot \overline{\mathbf{z}}_C$$

for each $\mathbf{z} \in Y$, where

$$Y := \{\mathbf{z} \in H(\operatorname{curl}; \Omega) \mid \operatorname{div}(\boldsymbol{\epsilon}_I \mathbf{z}_I) = 0 \text{ in } \Omega_I, \mathbf{z} \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}.$$

The existence of a solution $\mathbf{E}$ to problem (13) follows at once from what was already proved for the $\mathbf{H}$-formulation. Uniqueness is straightforward.

The magnetic field $\mathbf{H}$ is then determined in Ω as $\mathbf{H} = -\frac{1}{i\omega}\boldsymbol{\mu}^{-1}\operatorname{curl}\mathbf{E}$.

A formulation similar to (13) (but based on the source term $V\boldsymbol{\sigma}\operatorname{grad}\Phi_C$, the function Φ_C having been defined in Remark 3.1) has been presented in [23]. However, there the electric field is not the solution to (13), but it is corrected, only in Ω_C, by adding $V\operatorname{grad}\Phi_C$. Since a curl-free vector field in Ω_I that has tangential component on Γ equal to $V\operatorname{grad}\Phi_C \times \mathbf{n}$ does not exist (again, this is related to the solvability of problem (10)), this leads to the violation of the continuity of the tangential component of $\mathbf{E}$ through the interface Γ and thus to the violation of the Faraday law.

Electric ports: Current intensity excitation, $\mathbf{E}$*-formulation.* Since $\operatorname{curl}\mathbf{H}_C = \boldsymbol{\sigma}\mathbf{E}_C + V\boldsymbol{\sigma}\operatorname{grad}\phi_C$, the variational formulation (for the sole Case A) now reads as follows: for each given $I \in \mathbb{C}$, find $(\mathbf{E}, V) \in Y \times \mathbb{C}$ such that

$$(14)\qquad \begin{cases} \int_\Omega \boldsymbol{\mu}^{-1}\operatorname{curl}\mathbf{E}\cdot\operatorname{curl}\overline{\mathbf{z}} + \int_{\Omega_C} i\omega\boldsymbol{\sigma}\mathbf{E}_C\cdot\overline{\mathbf{z}}_C \\ \qquad\qquad + i\omega V\int_{\Omega_C}\boldsymbol{\sigma}\operatorname{grad}\phi_C\cdot\overline{\mathbf{z}}_C = 0, \\ \int_{\Omega_C}\operatorname{grad}\phi_C\cdot\boldsymbol{\sigma}\mathbf{E}_C + V\int_{\Omega_C}\boldsymbol{\sigma}\operatorname{grad}\phi_C\cdot\operatorname{grad}\phi_C = I \end{cases}$$

for each $\mathbf{z} \in Y$.

As before, existence of a solution is ensured by the correspondent result for the magnetic field $\mathbf{H}$. Instead, uniqueness is a more delicate point. In fact, multiplying $(14)_2$ by $i\omega\overline{U}$, where $U \in \mathbb{C}$, we find

$$\begin{aligned} &\int_\Omega \boldsymbol{\mu}^{-1}\operatorname{curl}\mathbf{E}\cdot\operatorname{curl}\overline{\mathbf{z}} \\ &\qquad + i\omega\int_{\Omega_C}\boldsymbol{\sigma}(\mathbf{E}_C + V\operatorname{grad}\phi_C)\cdot(\overline{\mathbf{z}}_C + \overline{U}\operatorname{grad}\phi_C) = i\omega I\,\overline{U}. \end{aligned}$$

Thus, putting $I = 0$ and choosing $\mathbf{z} = \mathbf{E}$ and $U = V$, we obtain $\operatorname{curl}\mathbf{E} = \mathbf{0}$ in Ω and $\mathbf{E}_C + V\operatorname{grad}\phi_C = \mathbf{0}$ in Ω_C. Since Ω is simply connected, we also have $\mathbf{E} = \operatorname{grad}\psi$ in Ω, and the boundary condition $\mathbf{E}\times\mathbf{n} = \mathbf{0}$ on $\partial\Omega$ gives $\psi = \text{const}$ on $\partial\Omega$. Therefore, integrating $\mathbf{E}_C$ on the path $\widehat{\gamma}$ joining Γ_E to Γ_J, we find

$$\begin{aligned} 0 &= \int_{\widehat{\gamma}}\operatorname{grad}\psi_C\cdot d\boldsymbol{\tau} = \int_{\widehat{\gamma}}\mathbf{E}_C\cdot d\boldsymbol{\tau}, \\ &= -\int_{\widehat{\gamma}} V\operatorname{grad}\phi_C\cdot d\boldsymbol{\tau} = -V. \end{aligned}$$

Thus $V = 0$, and consequently $\mathbf{E} = \mathbf{0}$.

Also in this case, the magnetic field $\mathbf{H}$ is obtained in Ω by setting

$$\mathbf{H} = -\frac{1}{i\omega}\boldsymbol{\mu}^{-1}\operatorname{curl}\mathbf{E}.$$

Again, a formulation like (14) (but based on the source term $V\boldsymbol{\sigma}\operatorname{grad}\Phi_C$) has been proposed in [23]. The remark at the end of the preceding subsection still applies.

Internal conductor: Voltage excitation, $\mathbf{H}$*-formulation.* We have already made explicit the "voltage rule": applying a voltage is equivalent to considering a current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\boldsymbol{\rho}_C$. Then, for the $\mathbf{H}$-based formulation, Cases D and E can be studied as in [3]. Let us focus on Case F.

The problem reads as follows: for each given $V \in \mathbb{C}$ find the unique solution $\mathbf{H} \in X$ to

$$(15)\qquad \int_{\Omega_C}\boldsymbol{\sigma}^{-1}\operatorname{curl}\mathbf{H}_C\cdot\operatorname{curl}\overline{\mathbf{w}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H}\cdot\overline{\mathbf{w}} = V\int_{\Omega_C}\boldsymbol{\rho}_C\cdot\operatorname{curl}\overline{\mathbf{w}}_C$$

for each $\mathbf{w} \in X$, where

$$X := \{\mathbf{w} \in H(\mathrm{curl};\Omega) \,|\, \mathrm{curl}\,\mathbf{w}_I = \mathbf{0} \text{ in } \Omega_I\}.$$

Then set $\mathbf{E}_C := \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C - V\boldsymbol{\rho}_C$ in Ω_C, and in Ω_I define $\mathbf{E}_I$ to be the solution to

$$(16) \qquad \begin{cases} \mathrm{curl}\,\mathbf{E}_I = -i\omega\boldsymbol{\mu}_I\mathbf{H}_I & \text{in } \Omega_I, \\ \mathrm{div}(\boldsymbol{\epsilon}_I\mathbf{E}_I) = 0 & \text{in } \Omega_I, \\ \mathbf{E}_I \times \mathbf{n}_I = -\mathbf{E}_C \times \mathbf{n}_C & \text{on } \Gamma, \\ \boldsymbol{\epsilon}_I\mathbf{E}_I \cdot \mathbf{n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Again, the main problem here is the solvability of (16). For the internal conductor, this has been already done in [3], to which we refer the reader.

Let us also recall that the same variational formulation (15) has been proposed in [20], [28], and [23]. However, there $\mathbf{E}_C := \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C$, leading to the violation of the Faraday equation on the surface Λ cutting the basic nonbounding cycle of Ω_I.

Internal conductor: Current intensity excitation, **H***-formulation.* Let us start focusing on Case F. Recalling (8), the problem is as follows: for each given $I \in \mathbb{C}$ find the unique solution $(\mathbf{H}, V) \in X \times \mathbb{C}$ to

$$(17) \qquad \begin{cases} \int_{\Omega_C} \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C + \int_\Omega i\omega\boldsymbol{\mu}\mathbf{H} \cdot \overline{\mathbf{w}} \\ \qquad\qquad - V\int_{\Omega_C} \boldsymbol{\rho}_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C = 0, \\ \int_{\Omega_C} \boldsymbol{\rho}_C \cdot \mathrm{curl}\,\mathbf{H}_C = I \end{cases}$$

for each $\mathbf{w} \in X$, where X is as in (15); then set $\mathbf{E}_C := \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C - V\boldsymbol{\rho}_C$ in Ω_C and determine $\mathbf{E}_I$ as in (16).

The well-posedness of problem (17) comes from the theory of saddle-point problems. Taking into account what we have already presented for the electric port case, it is enough to find $\mathbf{w}^* \in X$ such that

$$\left| \int_{\Omega_C} \boldsymbol{\rho}_C \cdot \mathrm{curl}\,\mathbf{w}_C^* \right| > 0.$$

This can be done by taking the solution $\mathbf{w}_C^*$ to

$$\begin{cases} \mathrm{curl}\,\mathbf{w}_C^* = \boldsymbol{\sigma}\boldsymbol{\rho}_C & \text{in } \Omega_C, \\ \mathrm{div}\,\mathbf{w}_C^* = 0 & \text{in } \Omega_C, \\ \mathbf{w}_C^* \times \mathbf{n}_C = c_0\boldsymbol{\rho}_I \times \mathbf{n}_C & \text{on } \Gamma \end{cases}$$

and the solution $\mathbf{w}_I^*$ to

$$\begin{cases} \mathrm{curl}\,\mathbf{w}_I^* = \mathbf{0} & \text{in } \Omega_I, \\ \mathrm{div}\,\mathbf{w}_I^* = 0 & \text{in } \Omega_I, \\ \mathbf{w}_I^* \times \mathbf{n}_I = c_0\boldsymbol{\rho}_I \times \mathbf{n}_I & \text{on } \Gamma, \\ \mathbf{w}_I^* \times \mathbf{n} = \mathbf{0} & \text{on } \partial\Omega, \\ \int_{\partial\Omega} \mathbf{w}_I^* \cdot \mathbf{n} = 0, \end{cases}$$

where $c_0 = \int_{\Omega_C} \boldsymbol{\sigma}\boldsymbol{\rho}_C \cdot \boldsymbol{\rho}_C$. Here $\boldsymbol{\rho}_I$ is the basis function of the space

$$\mathcal{H}(\Omega_I) := \{\boldsymbol{\eta} \in (L^2(\Omega_I))^3 \,|\, \mathrm{curl}\,\boldsymbol{\eta} = \mathbf{0}, \mathrm{div}\,\boldsymbol{\eta} = 0, \boldsymbol{\eta} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \cup \Gamma\},$$

normalized by $\int_{\widetilde{\gamma}} \boldsymbol{\rho}_I \cdot d\boldsymbol{\tau} = 1$, where $\widetilde{\gamma}$ is the basic nonbounding cycle entering the "handle" of Ω_C, and oriented consistently with the nonbounding cycle γ which runs in Ω_C (namely, each one is oriented counterclockwise with respect to the other). Note that the existence of the solution $\mathbf{w}_C^*$ is a consequence of the relation $\int_\Gamma (\boldsymbol{\rho}_C \times \mathbf{n}_C) \cdot \boldsymbol{\rho}_I = 1$.

To complete the presentation, let us note that, if interested in considering Case D, one has to substitute in (16) the boundary condition $\boldsymbol{\epsilon}_I \mathbf{E}_I \cdot \mathbf{n} = 0$ on $\partial\Omega$ with $\mathbf{E}_I \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$ and add the condition $\int_{\partial\Omega} \boldsymbol{\epsilon}_I \mathbf{E}_I \cdot \mathbf{n} = 0$.

Instead, concerning Case E, one has to use in (17) the space

$$X := \{\mathbf{w} \in H(\mathrm{curl};\Omega) \,|\, \mathrm{curl}\, \mathbf{w}_I = \mathbf{0} \text{ in } \Omega_I, \mathbf{w}_I \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}.$$

Internal conductor: Voltage excitation, **E**-*formulation.* We are not going to give details for this case. In fact, the "voltage rule" is telling us that we have only to consider a current density $\mathbf{J}_{e,C} = V\boldsymbol{\sigma}\boldsymbol{\rho}_C$; hence this formulation is easily devised (for instance, for Cases D and E one can follow what was done in [5]). Moreover, the case in which excitation is due to the current intensity can also illustrate the functional framework to be used for the voltage excitation case (in this respect, see the first equation in (18)).

Internal conductor: Current intensity excitation, **E**-*formulation.* The "current intensity rule" says that the given current intensity I is generating not only the electric field but also a current density $V\boldsymbol{\sigma}\boldsymbol{\rho}_C$. Moreover, we have $\mathrm{curl}\, \mathbf{H}_C = \boldsymbol{\sigma}\mathbf{E}_C + V\boldsymbol{\sigma}\boldsymbol{\rho}_C$. Then, the problem is as follows: for each given $I \in \mathbb{C}$ find $(\mathbf{E}, V) \in Y \times \mathbb{C}$ such that

$$(18)\qquad \begin{cases} \int_\Omega \boldsymbol{\mu}^{-1} \mathrm{curl}\, \mathbf{E} \cdot \mathrm{curl}\, \overline{\mathbf{z}} + \int_{\Omega_C} i\omega\boldsymbol{\sigma}\mathbf{E}_C \cdot \overline{\mathbf{z}}_C \\ \qquad\qquad\qquad + i\omega V \int_{\Omega_C} \boldsymbol{\sigma}\boldsymbol{\rho}_C \cdot \overline{\mathbf{z}}_C = 0, \\ \int_{\Omega_C} \boldsymbol{\rho}_C \cdot \boldsymbol{\sigma}\mathbf{E}_C + V \int_{\Omega_C} \boldsymbol{\sigma}\boldsymbol{\rho}_C \cdot \boldsymbol{\rho}_C = I \end{cases}$$

for each $\mathbf{z} \in Y$, where

$$Y := \begin{cases} \{\mathbf{z} \in H(\mathrm{curl};\Omega) \,|\, \mathrm{div}(\boldsymbol{\epsilon}_I \mathbf{z}_I) = 0 \text{ in } \Omega_I, \\ \qquad \mathbf{z} \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega, \int_{\partial\Omega} \boldsymbol{\epsilon}_I \mathbf{E}_I \cdot \mathbf{n} = 0\} & \text{for Case D,} \\ \{\mathbf{z} \in H(\mathrm{curl};\Omega) \,|\, \mathrm{div}(\boldsymbol{\epsilon}_I \mathbf{z}_I) = 0 \text{ in } \Omega_I, \\ \qquad \boldsymbol{\epsilon}_I \mathbf{z}_I \cdot \mathbf{n} = 0 \text{ on } \partial\Omega\} & \text{for Case E,} \\ \{\mathbf{z} \in H(\mathrm{curl};\Omega) \,|\, \mathrm{div}(\boldsymbol{\epsilon}_I \mathbf{z}_I) = 0 \text{ in } \Omega_I, \\ \qquad \boldsymbol{\epsilon}_I \mathbf{z}_I \cdot \mathbf{n} = 0 \text{ on } \partial\Omega, \mathrm{div}_\tau(\mathbf{z} \times \mathbf{n}) = \mathbf{0} \text{ on } \partial\Omega\} & \text{for Case F.} \end{cases}$$

As before, existence is a consequence of what was already proved for the **H**-formulation. Concerning uniqueness, by proceeding as in the electric port case we find $\mathrm{curl}\, \mathbf{E} = \mathbf{0}$ in Ω and $\mathbf{E}_C = -V\boldsymbol{\rho}_C$ in Ω_C. Since Ω is simply connected, we also have $\mathbf{E} = \mathrm{grad}\, \psi$ in Ω. Therefore, integrating $\mathbf{E}_C$ on the cycle γ, we find

$$\begin{aligned} 0 &= \int_\gamma \mathrm{grad}\, \psi_C \cdot d\boldsymbol{\tau} = \int_\gamma \mathbf{E}_C \cdot d\boldsymbol{\tau} \\ &= -\int_\gamma V\boldsymbol{\rho}_C \cdot d\boldsymbol{\tau} = -V. \end{aligned}$$

Thus $V = 0$, and consequently $\mathbf{E} = \mathbf{0}$.

Having solved (18), the magnetic field in Ω is as usual defined as

$$\mathbf{H} = -\frac{1}{i\omega}\boldsymbol{\mu}^{-1} \mathrm{curl}\, \mathbf{E}.$$

A similar formulation has been proposed in [23], [8] (in the former paper, by replacing the source $V\boldsymbol{\rho}_C$ by $V \,\mathrm{grad}\, \widetilde{\Phi}_C$, $\widetilde{\Phi}_C$ being a function jumping by 1 through a section Σ of Ω_C). However, in these papers the electric field is not the solution $\mathbf{E}_C$ to $(18)_1$ but is corrected in Ω_C by adding the source term. In this way the Faraday law is no longer verified across the interface Γ.

The same remark applies for the voltage excitation problem of the preceding subsection and the formulations proposed in [21], [23].

5. Numerical approximation. The variational formulations presented in the preceding section can be used as a starting point for devising finite element methods for approximating the solution.

In fact, the voltage excitation problem reduces to a standard problem with a given current density ($V\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_C$ or else $V\boldsymbol{\sigma}\boldsymbol{\rho}_C$); therefore, any method used for eddy-current problems can be applied. Without any attempt at being complete, let us mention only those proposed in [17], [9], [2], [4] for the $\mathbf{H}$-formulation and in [27], [11], [24], [25], [5] for the $\mathbf{E}$-formulation (or for the related magnetic vector potential formulation).

It is worthwhile to note that, when considering the $\mathbf{H}$-formulation, it is not necessary to construct the functions $\mathrm{grad}\,\phi_C$ or $\boldsymbol{\rho}_C$. In fact, to give an example for the electric port case, one can proceed in this way: consider a fixed (and coarse) mesh in Ω_C, and let $\mathcal{I}_*^C$ be the finite element interpolant taking value 0 everywhere, except on Γ_J, where it has value 1. Then define ϕ_* to be the solution to

$$\begin{cases} \mathrm{div}(\boldsymbol{\sigma}\,\mathrm{grad}\,\phi_*) = -\,\mathrm{div}(\boldsymbol{\sigma}\,\mathrm{grad}\,\mathcal{I}_*^C) & \text{in } \Omega_C, \\ \phi_* = 0 & \text{on } \Gamma_E \cup \Gamma_J, \\ \boldsymbol{\sigma}\,\mathrm{grad}\,\phi_* \cdot \mathbf{n} = -\boldsymbol{\sigma}\,\mathrm{grad}\,\mathcal{I}_*^C \cdot \mathbf{n} & \text{on } \Gamma. \end{cases}$$

Thus $\phi_C = \mathcal{I}_*^C + \phi_*$ in Ω_C, and

$$\begin{aligned} \textstyle\int_{\Omega_C} \mathrm{grad}\,\phi_C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C &= \textstyle\int_{\Omega_C} (\mathrm{grad}\,\mathcal{I}_*^C + \mathrm{grad}\,\phi_*) \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C, \\ &= \textstyle\int_{\Omega_C} \mathrm{grad}\,\mathcal{I}_*^C \cdot \mathrm{curl}\,\overline{\mathbf{w}}_C, \end{aligned}$$

as $\mathrm{div}\,\mathrm{curl}\,\overline{\mathbf{w}}_C = 0$, $\phi_* = 0$ on $\Gamma_E \cup \Gamma_J$, and $\mathrm{curl}\,\overline{\mathbf{w}}_C \cdot \mathbf{n} = 0$ on Γ.

Therefore, we have verified that in the $\mathbf{H}$-based variational formulations one can substitute ϕ_C by the easily computable $\mathcal{I}_*^C$, and the solution $\mathbf{H}$ remains the same. Clearly, the need to compute ϕ_C (namely, ϕ_*) comes into play again if one wants to recover $\mathbf{E}_C$, which is given by

$$\mathbf{E}_C = \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C - V\,\mathrm{grad}\,\phi_C = \boldsymbol{\sigma}^{-1}\,\mathrm{curl}\,\mathbf{H}_C - V\,\mathrm{grad}\,\mathcal{I}_*^C - V\,\mathrm{grad}\,\phi_*.$$

If the current intensity is given, the constraint $\int_{\Omega_C} \mathbf{Q}_C \cdot \mathrm{curl}\,\mathbf{H}_C = I$ has to be added (here $\mathbf{Q}_C = \mathrm{grad}\,\phi_C$ or else $\mathbf{Q}_C = \boldsymbol{\rho}_C$). In the $\mathbf{H}$-formulation, the voltage V plays the role of a Lagrange multiplier associated to this constraint, and the global problem is a saddle-point problem. For any type of conforming finite element discretization using edge elements in Ω_C, the presence of this Lagrange multiplier requires that an inf-sup condition like

$$\left| \int_{\Omega_C} \mathbf{Q}_C \cdot \mathrm{curl}\,\mathbf{w}_{C,h}^* \right| \geq \beta ||\mathbf{w}_h^*||_X$$

is satisfied for a constant $\beta > 0$, independent of h, and a suitable discrete vector function $\mathbf{w}_h^*$.

This can be done as follows (for instance, let us focus on the electric port case): expressing $\operatorname{grad} \phi_C$ in terms of $\operatorname{grad} \mathcal{I}^C_*$, as done before, we have by integration by parts and the Stokes theorem

$$\begin{aligned} \int_{\Omega_C} \operatorname{grad} \phi_C \cdot \operatorname{curl} \mathbf{w}^*_{C,h} &= \int_{\Omega_C} \operatorname{grad} \mathcal{I}^C_* \cdot \operatorname{curl} \mathbf{w}^*_{C,h} = \int_{\Gamma_J} \operatorname{curl} \mathbf{w}^*_{C,h} \cdot \mathbf{n} \\ &= \int_{\partial\Gamma_J} \mathbf{w}^*_{C,h} \cdot d\boldsymbol{\tau} = \int_{\partial\Gamma_J} \mathbf{w}^*_{I,h} \cdot d\boldsymbol{\tau}. \end{aligned}$$

Let us consider a fixed (and coarse) mesh in Ω, and let $\mathcal{I}^I_*$ be the finite element interpolant taking value 0 everywhere in Ω_I, except on the cutting surface Ξ, transversal to Γ_D and Γ, where it has a double value, 0 on one side and 1 on the other side (following the orientation of $\partial\Gamma_J$, that is, counterclockwise with respect to $\mathbf{n}$ on Γ_J). From now on we consider triangulations that are all obtained as a refinement of the basic coarse mesh in such a way that a discrete function on the coarse mesh is also a discrete function on all the other meshes. Then choose as $\mathbf{w}^*_{I,h}$ the $(L^2(\Omega_I))^3$-extension of $\operatorname{grad} \mathcal{I}^I_*$, computed in $\Omega_I \setminus \Xi$; note that $\operatorname{grad} \mathcal{I}^I_* \times \mathbf{n}$ is defined in a unique way on Ξ, as the jump of $\mathcal{I}^I_*$ on Ξ is equal to 1. Finally, take as $\mathbf{w}^*_{C,h}$ the edge element interpolant, on the coarse mesh in Ω_C, of the value $\mathbf{w}^*_{I,h} \times \mathbf{n}_I$ on Γ. It is easily checked that with this choice $\int_{\partial\Gamma_J} \mathbf{w}^*_{I,h} \cdot d\boldsymbol{\tau} = 1$ and that the norm $||\mathbf{w}^*_h||_X$ does not depend on h, and therefore the inf-sup condition is satisfied.

Coming to the **E**-formulation, when the current intensity is assigned it takes a nonstandard form: in fact, in (14) and (18) it is questionable if the sesquilinear forms on the left-hand sides are coercive, and, on the other hand, the current intensity condition is not a pure constraint, so that these problems are not saddle-point problems. In this paper we have proved existence and uniqueness of the solution for the infinite dimensional case, but a complete analysis of a finite element approximation method could be a more delicate point. However, this approach was used in [8] for an axisymmetric problem, with good numerical performances.

Remark 5.1. When the current intensity is assigned, it is possible to devise an alternative formulation in terms of a magnetic vector potential and an electric scalar potential, with the Coulomb gauge. Namely, one looks for $\mathbf{A}$ and v_C such that

$$\boldsymbol{\mu}\mathbf{H} = \operatorname{curl} \mathbf{A} \quad \text{in } \Omega, \quad \mathbf{E}_C = -i\omega \mathbf{A}_C - \operatorname{grad} v_C \quad \text{in } \Omega_C,$$

with $\operatorname{div} \mathbf{A} = 0$ in Ω and $\mathbf{A} \cdot \mathbf{n} = 0$ on $\partial\Omega$.

Writing (14) and (18) in terms of these unknowns, and inserting the gauging term as a penalization, as is usually done with this approach, one ends up with a sesquilinear form that can be proved to be coercive (for similar computations, see [14], where the analysis of the $\mathbf{A} - v_C$ method is presented when the excitation is due to a given current density $\mathbf{J}_e$).

Acknowledgments. It is a pleasure to thank Oszkár Bíró and Rafael Vázquez Hernández for some useful conversations about the subject of this paper.

REFERENCES

[1] A. Alonso, *A mathematical justification of the low-frequency heterogeneous time-harmonic Maxwell equations*, Math. Models Methods Appl. Sci., 9 (1999), pp. 475–489.

[2] A. Alonso Rodríguez, P. Fernandes, and A. Valli, *The time-harmonic eddy-current problem in general domains: Solvability via scalar potentials*, in Computational Electromagnetics, C. Carstensen, S. Funken, W. Hackbusch, R. H. W. Hoppe, and P. Monk, eds., Springer-Verlag, Berlin, 2003, pp. 143–163.

[3] A. Alonso Rodríguez, P. Fernandes, and A. Valli, *Weak and strong formulations for the time-harmonic eddy-current problem in general multi-connected domains*, European J. Appl. Math., 14 (2003), pp. 387–406.

[4] A. Alonso Rodríguez, R. Hiptmair, and A. Valli, *Mixed finite element approximation of eddy current problems*, IMA J. Numer. Anal., 24 (2004), pp. 255–271.

[5] A. Alonso Rodríguez and A. Valli, *Mixed finite element approximation of eddy current problems based on the electric field*, in ECCOMAS 2004: European Congress on Computational Methods in Applied Sciences and Engineering, P. Neittaanmäki, T. Rossi, K. Majava, and O. Pironneau, eds., University of Jyvškylä, Jyväskylä, Finland, 2004, http://www.mit.jyu.fi/eccomas2004/proceedings/pdf/185.pdf.

[6] A. Alonso Rodríguez, A. Valli, and R. Vázquez Hernández, *A formulation of the eddy-current problem in the presence of electric ports*, Report UTM 720, Department of Mathematics, University of Trento, Trento, Italy, 2008.

[7] H. Ammari, A. Buffa, and J.-C. Nédélec, *A justification of eddy currents model for the Maxwell equations*, SIAM J. Appl. Math., 60 (2000), pp. 1805–1823.

[8] A. Bermúdez, D. Gómez, M. C. Muñiz, and P. Salgado, *A FEM/BEM for axisymmetric electromagnetic and thermal modelling of induction furnaces*, Internat. J. Numer. Methods Engrg., 71 (2007), pp. 856–878, 879–882.

[9] A. Bermúdez, R. Rodríguez, and P. Salgado, *A finite element method with Lagrange multipliers for low-frequency harmonic Maxwell equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1823–1849.

[10] A. Bermúdez, R. Rodríguez, and P. Salgado, *Numerical solution of eddy-current problems in bounded domains using realistic boundary conditions*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 411–426.

[11] O. Bíró and K. Preis, *On the use of magnetic vector potential in the finite element analysis of three-dimensional eddy currents*, IEEE Trans. Magn., 25 (1989), pp. 3145–3159.

[12] O. Bíró, K. Preis, G. Buchgraber, and I. Tičar, *Voltage-driven coils in finite-element formulations using a current vector and a magnetic scalar potential*, IEEE Trans. Magn., 40 (2004), pp. 1286–1289.

[13] O. Bíró, K. Preis, W. Renhart, G. Vrisk, and K. R. Richter, *Computation of* 3*-D current driven skin effect problems using a current vector potential*, IEEE Trans. Magn., 29 (1993), pp. 1325–1328.

[14] O. Bíro and A. Valli, *The Coulomb gauged vector potential formulation for the eddy-current problem in general geometry: Well-posedness and numerical approximation*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1890–1904.

[15] A. Bossavit, *Computational Electromagnetism. Variational Formulation, Complementarity, Edge Elements*, Academic Press, San Diego, 1998.

[16] A. Bossavit, *Most general 'non-local' boundary conditions for the Maxwell equations in a bounded region*, COMPEL, 19 (2000), pp. 239–245.

[17] A. Bossavit and J. C. Vérité, *A mixed FEM/BIEM method to solve eddy-current problems*, IEEE Trans. Magn., MAG-18 (1982), pp. 431–435.

[18] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of eddy current problems*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 807–831.

[19] P. Dular, *The benefits of nodal and edge elements coupling for discretizing global constraints in dual magnetodynamic formulations*, J. Comput. Appl. Math., 168 (2004), pp. 165–178.

[20] P. Dular, C. Geuzaine, and W. Legros, *A natural method for coupling magnetodynamic* **H***-formulations and circuits equations*, IEEE Trans. Magn., 35 (1999), pp. 1626–1629.

[21] P. Dular, F. Henrotte, and W. Legros, *A general and natural method to define circuits relations associated with magnetic vector potential formulations*, IEEE Trans. Magn., 35 (1999), pp. 1630–1633.

[22] P. Dular, W. Legros, and A. Nicolet, *Coupling of local and global quantities in various finite element formulations and its application to electrostatics, magnetostatics and magnetodynamics*, IEEE Trans. Magn., 34 (1998), pp. 3078–3081.

[23] R. Hiptmair and O. Sterz, *Current and voltage excitations for the eddy current model*, Int. J. Numer. Modelling, 18 (2005), pp. 1–21.

[24] H. Kanayama and F. Kikuchi, 3*-D eddy current computation using Nedelec elements*, Information, 2 (1999), pp. 37–45.

[25] H. Kanayama, D. Tagami, M. Saito, and F. Kikuchi, *A numerical method for* 3*-D eddy current problems*, Japan J. Indust. Appl. Math., 18 (2001), pp. 603–612.

[26] G. Meunier, Y. Le Floch, and C. Guérin, *A nonlinear circuit coupled* $\mathbf{t}$-$\mathbf{t_0}$-ϕ *formulation for solid conductors*, IEEE Trans. Magn., 39 (2003), pp. 1729–1732.

[27] T. Morisue, *Magnetic vector potential and electric scalar potential in three-dimensional eddy current problem*, IEEE Trans. Magn., MAG-18 (1982), pp. 531–535.

[28] J. Rappaz, M. Swierkosz, and C. Trophime, *Un modèle mathématique et numérique pour un logiciel de simulation tridimensionnelle d'induction électromagnétique*, Tech. report 05.99, Département de Mathématiques, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 1999.

[29] G. Rubinacci, A. Tamburrino, and F. Villone, *Circuits/fields coupling and multiply connected domains in integral formulations*, IEEE Trans. Magn., 38 (2002), pp. 581–584.

SIAM J. APPL. MATH.
Vol. 68, No. 5, pp. 1495–1502

GLOBAL STABILITY OF VIRUS SPREADING IN COMPLEX HETEROGENEOUS NETWORKS*

LIN WANG† AND GUAN-ZHONG DAI‡

Abstract. Various networks are possessed of an obvious heterogeneity in the connectivity properties, and it is of practical significance to study epidemic spreading in networks of this kind. Pastor-Satorras and Vespignani established the dynamical mean-field reaction rate equations for the spreading of infections in complex heterogeneous networks based on the well-known SIS model, and figured out an epidemic threshold λ_c such that if λ (effective spreading rate) is above λ_c, the infection spreads and becomes endemic. The significance of this result is far-reaching; however, the authors have not found a strict mathematical proof of their conclusion in the literature. In this paper, we approach this problem by proving that if λ is above λ_c, the infection spreads and approaches the unique positive stationary point of the reaction rate equations as long as there exist infected nodes in the network initially; i.e., the virus infection process is globally stable.

Key words. heterogeneity, spreading of infections, SIS, stability

AMS subject classifications. 92D30, 34D23

DOI. 10.1137/070694582

1. Introduction.

In recent years, complex networks have been widely researched, with research mainly focused on network topology, network evolutionary models, network dynamics, and empirical studies on real networks; see, e.g., [1, 2, 3]. Various networks are complex heterogeneous networks, where there is an obvious heterogeneity in the connectivity properties of network nodes. The node degree distribution of these networks is found, through empirical studies, to follow power law, which implies an unexpected statistical abundance of vertices, so-called hubs, with very large degrees. Examples of such networks include the Internet [4], Reply Networks on Bulletin Board System [5], the network of airline connections [6], and the web of sexual contacts [7], just to list a few which are relevant to epidemic spreading. Thus, it is of practical significance to study epidemic spreading on complex heterogeneous networks.

The extreme heterogeneity of real networks relevant to epidemic spreading implies that traditional epidemiological models such as the SI model, SIS model, and SIR model need scrutiny in the framework of complex heterogeneous networks, and it has been discovered that in heterogeneous networks, these models behave much differently from those in homogeneous ones [8, 6, 9, 10, 11]. The work done by Pastor-Satorras and Vespignani is the famous representative research in this area, and they present a detailed analytical and numerical study on the SIS model in the framework of complex networks [10, 11].

In the SIS model, each node exists in only two discrete states, "healthy" or "infected." At each time step, the susceptible (healthy) node is infected with rate $\nu \times k'$ if it is connected to k' infected nodes. At the same time, infected nodes are cured and become again susceptible with rate δ. We define an effective spreading rate $\lambda = \nu/\delta$.

*Received by the editors June 15, 2007; accepted for publication (in revised form) January 29, 2008; published electronically May 21, 2008. This work was supported by NSF of Shaanxi province in China (grant 2007F14).

http://www.siam.org/journals/siap/68-5/69458.html

†Department of Automation, Xi'an University of Technology, Xi'an 710048, China (wanglin@xaut.edu.cn).

‡Department of Automation, Northwestern Polytechnical University, Xi'an 710072, China (daigz@nwpu.edu.cn).

Without loss of generality, we can set $\delta = 1$. Since there exists a large difference in the degree values of nodes in complex networks, and since nodes with different degree values also have different probabilities of being connected to infected nodes, it is therefore very inaccurate to describe a virus infection process in complex networks with only one unifying dynamical reaction rate equation as in traditional epidemiological models. Pastor-Satorras and Vespignani established the following dynamical mean-field reaction rate equation based on the SIS model [10, 11]:

$$\frac{d\rho_k(t)}{dt} = -\rho_k(t) + \lambda k[1-\rho_k(t)]\Theta(t), \qquad k = 1, 2, \ldots, n, \qquad \lambda > 0, \tag{1.1}$$

where $\rho_k(t)$ denotes the relative density of infected nodes with given degree k, $k = 1, 2, \ldots, n$, n denotes the maximum degree values of all nodes, $P(i) \geq 0$ denotes the density of nodes with given degree i, $\Theta(t) = \langle k\rangle^{-1}\sum_{i=1}^{n} iP(i)\rho_i(t)$, and $\langle k\rangle$ denotes the mean of degree values; i.e., $\langle k\rangle = \sum_{i=1}^{n} iP(i)$. For convenience, we call (1.1) the Satorras–Vespignani (SV) rate equation.

The stationary point of the SV rate equation satisfies the following algebraic equation:

$$-\rho_k + \lambda k\left(1-\rho_k\right)\langle k\rangle^{-1}\sum_{i=1}^{n} iP(i)\rho_i = 0, \qquad \rho_k \geq 0. \tag{1.2}$$

Pastor-Satorra and Vespignani noticed the fact that if and only if $\lambda > \lambda_c = \langle k\rangle / \langle k^2\rangle$, the SV rate equation allows a positive stationary point, and in this case, there exists only one positive stationary point, where $\langle k^2\rangle = \sum_{k=1}^{n} k^2 P(k)$ denotes the second order moment of the node degree distribution. By virtue of this fact, Pastor-Satorras and Vespignani inferred that if $\lambda > \lambda_c$, the infection spreads and becomes endemic; i.e., the SV equation is globally stable. However, so far as the authors know, no current literature presents a theoretical proof of this conclusion.

In this paper, we give a detailed analytical solution to this problem. We prove that if $\lambda > \lambda_c$ and there exist infected nodes initially, i.e., $\sum_{i=1}^{n} iP(i)\rho_i(0) > 0$, the relative density of infected nodes with given connectivity k will approach the unique stationary point of the SV rate equation, i.e., $\lim_{t\to\infty}\rho_k(t) = \rho_k$.

2. Global stability of epidemic spreading. Before stating and proving the main theorem of this paper, we first prove the following lemma.

LEMMA 1. *Suppose that the initial relative infected densities* $0 \leq \rho_k(0) \leq 1$ *satisfy* $\sum_{i=1}^{n} iP(i)\rho_i(0) > 0$*; then as* $t > 0$*, the solution* $\rho_k(t)$ *of* (1.1) *satisfies* $0 < \Theta(t) < 1$, $0 < \rho_k(t) < 1$.

Proof. By (1.1), $\Theta(t)$ satisfies the following equation:

$$\frac{d\Theta(t)}{dt} = -\Theta(t) + \lambda\langle k\rangle^{-1}\sum_{k=1}^{n} k^2 P(k)[1-\rho_k(t)]\Theta(t). \tag{2.1}$$

Since $\Theta(0) = \langle k\rangle^{-1}\sum_{i=1}^{n} iP(i)\rho_i(0) > 0$, by virtue of the Picard–Lindelöf theorem on existence and uniqueness of solutions of differential equations, we know that $\Theta(t) \neq 0$ for any $t > 0$; therefore, as $t > 0$, $\Theta(t) > 0$.

Equation (1.1) can be rewritten as

$$\frac{d\rho_k(t)}{dt} = -\left[1 + \lambda k\Theta(t)\right]\rho_k(t) + \lambda k\Theta(t). \tag{2.2}$$

Since $\Theta(t) > 0$, we have

$$\frac{d\rho_k(t)}{dt} + [1 + \lambda k \Theta(t)]\, \rho_k(t) > 0. \tag{2.3}$$

Multiplying the above inequality by $\exp[t + \lambda k \int_0^t \Theta(s) ds]$ and integrating from 0 to t, we get

$$\rho_k(t) > \rho_k(0) \exp\left[-t - \lambda k \int_0^t \Theta(s) ds\right] \geq 0, \tag{2.4}$$

where $t > 0$.

On the other hand, it can be verified that the function $1 - \rho_k(t)$ satisfies the equation

$$\frac{d\,[1 - \rho_k(t)]}{dt} = -\,[1 + \lambda k \Theta(t)]\,[1 - \rho_k(t)] + 1. \tag{2.5}$$

Similarly to the above proof, we have $1 - \rho_k(t) > 0$. Thus, as $t > 0$, it follows that $0 < \rho_k(t) < 1$. □

The above lemma signifies that if there are infected nodes in the network initially, no matter how many and how distributed, then at any time after the infection process starts, there will appear infected nodes with any given degree.

Now, we prove the main theorem of this paper.

THEOREM 1. *Suppose that the initial relative infected densities* $0 \leq \rho_k(0) \leq 1$ *satisfy* $\sum_{i=1}^n iP(i)\rho_i(0) > 0$ *and that* $\lambda > \langle k \rangle \big/ \langle k^2 \rangle$*; then* $\rho_k(t)$*, the solution of the SV rate equation, satisfies* $\lim_{t\to\infty} \rho_k(t) = \rho_k$*, where* $\rho_1, \rho_2, \ldots, \rho_n$ *are the unique nonzero stationary points of the SV rate equation.*

Proof. In order to prove this theorem, we first prove that the limit $\lim_{t\to\infty} \rho_k(t)$ exists. For this purpose, we have to prove that

$$\liminf_{t\to+\infty} \rho_k(t) = \limsup_{t\to+\infty} \rho_k(t).$$

Letting $u_k^{(1)} = 1$, define the sequence

$$u_k^{(m+1)} = \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m)}}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m)}}, \qquad 1 \leq k \leq n, \qquad m = 1, 2, \ldots. \tag{2.6}$$

According to Lemma 1, for $1 \leq k \leq n$, $\limsup_{t\to+\infty} \rho_k(t) \leq 1 = u_k^{(1)}$. Applying Proposition 2 in the appendix repeatedly, we obtain

$$\limsup_{t\to+\infty} \rho_k(t) \leq u_k^{(m)}, \qquad 1 \leq k \leq n, \qquad m = 1, 2, \ldots. \tag{2.7}$$

Consider the convergence of the sequence defined by (2.6). By (2.6), for all k, $u_k^{(2)} \leq 1 = u_k^{(1)}$. If, for all k, $u_k^{(m+1)} \leq u_k^{(m)}$, it follows from (2.6) that

$$\begin{aligned} \forall k,\ u_k^{(m+2)} &= \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m+1)}}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m+1)}} \\ &\leq \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m)}}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) u_i^{(m)}} = u_k^{(m+1)}. \end{aligned} \tag{2.8}$$

By induction, we know that for each k, the sequence $u_k^{(m)}$ is decreasing, so its limit exists and is denoted by $u_k = \lim_{m\to\infty} u_k^{(m)}$. Leting $m \to \infty$ on both sides of (2.6)–(2.7), we deduce that $u_k = \lim_{m\to\infty} u_k^{(m)}$ satisfies the following relations:

$$\begin{aligned} &-u_k + \lambda k\,(1-u_k)\,\langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)u_i = 0, \\ &\limsup_{t\to+\infty} \rho_k(t) \le u_k, \qquad 1 \le k \le n. \end{aligned} \tag{2.9}$$

On the other hand, we consider the function

$$f(x) = \langle k\rangle^{-1} \sum_{k=1}^{n} \frac{\lambda k^2 P(k)x}{1+\lambda kx} - x. \tag{2.10}$$

By simple calculations, we obtain

$$\begin{aligned} f(0) &= 0, \\ f'(0) &= \langle k\rangle^{-1} \sum_{k=1}^{n} \lambda k^2 P(k) - 1 \\ &= \lambda \langle k\rangle^{-1} \left\langle k^2\right\rangle - 1 > 0. \end{aligned} \tag{2.11}$$

By the definition of derivatives, if $x > 0$ is sufficiently small, then $f(x) > f(0) = 0$.

According to Proposition 1 in the appendix and (2.11), we can take $l_k^{(1)}$ such that

$$\forall k,\ 0 < l_k^{(1)} < \liminf_{t\to\infty} \rho_k(t), \qquad f\left(\langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(1)}\right) > 0. \tag{2.12}$$

We define the following sequence:

$$l_k^{(m+1)} = \frac{\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(m)}}{1+\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(m)}}, \qquad 1 \le k \le n, \qquad m = 1, 2, \ldots. \tag{2.13}$$

By (2.12) and applying Proposition 2 in the appendix repeatedly, we have

$$\liminf_{t\to+\infty} \rho_k(t) \ge l_k^{(m)}, \qquad 1 \le k \le n, \qquad m = 1, 2, \ldots. \tag{2.14}$$

Now consider the convergence of the sequence defined by (2.13). First, according to (2.10), (2.12), and (2.13), we have

$$\langle k\rangle^{-1} \sum_{k=1}^{n} kP(k)l_i^{(2)} > \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(1)}. \tag{2.15}$$

By (2.13) and (2.15), we obtain

$$\begin{aligned} \forall k,\ l_k^{(3)} &= \frac{\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(2)}}{1+\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(2)}} \\ &> \frac{\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(1)}}{1+\lambda k \langle k\rangle^{-1} \sum_{i=1}^{n} iP(i)l_i^{(1)}} = l_k^{(2)}. \end{aligned} \tag{2.16}$$

If for all k, $l_k^{(m+1)} > l_k^{(m)}$, it follows from (2.13) that

$$\begin{aligned}\forall k,\ l_k^{(m+2)} &= \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) l_i^{(m+1)}}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) l_i^{(m+1)}} \\ &> \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) l_i^{(m)}}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^n iP(i) l_i^{(m)}} = l_k^{(m+1)}. \qquad (2.17)\end{aligned}$$

Thus, by induction, we know that, for each k, the sequence $l_k^{(m)}$, $m \geq 2$, is increasing, so its limit exists and is denoted by $l_k = \lim_{m\to\infty} l_k^{(m)}$. Letting $m \to \infty$ on both sides of (2.13)–(2.14), we deduce that the limit $l_k = \lim_{m\to\infty} l_k^{(m)}$ satisfies the following relations:

$$\begin{aligned} &-l_k + \lambda k (1 - l_k) \langle k \rangle^{-1} \sum_{i=1}^n iP(i) l_i = 0, \\ &l_k \leq \liminf_{t\to+\infty} \rho_k(t), \qquad 1 \leq k \leq n. \qquad (2.18)\end{aligned}$$

By (2.9) and (2.18), both $u_k = \lim_{m\to\infty} u_k^{(m)}$ and $l_k = \lim_{m\to\infty} l_k^{(m)}$ are positive stationary points of the SV rate equation; thus by the uniqueness of the positive stationary point of the SV rate equation, we have that $u_k = l_k = \rho_k$ and

$$\rho_k \leq \liminf_{t\to+\infty} \rho_k(t) \leq \limsup_{t\to+\infty} \rho_k(t) \leq \rho_k, \qquad 1 \leq k \leq n. \qquad (2.19)$$

That is, $\lim_{t\to\infty} \rho_k(t) = \rho_k$, and Theorem 1 is proved. □

Theorem 1 indicates that if $\lambda > \langle k \rangle / \langle k^2 \rangle$ and there exist infected nodes (no matter how few) in the network initially, then after a duration of transitions between healthy nodes and infected nodes, the infection process becomes stable and the relative density $\rho_k(t)$ of infected nodes with given degree k approaches the unique positive stationary point of the SV rate equation.

3. Conclusions. In this paper, we present a strict proof of global stability of dynamical mean-field reaction rate equations for the spreading of infections in complex heterogeneous networks based on the well-known SIS model established by Pastor-Satorras and Vespignani. According to our theorem, if the virus spread speed is above the threshold ($\langle k \rangle / \langle k^2 \rangle$), then as long as there exist infected nodes in the system, infections will spread, and the eventual proportion of infected nodes with given degree is independent of the initial number of infected nodes; i.e., the infection process is globally stable.

Appendix. Proof of two propositions. In the appendix, we will prove some properties about the solution and stationary point of the SV rate equation:

$$\frac{d\rho_k(t)}{dt} = -\rho_k(t) + \lambda k[1 - \rho_k(t)]\Theta(t), \qquad k = 1, 2, \ldots, n, \qquad \lambda > 0, \qquad (A.1)$$

where $\Theta(t) = \langle k \rangle^{-1} \sum_{i=1}^n iP(i)\rho_i(t)$.

PROPOSITION 1. *Suppose that the initial relative infected densities* $0 \leq \rho_k(0) \leq 1$ *satisfy* $\sum_{i=1}^n iP(i)\rho_i(0) > 0$ *and that* $\lambda > \langle k \rangle / \langle k^2 \rangle$*; then the solution* $\rho_k(t)$ *of* (A.1) *satisfies* $\inf_{t\geq 0} \Theta(t) > 0$*; for any* $\tau > 0$*,* $\inf_{t\geq\tau} \rho_k(t) > 0$.

Proof. According to Lemma 1, as $t \geq 0$, $\Theta(t) > 0$. By (2.1), we have

$$\frac{\frac{d\Theta(t)}{dt}}{\Theta(t)} = \left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda \langle k \rangle^{-1} \sum_{k=1}^{n} k^2 P(k) \rho_k(t). \tag{A.2}$$

It follows easily that

$$\frac{\frac{d\Theta(t)}{dt}}{\Theta(t)} \geq \left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda n \Theta(t). \tag{A.3}$$

We will prove that there exists $\bar{t} > 0$ such that as $t \geq \bar{t}$,

$$\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda n \Theta(t) < \frac{\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2}. \tag{A.4}$$

First notice that there exists $\bar{t} > 0$ such that as $t = \bar{t}$, (A.9) holds. Otherwise, for all $t > 0$, the following holds:

$$\begin{aligned} \frac{\frac{d\Theta(t)}{dt}}{\Theta(t)} &\geq \left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda n \Theta(t) \\ &\geq \frac{\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2}. \end{aligned} \tag{A.5}$$

By integrating on both sides of (A.5), we have, for all $t > 0$,

$$\Theta(t) \geq \Theta(0) \exp\left[\frac{t\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2}\right] \to \infty,$$

which contradicts $\Theta(t) < 1$. So, there exists $\bar{t} > 0$, such that as $t = \bar{t}$, (A.4) holds. Now, we further prove that as $t \geq \bar{t}$, (A.4) holds. If this were not true, there would exist $t_1 > \bar{t}$ such that

$$t_1 = \inf\left\{ t > \tau \,\middle|\, \left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda n \Theta(t) \geq \frac{\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2} \right\}. \tag{A.6}$$

Thus, as $0 < t < t_1$, (A.4) holds, while as $t = t_1$,

$$\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right) - \lambda n \Theta(t) = \frac{\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2}. \tag{A.7}$$

By (A.3), we have

$$\left.\frac{d\Theta(t)}{dt}\right|_{t=t_1} \geq \Theta(t_1) \frac{\left(\lambda \frac{\langle k^2 \rangle}{\langle k \rangle} - 1\right)}{2} > 0. \tag{A.8}$$

By the definition of derivatives, as t approaches t_1 from the left of t_1, $\Theta(t) < \Theta(t_1)$, so

$$\begin{aligned}\left(\lambda\frac{\langle k^2\rangle}{\langle k\rangle}-1\right)-\lambda n\Theta(t) &> \left(\lambda\frac{\langle k^2\rangle}{\langle k\rangle}-1\right)-\lambda n\Theta(t_1)\\ &= \frac{\left(\lambda\frac{\langle k^2\rangle}{\langle k\rangle}-1\right)}{2}, \qquad \text{(A.9)}\end{aligned}$$

which contradicts the definition of t_1, so, as $t \geq \bar{t}$, (A.4) holds. That is, as $t \geq \bar{t}$, the following holds:

$$\Theta(t) > \frac{1}{2\lambda n}\left(\lambda\frac{\langle k^2\rangle}{\langle k\rangle}-1\right). \qquad \text{(A.10)}$$

By virtue of Lemma 1 and continuity of $\Theta(t)$, we have

$$\inf_{t\geq 0}\Theta(t) = \min\left[\inf_{0\leq t\leq \bar{t}}\Theta(t), \inf_{t>\bar{t}}\Theta(t)\right] > 0. \qquad \text{(A.11)}$$

We prove the second conclusion next. Letting $\sigma = \inf_{t\geq 0}\Theta(t)$, by (A.1), as $t > 0$, the following holds:

$$\begin{aligned}\frac{d\rho_k(t)}{dt} &= -\left(1+\lambda k\Theta(t)\right)\rho_k(t)+\lambda k\Theta(t)\\ &\geq -\left(1+\lambda k\right)\rho_k(t)+\lambda k\sigma. \qquad \text{(A.12)}\end{aligned}$$

Multiplying both sides of the above inequality by $\exp\left[\left(1+\lambda k\right)t\right]$ and taking integration, we have

$$\begin{aligned}\rho_k(t) \geq\ & \exp\left[-\left(1+\lambda k\right)t\right]\rho_k(0)\\ &+\frac{\lambda k\sigma}{1+\lambda k}\left(1-\exp\left[-\left(1+\lambda k\right)t\right]\right), \qquad \text{(A.13)}\end{aligned}$$

from which we know that, for any $\tau > 0$, $\inf_{t\geq\tau}\rho_k(t) > 0$. □

The main theorem of this paper is proved by virtue of the following proposition.

PROPOSITION 2. *Suppose the solution* $\rho_k(t)$ *of* (A1) *satisfies* $\limsup_{t\to+\infty}\rho_k(t) \leq u_k$ *and* $\liminf_{t\to+\infty}\rho_k(t) \geq l_k$*, where* $u_k \geq 0$*,* $l_k \geq 0$*; then*

$$\begin{aligned}\limsup_{t\to+\infty}\rho_k(t) &\leq \frac{\frac{\lambda k}{\langle k\rangle}\sum_{i=1}^{n} iP(i)u_i}{1+\frac{\lambda k}{\langle k\rangle}\sum_{i=1}^{n} iP(i)u_i},\\ \liminf_{t\to+\infty}\rho_k(t) &\geq \frac{\frac{\lambda k}{\langle k\rangle}\sum_{i=1}^{n} iP(i)l_i}{1+\frac{\lambda k}{\langle k\rangle}\sum_{i=1}^{n} iP(i)l_i}. \qquad \text{(A.14)}\end{aligned}$$

Proof. Since $\limsup_{t\to+\infty}\rho_k(t) \leq u_k$, for all $\varepsilon > 0$, there exists $\tau > 0$ such that as $t \geq \tau$, $\rho_k(t) \leq u_k+\varepsilon$. According to Lemma 1, we have $1-\rho_k(t) > 0$, so, by (A.1), we know that, as $t \geq \tau$, the following holds:

$$\frac{d\rho_k(t)}{dt} \leq -\rho_k(t)+\lambda k[1-\rho_k(t)]\langle k\rangle^{-1}\sum_{i=1}^{n} iP(i)(u_i+\varepsilon). \qquad \text{(A.15)}$$

It follows easily that

$$\begin{aligned}\frac{d\rho_k(t)}{dt} \leq & -\left[1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)\right] \rho_k(t) \\ & + \lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon).\end{aligned} \tag{A.16}$$

It follows from (A.16) that

$$\begin{aligned}\rho_k(t) \leq & \frac{\rho_k(\tau)}{\exp\left\{\left(1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)\right)(t - \tau)\right\}} \\ & + \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)}.\end{aligned} \tag{A.17}$$

Taking the limit as $t \to +\infty$, we obtain

$$\limsup_{t \to +\infty} \rho_k(t) \leq \frac{\lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)}{1 + \lambda k \langle k \rangle^{-1} \sum_{i=1}^{n} iP(i)(u_i + \varepsilon)}. \tag{A.18}$$

Letting $\varepsilon \to +\infty$, we obtain the first inequality in (A.14). The second inequality in (A.14) can be proved similarly. □

REFERENCES

[1] R. Albert and A.-L. Barabasi, *Statistical mechanics of complex networks*, Rev. Modern Phys., 74 (2002), pp. 47–97.

[2] M. E. J. Newmann, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256.

[3] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of networks*, Adv. in Phys., 51 (2002), pp. 1079–1187.

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *On power-law relationships of the Internet topology*, ACM SIGCOMM Computer Communication Review, 29 (1999), pp. 251–262.

[5] Z. Kou and C. Zhang, *Reply networks on bulletin board system*, Phys. Rev. E, 67 (2003), pp. 036117/1–036117/6.

[6] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, *Dynamical patterns of epidemic outbreaks in complex heterogeneous networks*, J. Theoret. Bio., 235 (2005), pp. 275–288.

[7] J. H. Jones and M. S. Handcock, *Sexual contacts and epidemic thresholds*, Nature, 423 (2003), pp. 605–606.

[8] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, *The architecture of complex weighted networks*, Proc. Nat. Acad. Sci. USA, 101 (2004), pp. 3747–3752.

[9] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, *Absence of epidemic threshold in scale-free networks with degree correlations*, Phys. Rev. Lett., 90 (2003), pp. 028701/1–028701/4.

[10] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale free networks*, Phys. Rev. Lett., 86 (2001), pp. 3200–3203.

[11] R. Pastor-Satorras and A. Vespignani, *Epidemic dynamics in finite size scale-free networks*, Phys. Rev. E, 65 (2002), pp. 035108/1–035108/4.

SIAM J. Appl. Math.
Vol. 68, No. 6, pp. 1503–1517

ELECTRIC DISCHARGE SINTERING: A MATHEMATICAL MODEL[*]

G. A. KRIEGSMANN[†]

Abstract. In this paper we mathematically model the densification of metallic powders and the sintering of ceramic powders by electric discharge. The ordinary and partial differential equations governing these processes are the same with the exception of the effective electrical conductivity. This function is a monotonically decreasing (increasing) function of temperature for the metallic (ceramic) powders. We employ asymptotic methods to approximate the solution to these equations in the limit as $\epsilon \to 0$, where ϵ is the ratio of the discharge to diffusion time scales. We find on the shortest time scale that the temperature, voltage, and density satisfy a system of nonlinear, coupled ordinary equations. We solve these and find the relationship between the temperature and density, as functions of the input energy. The results on the short or discharge time scale do not take into account diffusion and heat loss into the surrounding medium. These occur on a much longer time scale, which we identify and exploit to deduce a new approximation. On this time scale the capacitor has no more energy to deposit into the powder. The temperature relaxes to that of its surroundings and the density increases to its final value. Our results show the functional relationship between the final density and the initial energy stored in the capacitor, as well as the initial density of the powder.

Key words. heat transfer, electric discharge, asymptotics, partial differential equations, differential equations

AMS subject classifications. 34E10, 34E13, 35K20, 78A30

DOI. 10.1137/070706689

1. Introduction.

Over the last several years researchers have used electric discharges to sinter ceramic materials and to compact metal powders [1, 2, 3]. The experimental configurations have cylindrical dies, which contain a powdered material. The green powders are compacted with punches which also act as electrodes through which a short, powerful current pulse is passed. The source of the current is a capacitor bank that is charged to a few hundred volts and in some applications to metal powders, 20 kilovolts [2]. In all cases the dies are surrounded by a thermally insulating layer such as alumina, Al_2O_3. The short current pulse passing through the powder rapidly heats it to a temperature at which the ceramic powder sinters or the metallic powders meld together. A schematic of the experimental set-up is shown in Figure 1.

In this paper we introduce a mathematical model to describe this sintering or densification process. Specifically, our model takes the form of a heat equation, a Laplace equation describing the electric potential, and an evolution equation describing the densification of the powder. All of these equations are coupled and nonlinear. In addition, there is a differential equation relating the change in the capacitor voltage across the punches to the current flowing through them. Implicit in this mathematical description is the assumption that the powdered material can be treated as a continuum.

Our mathematical model is very similar to the one developed and analyzed to describe temperature surges in thermistors [4]. There are two obvious differences; the first is the addition of the evolution equation for the density, and the second, the

[*] Received by the editors October 29, 2007; accepted for publication February 19, 2008; published electronically May 28, 2008. This work was sponsored by the Department of Energy, under grant DE-FG02-04ER25654, and the National Science Foundation, under grant DMS-0708065.
http://www.siam.org/journals/siap/68-6/70668.html

[†] Department of Mathematical Sciences, Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 (grkrie@oak.njit.edu).

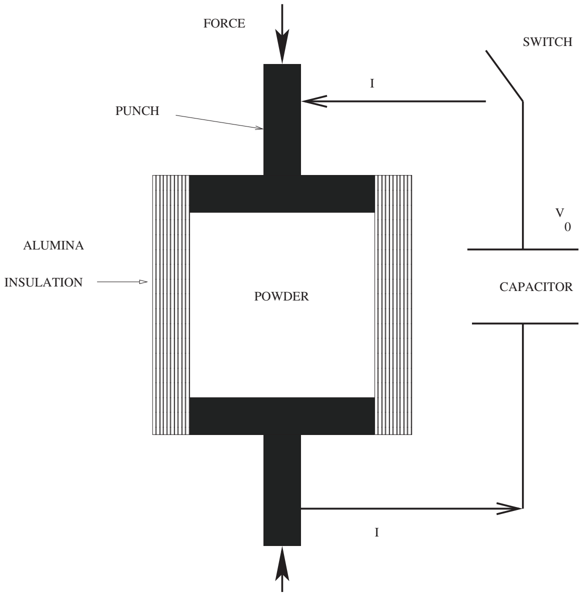


Fig. 1. *Schematic of experimental set-up.*

equation for the capacitor voltage. In reference [4] the external circuit was a battery and resistor in series with the thermistor. More subtle and important differences are the time scales. These produce different physical balances and hence different physics. For example, the time scale describing the source in our heat equation is commensurate with the capacitive discharge time scale; both are very small compared with the diffusion time scale along the axis of the experimental device. In the thermistor model the source and diffusion time scales are of the same order; the result is a nonlinear, nonlocal heat equation.

Denoting the ratio of the source to diffusion time scales by ϵ, we perform an asymptotic analysis of the solutions to our equations as $\epsilon \to 0$. Assuming that the sintering dynamics occur on the same time scale as the source, our analysis shows that the temperature, voltage, and density satisfy a system of three nonlinear ordinary differential equations to leading order. These are analyzed, and the qualitative features of the solutions are described. In particular, the capacitor voltage decays to zero, the density reaches a maximum, and the temperature achieves a maximum, too; all depend in a critical fashion on the initial energy stored in the capacitor. These results are valid only on the source time scale which we call the heating regime. On the much longer convective time scale, the thermal energy decays as the sample loses heat to the surrounding medium through the metallic punches. In this span of time the temperature of the sample decays to that of the surrounding environment. On this time scale, diffusion plays a minor role, yielding an $O(\epsilon^2)$ spatial variation along the axis of the experimental device.

If the sintering or the densification dynamics occur on the convective time scale, the results are more complicated. In the heating regime, the voltage again decays to zero, the density remains at its initial value, and the temperature achieves a maximum. This maximum depends upon the initial density and the initial energy stored in the capacitor. These results again are valid only in the heating regime. On the convective time scale the density increases to its final value, and the temperature decays to that of the surrounding medium. The final density depends critically upon its initial value and the energy stored in the capacitor. Our results explicitly show this dependence.

We close this introduction by giving a brief outline of the paper. Section 2 contains the mathematical formulation of our model. In section 3 a careful dimensional analysis is given identifying the relevant time scales for our problem. Section 4 contains our asymptotic analyses in both the heating and convective regimes. Here the time scale associated with the sintering dynamics is taken to be commensurate with the source time scale. In section 5 we present our analysis for the slow densification case. Here the sintering time scale is commensurate with the convective time scale. Finally, section 6 contains our conclusions.

2. Problem formulation. In Figure 1 we show a schematic of an electrodischarge consolidation experiment. The powder specimen to be densified is contained in an electrically nonconducting cylinder, of radius R_0 and height L, surrounded by a layer of alumina. Two punches compress the powder through an external force F. The punches also serve as electrodes connected to a capacitor charged to an initial voltage V_0 which can be on the order of a few hundred to several thousand volts. When the switch is closed the capacitor discharges and creates a current, I, which flows through the powder specimen. This rapidly heats the powder and causes it to sinter into a dense material.

We model the compacted powder as a continuum with a temperature T, defined inside the cylindrical region $0 < Z < L$, $0 < R < R_0$, satisfying

$$\frac{\partial}{\partial t}[\rho C_p T] = \nabla \cdot (K \nabla T) + \sigma(T)|\nabla \Phi|^2, \tag{1a}$$

where ρ is an averaged density, C_p the thermal capacity, K the thermal conductivity, and σ the electrical conductivity of the powder, which all can depend upon the temperature and the external force F, with the exception of C_p. In this equation Φ is the electric potential, caused by the discharging capacitor, which satisfies in the same region

$$\nabla \cdot (\sigma \nabla \Phi) = 0. \tag{1b}$$

Denoting the voltage across the capacitor as $V(t)$ and the current flowing from it as $I(t)$, we have from Kirchoff's current law

$$C\frac{dV}{dt} = -I, \tag{1c}$$

where C is the capacitance.

We shall now determine a relationship between the current, the temperature, and the potential. First, we recall that the current density is defined by $\mathbf{J} = \sigma \nabla \Phi$, where $\nabla \Phi$ is the electric field. Next, we assume that the R component of this current density vanishes at $R = R_0$, that is, $\sigma \frac{\partial \Phi}{\partial R} = 0$ there. This assumes no current flows out of the powder into the insulation. Next, we integrate (1b) over the cylindrical domain $Z_0 < Z < Z_1$, $0 < R < R_0$, where Z_0 and Z_1 are arbitrary, apply the divergence theorem, and use $\Phi_R = 0$ at $R = R_0$ to obtain

$$\iint_{Z=Z_1} \sigma \frac{\partial \Phi}{\partial Z}\, dX\, dY = \iint_{Z=Z_0} \sigma \frac{\partial \Phi}{\partial Z}\, dX\, dY.$$

This shows that the integral of $\sigma\Phi_Z$, the Z component of the current density, is independent of Z and hence a constant. This constant is the current in (1c) and is defined by

$$I = \iint_\Omega \sigma \frac{\partial \Phi}{\partial Z}\, dX\, dY, \tag{1d}$$

where Ω is an arbitrary circular cross-section of the powder sample.

Finally, to close these equations we must model the evolution of the powder density as it sinters. We assume the phenomenological equation

$$\frac{\partial \rho}{\partial t} = G(T)[\rho_1 - \rho] \tag{1e}$$

qualitatively describes this evolution [5], where the reaction rate G depends upon temperature and ρ_1 is the bulk density for a single species powder, or an average bulk density of a multispecies powder.

The boundary conditions we apply are

$$\frac{\partial T}{\partial R} = 0, \qquad \frac{\partial \Phi}{\partial R} = 0, \quad R = R_0, \quad 0 < Z < L, \tag{2a}$$

where the first assumes a perfect thermal insulation and the second is a restatement that no current flows radially out of the sample,

$$\Phi(X,Y,0,t) = 0, \quad \Phi(X,Y,L,t) = V(t), \tag{2b}$$

which relates the potential to the capacitor voltage, and

$$K\frac{\partial T}{\partial Z} - h(T - T_A) = 0, \quad Z = 0, \qquad K\frac{\partial T}{\partial Z} + h(T - T_A) = 0, \quad Z = L, \qquad 0 < R < R_0. \tag{2c}$$

Here, we thermally model the punches by a linear Newton law of cooling where T_A is the ambient temperature of the surrounding air and h is the effective heat transfer coefficient. This boundary condition removes the need to study heat diffusion in the punch and ultimately heat convection from the punch into the surrounding medium. It is valid when the mass of the punch is small and its thermal conductivity is large. Such are the cases in [1, 2, 3].

Finally, the initial conditions required are

$$T = T_A, \quad 0 < R < R_0, \quad 0 < Z < L, \quad V(0) = V_0, \quad \rho(0) = \rho_0(F), \tag{2d}$$

where V_0 is the initial voltage of the charged capacitor and $\rho_0(F)$ is the initial density of the powder which depends upon the forces on the punches.

3. Nondimensional analysis.

We begin this section by introducing the dimensionless dependent and independent variables

$$u = \frac{T}{T_A} - 1, \quad \phi = \frac{\Phi}{V_0}, \quad v = \frac{V}{V_0}, \quad i = \frac{I}{\sigma_A V_0 L}, \quad w = \frac{\rho}{\rho_1} \tag{3a}$$

and

$$\tau = \frac{t}{\theta_S}, \quad \mathbf{x} = \frac{\mathbf{X}}{L}, \quad r = \frac{R}{L}, \tag{3b}$$

respectively. Here and henceforth, the subscript A denotes that a function is evaluated at the ambient temperature, and $\theta_S = \frac{C_p \rho_1 T_A L^2}{\sigma_A V_0^2}$ is the time scale associated with the capacitive source. We also define the dimensionless functions

$$k(u) = \frac{K(T)}{K_A}, \quad f(u) = \frac{\sigma(T)}{\sigma_A}, \quad g(u) = \frac{G(T)}{G_A}. \tag{3c}$$

Introducing (3a)–(3c) into (1a)–(1e) and combining (1c) and (1d), we find that our dimensionless equations are

$$\frac{\partial[wu]}{\partial\tau} = \epsilon\nabla\cdot(k\nabla u) + f(u)|\nabla\phi|^2, \quad 0 < z < 1, \quad 0 < r < r_0 = \frac{R_0}{L}, \quad \tau > 0, \tag{4a}$$

$$\nabla\cdot(f\nabla\phi) = 0, \quad 0 < z < 1, \quad 0 < r < r_0, \quad \tau > 0, \tag{4b}$$

$$\frac{dv}{d\tau} = -\lambda \iint_{r<r_0} f(u)\frac{\partial\phi}{\partial z}\,dx\,dy, \tag{4c}$$

$$\frac{\partial w}{\partial\tau} = \gamma g(u)(1-w). \tag{4d}$$

Here we have defined three new parameters, each a ratio of time scales. The first is $\lambda = \theta_S/\theta_C$, where $\theta_C = C/\sigma_A L$ is the capacitive time scale. The second is $\gamma = \theta_S G_A$. We assume here that both these parameters are order one quantities. On the other hand, the third parameter $\epsilon = \theta_S/\theta_Z << 1$, where $\theta_Z = \rho_0 C_p/K_A L^2$ is the diffusive time scale in the Z direction. That is, we are assuming that heat diffuses much more slowly than the source time scale. The corresponding boundary conditions become

$$\frac{\partial u}{\partial r} = 0, \qquad \frac{\partial\phi}{\partial r} = 0, \quad r = r_0, \quad 0 < z < 1, \tag{5a}$$

$$\phi(x,y,0,\tau) = 0, \qquad \phi(x,y,1,\tau) = v(t), \quad 0 < r < r_0, \tag{5b}$$

$$k(u)\frac{\partial u}{\partial z} - Bu = 0, \quad z = 0, \qquad k(u)\frac{\partial u}{\partial z} + Bu = 0, \quad z = 1. \tag{5c}$$

In (5c) the Biot number is defined by $B = \theta_Z/\theta_{con}$, where $\theta_{con} = \rho_1 C_p L/h$ is the convective time scale at which the sample cools. The Biot number is also a small parameter for $L \sim$ several centimeters. We relate it to ϵ by $B = \beta\epsilon$, where $\beta = (\theta_Z/\theta_{con})(\theta_Z/\theta_S) = O(1)$. Finally, the dimensionless initial conditions are

$$u = 0, \quad 0 < r < r_0, \quad 0 < z < 1, \qquad v(0) = 1, \qquad w(0) = \frac{\rho_0(F)}{\rho_1}. \tag{5d}$$

Equations (4)–(5) constitute our dimensionless initial boundary value problem.

4. Analysis. This section contains two parts. In the first we analyze the evolution of u, v, and w on a time scale where $\tau = O(1)$. In this time frame diffusion and convective heat losses play a minor role as the powder sample rapidly heats. We shall provide a leading order asymptotic analysis of our problem. This result becomes nonuniform for large τ when diffusion and convection become important. The analysis on a much larger time scale is the subject of the second part of this section.

4.1. The heating regime.

We begin our analysis with the observation that the boundary data and initial conditions are all independent of r and the angle θ. Thus, we shall seek a solution of (4)–(5) that depends only upon z and τ. Accordingly, (4a)–(4c) become

$$\frac{\partial[uw]}{\partial\tau} = \epsilon\frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right) + f(u)\left|\frac{\partial\phi}{\partial z}\right|^2, \quad 0 < z < 1, \quad \tau > 0, \tag{6a}$$

$$\frac{\partial}{\partial z}\left(f(u)\frac{\partial\phi}{\partial z}\right) = 0, \quad 0 < z < 1, \quad \tau > 0, \tag{6b}$$

$$\frac{dv}{d\tau} = -\Lambda f(u)\frac{\partial\phi}{\partial z}, \tag{6c}$$

where $\Lambda = \pi r_0^2\lambda$. Equation (4d) remains the same as do the boundary and initial equations, except now (5c) is replaced by

$$k(u)\frac{\partial u}{\partial z} - \beta\epsilon u = 0, \quad z = 0, \qquad k(u)\frac{\partial u}{\partial z} + \beta\epsilon u = 0, \quad z = 1. \tag{6d}$$

We now exploit the smallness of the parameter ϵ and seek an asymptotic solution of the form

$$u \sim u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots \tag{7a}$$

$$\phi{\sim}\phi_0 + \epsilon\phi_1 + \epsilon^2\phi_2 + \cdots \tag{7b}$$

$$v \sim v_0 + \epsilon v_1 + \epsilon^2 v_2 + \cdots \tag{7c}$$

$$w \sim w_0 + \epsilon w_1 + \epsilon^2 w_2 + \cdots . \tag{7d}$$

Inserting these expansions into our equations, boundary, and initial conditions, we find to leading order that u_0, v_0, ϕ_0, and w_0 satisfy

$$\frac{\partial[u_0w_0]}{\partial\tau} = f(u_0)\left|\frac{\partial\phi_0}{\partial z}\right|^2, \quad \frac{\partial u_0}{\partial z} = 0,\ z = 0, 1, \quad u_0(z,0) = 0, \tag{8a}$$

$$\frac{\partial}{\partial z}\left(f(u_0)\frac{\partial\phi_0}{\partial z}\right) = 0, \quad \phi_0(0,\tau) = 0, \quad \phi(1,\tau) = v(\tau), \quad \tau > 0, \tag{8b}$$

$$\frac{dv_0}{d\tau} = -\Lambda f(u_0)\frac{\partial\phi_0}{\partial z}, \quad v_0(0) = 1, \tag{8c}$$

$$\frac{\partial w_0}{\partial\tau} = \gamma g(u_0)(1 - w_0), \quad w_0(0) = \frac{\rho_0(F)}{\rho_1}. \tag{8d}$$

We now integrate (8b) twice and employ the boundary conditions at $z = 0$ and $z = 1$ to find

$$\phi_0 = v_0(\tau)\,\frac{Q(z)}{Q(1)}, \quad Q(z) = \int_0^z \frac{1}{f(u_0)}\,dz'.$$

Now from this result we observe that $\frac{\partial\phi_0}{\partial z} = v_0/f(u_0)Q(1)$, which is independent of z. Thus, the right-hand side of (8a) is independent of z, too, and this implies that u_0 and w_0 are functions of τ alone. This observation yields

$$\phi_0 = z\,v_0(\tau), \tag{9}$$

and upon inserting this result into (8a), (8c), and (8d) we arrive at the third order system of ordinary differential equations

$$\text{(10a)} \qquad \frac{d[u_0 w_0]}{d\tau} = f(u_0)v_0^2, \quad u_0(0) = 0,$$

$$\text{(10b)} \qquad \frac{dv_0}{d\tau} = -\Lambda f(u_0)v_0, \quad v_0(0) = 1,$$

$$\text{(10c)} \qquad \frac{dw_0}{d\tau} = \gamma g(u_0)[1 - w_0], \quad w_0(0) = \frac{\rho_0(F)}{\rho_1}.$$

We next divide (10a) by (10b), integrate this result, and use the initial conditions to obtain the first integral

$$\text{(11)} \qquad v_0^2 = 1 - 2\Lambda u_0 w_0.$$

Combining this with (10a), differentiating the left-hand side, and using (10c) to eliminate the derivative of w_0, we find

$$\text{(12)} \qquad \frac{du_0}{d\tau} = \frac{1}{w_0}\{(1 - 2\Lambda u_0 w_0)f(u_0) - \gamma u_0 g(u_0)(1 - w_0)\}, \quad w_0(0) = \frac{\rho_0(F)}{\rho_1},$$

and this combined with (10c) yields a second order system, which describes the leading order heating process.

We begin by noting that $f(u)$ is a positive function; it may increase or decrease with u depending upon the powdered material. If we assume for the moment that $g(u)$ is also a positive function, then a simple phase plane analysis of our system reveals that u_0 and w_0 both evolve from their initial conditions $(0, \rho_0/\rho_1)$ to their steady state values (u_0^*, w_0^*), where $u_0^* = 1/2\Lambda$ and $w_0^* = 1$, as $\tau \to \infty$. Coupling this information with (10b) it is easy to deduce that v_0 decreases monotonically from its initial value of 1 to its steady state $v_0^* = 0$ as $\tau \to \infty$. It is interesting to note that the final state is independent of the dimensionless electrical conductivity $f(u)$ and the dimensionless reaction rate $g(u)$. These functions will of course affect the rate at which u_0, w_0, and v_0 approach their steady state values.

A more physically realistic assumption about the function $g(u)$ is to require it to vanish for $u < u_T$ and increase monotonically and smoothly for $u > u_T$. Here u_T denotes a dimensionless threshold temperature. This characterization is equivalent to assuming that no sintering or densification occurs below this temperature. For simplicity and concreteness of presentation we take

$$g(u) = \begin{cases} 0, & 0 \le u < u_T, \\ \tanh^2(\nu(u - u_T)), & u_T \ge u, \end{cases}$$

where ν controls how rapidly g switches. For this choice of g, or any other smooth switching function, there will be two cases to consider: low power, where $\frac{1}{2\Lambda} < u_T$, and high power, where $\frac{1}{2\Lambda} > u_T$.

We begin by considering the high power case. Figure 2a shows two trajectories for different values of $w_0(0) = \rho_0/\rho_1$. Both trajectories remain horizontal, where w_0 is essentially fixed at its initial value, until they cross the threshold u_T, where w_0 increases with u_0. This continues until the right-hand side of (12) changes sign and u_0 decreases with increasing w_0. The functions u_0 and w_0 then approach the steady state

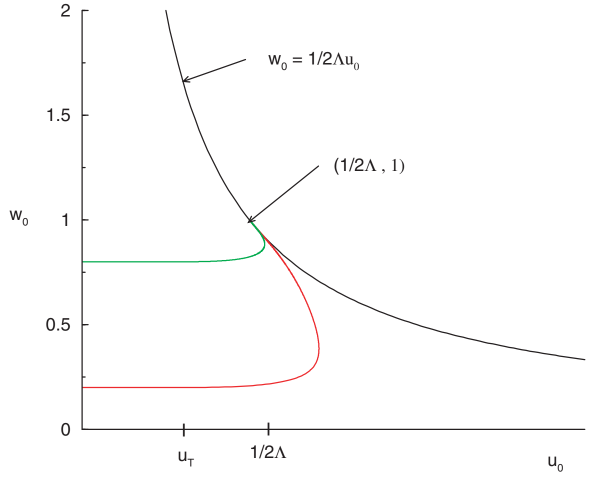


FIG. 2a. *Phase plane: High power.*

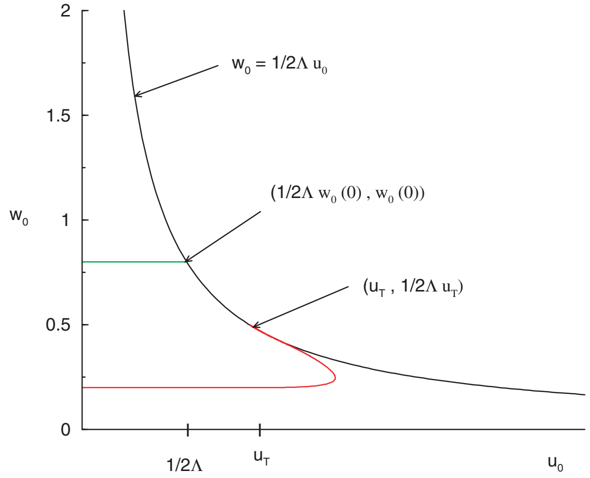


FIG. 2b. *Phase plane: Low power.*

values, $u_0^* = 1/2\Lambda$ and $w_0^* = 1$, respectively. The function v_0 converges to $v_0^* = 0$, as can be deduced from (11).

We next consider the case of low power, where the dynamics are more interesting. Figure 2b also shows two trajectories for different values of $w_0(0)$. For the larger initial condition the trajectory never crosses the threshold temperature u_T, remains horizontal, and approaches the hyperbola $w_0 = 1/2\Lambda u_0$ as $\tau \to \infty$. There $w_0^* = w_0(0)$ and $u_0^* = 1/2\Lambda w_0(0)$. The trajectory for the lower initial condition remains horizontal until it crosses u_T, where w_0 increases with u_0. This continues until the right-hand side of (12) changes sign and u_0 decreases with increasing w_0. The trajectory approaches its steady state on the hyperbola where $u_0^* = u_T$ and $w_0^* = 1/2\Lambda\, u_T$. Again, we deduce from (11) that $v_0^* = 0$.

In closing this subsection we express the final state of the system in terms of its dimensional quantities. We first consider the high power case $1/2\Lambda > u_T$ or, in dimensional terms,

$$\frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S} > T_T - T_A.$$

Here, T_T is the threshold temperature, V_S is the volume of the sample, and $\frac{1}{2}CV_0^2$ is the total energy initially stored in the capacitor. The steady state values of the density and temperature for this case are

$$\rho^* = \rho_1 w_0^* = \rho_1 \tag{13a}$$

and

$$T^* = T_A(1 + u_0^*) = T_A + \frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S}, \tag{13b}$$

respectively. The result (13a) is intuitively obvious, but (13b) is not. It states that the increase in the system's temperature above its ambient value depends very little upon the physical properties of the powder sample. In fact, it depends only upon the volume of the sample, the final density ρ_1, and the total energy initially stored in the capacitor.

In dimensional terms, the low power case occurs when

$$\frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S} < T_T - T_A.$$

If the initial density ρ_0 is sufficiently large, then, according to Figure 2b, the density remains constant and the temperature increases to $\frac{1}{2\Lambda w_0(0)}$. In terms of dimensional quantities, the steady state density and temperature are

$$\rho^* = \rho_1 w_0(0) = \rho_0 \tag{13c}$$

and

$$T^* = T_A + \frac{1}{2}\frac{CV_0^2}{\rho_0 C_p V_S}, \tag{13d}$$

respectively. The only difference between this final temperature and the one given for higher power (13b) is the presence of ρ_0 in the denominator. Finally, if the initial density is low enough, then the system evolves along the lower trajectory shown in Figure 2b. The steady state density and temperature are now given by

$$\rho^* = \rho_1 \frac{1}{2\Lambda u_T} = \frac{\rho_1}{2}\left\{\frac{CV_0^2}{\rho_1 C_P V_S}\right\}\frac{1}{T_T - T_A} \tag{13e}$$

and

$$T^* = T_A(1 + u_T) = T_T, \tag{13f}$$

respectively, where it must be recalled that the low power constraint implies $\rho^* < \rho_1$.

4.2. The cooling period. The analysis in the preceding section was concerned with times on the order of θ_S, the source time scale. Convective heat losses, modeled by Newton's law of cooling, are unimportant here. However, they do become important on a much longer time period where the powder sample cools back to the ambient temperature, T_A. We can analyze this behavior by introducing the long time scale $\bar{t} = \epsilon^2\tau$. In dimensional terms, this $\bar{t} = t/\beta\theta_{con}$; that is, we are now considering the temperature evolution on the convective time scale. Inserting this change of variable into (6a) gives

$$\epsilon\frac{\partial}{\partial\bar{t}}(uw) = \frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right) + \frac{1}{\epsilon}S(z,\bar{t}), \quad 0 < z < 1, \quad \bar{t} > 0, \tag{14a}$$

where the source S is defined by $S = f(u)|\frac{\partial\phi}{\partial z}|^2$. The source term can be simplified by integrating (6b) twice and using the boundary conditions (5b) to obtain $\phi = v(\tau)Q(z)/Q(1)$, where now $Q(z) = \int_0^z \frac{1}{f(u)}\,dz'$. Differentiating this result with respect to z and replacing τ by $\bar{t}/\epsilon^2$, we find that the source term becomes

$$S(z,\bar{t}) = \frac{v^2(\bar{t}/\epsilon^2)}{f(u)Q^2(1)}. \tag{14b}$$

Now, on the cooling time scale $\bar{t} = O(1)$, which implies that the argument of v in (14b) is very large; alternatively, $\tau >> 1$. But for large values of τ we have shown that $v \sim v_0 \to 0$ exponentially. Thus, the S/ϵ is negligible and (14a) becomes a homogeneous diffusion equation on this time scale. Also, by inserting $\tau = \bar{t}/\epsilon^2$ into (14) and letting $\epsilon \to 0$ we obtain $\gamma g(u)\{1 - w\} = 0$. This is satisfied in both the high and low power cases. In the former, $w \sim w_0^* = 1$. In the latter, $u \sim u_0^* \leq u_T$ so that $g = 0$.

Taking the facts that $S = 0$ and $w = w_0^*$ into consideration, we find that u satisfies

$$\epsilon w_0^*\frac{\partial}{\partial\bar{t}}(u) = \frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right), \quad 0 < z < 1, \quad \bar{t} > 0. \tag{15a}$$

In addition to this equation u must still satisfy the boundary conditions (6c) at $z = 0$ and $z = 1$. Finally, an initial condition must be prescribed to close the initial boundary value problem for u. Intuitively, this is given by the large τ value of u_0, namely,

$$u|_{\bar{t}=0} = u_0^*. \tag{15b}$$

We note here that the arguments of this paragraph can be put on more formal grounds by asymptotically matching the long time behavior of the solution on the heating time scale with the short time behavior of the solution on the cooling time scale.

We now proceed to find the behavior of $u(z,\bar{t})$ as $\epsilon \to 0$. As usual, we take

$$u \sim U_0(z,\bar{t}) + \epsilon U_1(z,\bar{t}) + \epsilon^2 U_2(z,\bar{t}) + \cdots$$

and insert this expression into (15) and (6c). Equating to zero the coefficients of the powers of ϵ yields a sequence of initial boundary value problems. The leading order problem is

$$\frac{\partial}{\partial z}\left(k\frac{\partial U_0}{\partial z}\right) = 0, \quad 0 < z < 1, \quad k\frac{\partial U_0}{\partial z} = 0, \quad z = 0,1, \quad U_0(z,0) = u_0^*. \tag{16}$$

Integrating this equation once and applying the boundary conditions, we find that $U_0(z,\bar{t}) = U_0(\bar{t})$, a function of $\bar{t}$ alone.

The first order correction U_1 satisfies

$$\frac{\partial}{\partial z}\left(k\frac{\partial U_1}{\partial z}\right) = w_0^*\frac{dU_0}{d\bar{t}}, \quad 0 < z < 1, \quad U_1(z,0) = 0, \tag{17a}$$

and the inhomogeneous boundary conditions

$$k\frac{\partial U_1}{\partial z} = \beta U_0, \quad z = 0, \qquad k\frac{\partial U_1}{\partial z} = -\beta U_0, \quad z = 1. \tag{17b}$$

Integrating the equation in (17a) between $z = 0$ and $z = 1$ and applying the boundary conditions in (17b), we find that U_0 satisfies the ordinary differential equation $\frac{dU_0}{d\bar{t}} = -2\frac{\beta}{w_0^*}U_0$. The solution of this equation satisfying the initial condition in (16) is

$$U_0 = u_0^* e^{-2\frac{\beta}{w_0^*}\bar{t}}. \tag{18}$$

The result given in (18) is the leading order approximation to u on the cooling time scale. For the high power case the result in terms of dimensional quantities becomes

$$T \sim T_A + \frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S}e^{-2t/\theta_{con}}, \tag{19a}$$

which explicitly shows that the temperature decays back to its ambient value on the cooling time scale θ_{con}. This scale depends upon the punches' ability to lose heat to their surroundings. For the low power, high initial density case we have

$$T \sim T_A + \frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S}e^{-2\frac{\rho_1}{\rho_0}t/\theta_{con}}, \tag{19b}$$

and for the low power, low initial density case,

$$T \sim T_A + \frac{1}{2}\frac{CV_0^2}{\rho_1 C_p V_S}e^{-2\frac{\rho^*}{\rho_0}t/\theta_{con}}. \tag{19c}$$

In (19c) the final density ρ^* is given in (13e). It is interesting to note that the low power cases decay more quickly to the ambient temperature than the high power scenario.

5. Slow sintering. Up until now, we have taken the dimensionless parameter γ, appearing in (4d), as an $O(1)$ quantity. That is, we have assumed that sintering occurs on the source time scale, τ. In this section we investigate the case where this process occurs on the much longer cooling time scale, $\bar{t}$. This assumption implies that $\gamma = \gamma_0\epsilon^2$, where γ_0 is now an order one quantity.

The heating regime analysis proceeds as before, the only difference being that (10c) is replaced by

$$\frac{dw_0}{d\tau} = 0, \qquad w_0(0) = \frac{\rho_0(F)}{\rho_1}, \tag{20}$$

from which it follows that $w_0 = \frac{\rho_0(F)}{\rho_1}$; that is, the density remains a constant on the source time scale. Equations (11) and (12) still remain the same. However,

$u_0 \to u_0^* = \frac{1}{2\Lambda w_0}$ as $\tau \to \infty$. Thus, the final temperature in the heating regime depends upon w_0. In terms of dimensional quantities, (13b) is now replaced by

$$T^* = T_A(1 + u_0^*) = T_A + \frac{1}{2}\frac{CV_0^2}{\rho_0(F)C_p V_S}. \tag{21}$$

Since $\rho_1/\rho_0 > 1$, the final temperature is higher in this case, although increasing the external force F on the punches diminishes the difference. Finally, we observe that v_0 still satisfies (11) and hence approaches zero as $\tau \to \infty$.

The cooling period analysis again proceeds as before. Equation (15) is now replaced by

$$\epsilon\frac{\partial(u\,w)}{\partial\bar{t}} = \frac{\partial}{\partial z}\left(k\frac{\partial}{\partial z}u\right), \quad u|_{\bar{t}=0} = \frac{1}{2\Lambda w_0}, \tag{22a}$$

and equation (4d) by

$$\frac{dw}{d\bar{t}} = \gamma_0 g(u)(1-w), \quad w|_{\bar{t}=0} = w_0 = \frac{\rho_0(F)}{\rho_1}, \tag{22b}$$

where we have replaced γ by $\gamma_0\epsilon^2$ and changed the time variable to $\bar{t}$. The temperature u still satisfies the boundary conditions contained in (6c).

We now expand both u and w in the asymptotic series $u \sim U_0 + \epsilon U_1 + \dots$ and $w \sim W_0 + \epsilon W_1 + \dots$, respectively, and perform the same analysis as in section 4.2. We find to leading order that U_0 and W_0 satisfy the ordinary differential equations and initial conditions

$$\frac{d}{d\bar{t}}W_0 = \gamma_0 g(U_0)\,(1 - W_0), \quad W_0(0) = \frac{\rho_0(F)}{\rho_1}, \tag{23a}$$

$$\frac{d}{d\bar{t}}U_0 = -\frac{1}{W_0}\{2\beta U_0 + \gamma U_0 g(U_0)(1 - W_0)\}, \quad U_0(0) = \frac{1}{2\Lambda W_0(0)}. \tag{23b}$$

Equation (23a) is the same as (10c), and (23b) is similar to (12). The difference in the latter is the new initial condition and the change in its right-hand side.

We again have the high and low power cases to consider, $\frac{1}{2\Lambda} \geq u_T$ and $\frac{1}{2\Lambda} \leq u_T$, respectively. The phase plane with typical trajectories is shown in Figure 3a for the former case. These trajectories begin on the hyperbola $W = 1/2\Lambda U$ and flow to the left with U_0 decreasing and W_0 increasing. When this curve crosses $U_0 = u_T$, it becomes horizontal, W_0 remains constant, and U_0 approaches zero. The final state is then $U_0^* = 0$ and $W_0^* = \psi(W_0(0))$, where ψ denotes the function mapping the initial data to its final state. This function is shown in Figure 3b for different values of Λ. Since $1/\Lambda$ is proportional to the power initially stored in the capacitor, i.e., CV_0^2, our results show that higher power levels produce higher densification.

The trajectories for the low power case are shown in Figure 4a, each again beginning on the hyperbola. If the initial density is large enough, the trajectory remains horizontal, U_0 monotonically decreases, and $W_0 = \rho_0/\rho_1$. If the initial density is sufficiently low, then W_0 initially increases and U_0 decreases until the threshold $U_0 = u_T$ is crossed. After this time W_0 remains fixed and U_0 approaches zero as $\bar{t} \to \infty$.

The final density W_0^* for the lower power case is also a function of the initial density $W_0(0)$, i.e., $W_0^* = \psi(W_0(0))$. This function is shown in Figure 4b for several values of Λ. Again, higher power levels for a given initial density produce more densification. However, for densities sufficiently large, the temperature is lower than the threshold and no sintering occurs. This is born out in the linear behavior of ψ as a function of ρ_0/ρ_1.

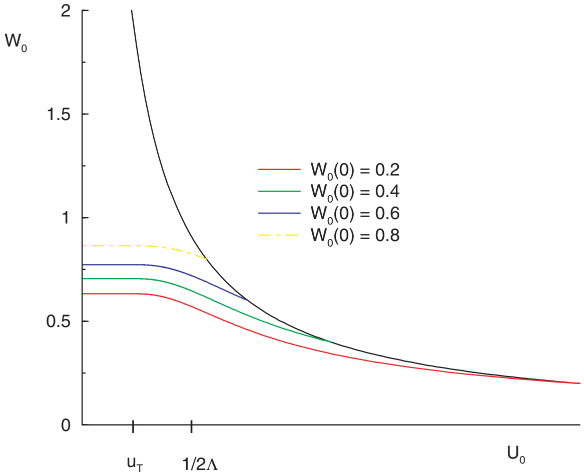


FIG. 3a. *Phase plane: High power.*

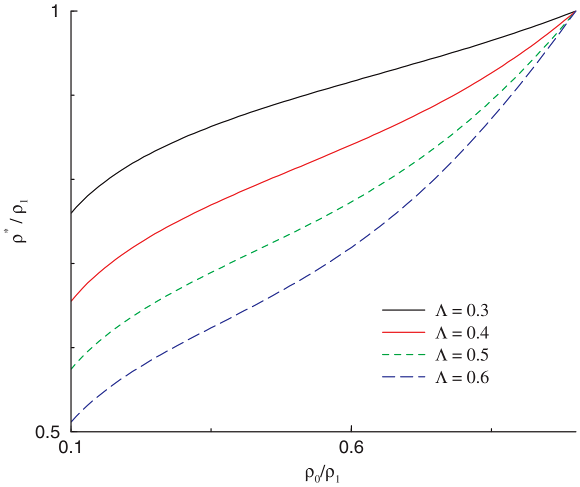


FIG. 3b. *Final density: High power.*

6. Conclusion. We have developed a model to describe a class of experiments that use an electric discharge to sinter or compact ceramic and metallic powders, respectively. Our mathematical description begins with the heat equation, Laplace's equation, and an evolution equation describing the densification of the powder. All these equations are coupled and nonlinear and have appended to them the appropriate initial and boundary conditions.

We have performed an asymptotic analysis to approximate the solutions of these equations. The appropriate small parameter is the ratio of the source to diffusion time scales. Our analysis shows that the leading order temperature, voltage across the sample, and relative density satisfy a third order system of nonlinear, ordinary differential equations. These are analyzed for a particular densification rate which possesses a temperature threshold, below which sintering or densification does not

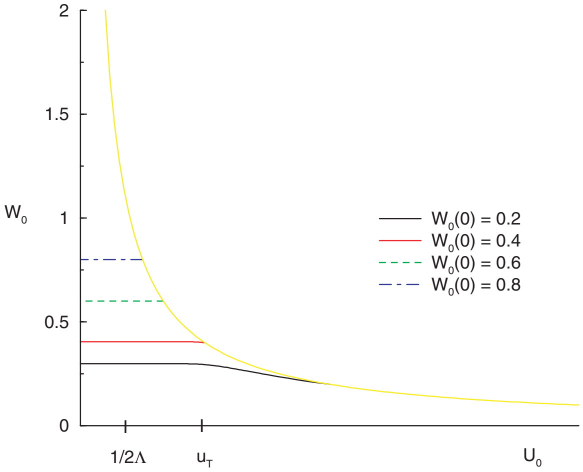


Fig. 4a. *Phase plane: Low power.*

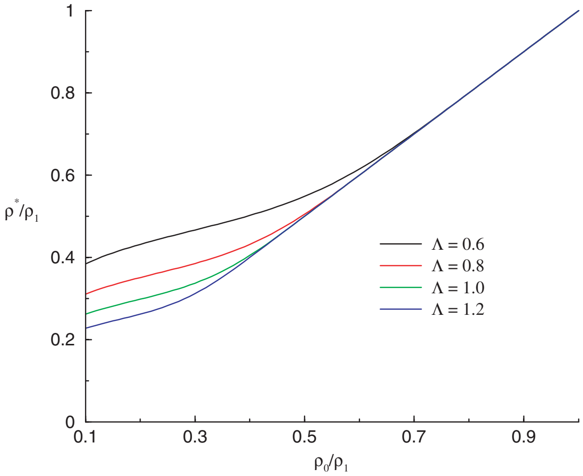


Fig. 4b. *Final density: Low power.*

occur. In particular, we have identified a critical power below which no densification occurs. We have presented a physical interpretation of our results.

The results described above are valid only on the source time scale where diffusion and heat loss to the surrounding environment are unimportant. We have identified a much longer convective time scale on which the heated powder sample loses energy through the punches to the surrounding medium. We have analyzed our equations on this time scale and found that the temperature decays back to its ambient value. The rate depends upon the energy initially stored in the capacitor as does the final density.

All of the preceding results are based on the assumption that the time scale associated with the densification rate is commensurate with the source time scale. We have also modified our analysis to take into account a much slower densification process that occurs on the convective time scale. During the initial heating of the powder, the temperature evolves to a steady state that depends upon the initial density and the energy stored in the capacitor. The density remains at its initial value. Then on the long convective time scale the temperature decays back to its ambient value and the density increases to its final value. This value depends upon the initial density and the energy stored in the capacitor.

We note that our model assumes that the die, which holds the powder sample, is perfectly insulated. If this restriction is removed, then ultimately heat will escape through the cylindrical sides. If the time scale associated with this process is commensurate with the other convective time scales, then a refined theory can be derived. Finally, we observe that heat conduction in the punches will become important when the experimental apparatus becomes larger and the punches more massive. Indeed, if the punches are massive enough, the temperature in them is T_A, the ambient value, and (6c) is replaced by $u = 0$ at $z = 0$ and $z = 1$. This and (6a) strongly suggest that boundary layers occur at the ends of the sample, and through them heat is transferred into the punches. This and the incorporation of more complicated die structures are topics of ongoing research.

REFERENCES

[1] M. Suganuma, Y. Kitawaga, S. Wada, and N. Murayama, *Pulsed electric current sintering of silicon nitride*, J. American Ceramic Society, 86 (2003), pp. 387–396.

[2] P. K. Rajagopalan, S. V. Desai, R. S. Kalghatgi, T. S. Krishnan, and D. K. Bose, *Studies on the electric compaction of metal powders*, Material Science Engineering A, 280 (2000), pp. 289–298.

[3] K. Okazaki, *Electro-discharge consolidation applied to nanocrystalline powders*, Material Science Engineering A, 287 (2000), pp. 189–198.

[4] A. C. Fowler, I. Frigaard, and S. D. Howison, *Temperature surges in current-limiting circuit devices*, SIAM J. Appl. Math., 52 (1992), pp. 998–1011.

[5] W. D. Kingery, H. K. Bowen, and D. R. Uhlmann, *Introduction to Ceramics*, John Wiley and Sons, New York, 1976.

SIAM J. Appl. Math.
Vol. 68, No. 6, pp. 1518–1534

NUMERICAL TESTS OF A PHASE FIELD MODEL WITH SECOND ORDER ACCURACY[*]

GUNDUZ CAGINALP[†], XINFU CHEN[†], and CHRISTOF ECK[‡]

Abstract. Numerical computations are performed for a recently derived phase field model for the interface between two phases. The rigorous results indicate that solutions to this new phase field model should converge more rapidly than traditional ones to solutions of the corresponding sharp interface (free boundary) formulation for sufficiently small values of the approximation parameter ε representing the thickness of the interfacial region. In particular, the distance between the sharp interface of the limiting model and the zero level set of the phase function in the phase field model is of order ε^2 rather than ε. Numerical computations within a three-dimensional spherically symmetric setting compare the computed solutions of this new model with the known exact solutions for the limiting free boundary problem and confirm the second order accuracy predictions of the theory for sufficiently small ε. The sets of parameters include those of succinonitrile used in dendritic experiments.

Key words. phase field model, matched asymptotic expansion, second order accuracy

AMS subject classifications. 80A22, 80M35

DOI. 10.1137/070680965

1. Introduction.

Phase field models are now established as one of the most popular approaches for the computation of various types of dynamical phase transition models and problems with moving interfaces [4, 12, 13, 19, 20, 21, 22, 23, 35].

From the perspective of numerical simulation, these models can be interpreted as approximations of free boundary problems by problems without explicit interface conditions. This simplifies the numerical implementation of the model and, in particular, renders possible the application of standard software packages for partial differential equations (PDEs) to free boundary problems without implementing special front tracking and difficult treatment of topological change techniques. The required resolution for the diffuse interface that arises in the phase field models can be achieved by adaptive mesh refinement, a feature that is typically available in modern software packages.

Although the phase field (diffuse interface) approach can be used within a number of physical applications, many of the key ideas can be understood in terms of the two-phase problem with surface tension and kinetic undercooling. Starting with the free boundary approach for this physical problem, we consider a material in a spatial region $\Omega \subset \mathbb{R}^n$ $(n \geqslant 1)$ that can be in either of two phases (e.g., liquid or solid) separated by an interface, $\Gamma(t)$. Hence the mathematical problem consists of determining both the temperature $T(x,t)$ and the interface $\Gamma(t)$ from the system, in its full dimensional form,

[*] Received by the editors January 25, 2007; accepted for publication (in revised form) January 28, 2008; published electronically May 28, 2008.
http://www.siam.org/journals/siap/68-6/68096.html

[†] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (caginalp@pitt.edu, http://www.pitt.edu/~caginalp; xinfu@pitt.edu, http://www.pitt.edu/~xfc). The second author was supported by National Science Foundation grant DMS–0504691.

[‡] Faculty of Mathematics, University of Bielefeld, Universitätsstr. 25, 33615 Bielefeld, Germany (eck@math.uni-bielefeld.de, http://www.math.uni-bielefeld.de/~eck).

$$
(1.1) \qquad \begin{cases} (\rho\, c\, T)_t = \operatorname{div}(K\, \nabla T) & \text{in} \quad \Omega \backslash \Gamma(t), \\ \rho\, \ell\, v_n = [\![K \nabla T \cdot \mathbf{n}]\!]_+^- & \text{on} \quad \Gamma(t), \\ T = T_E - \frac{\sigma}{[s]_{\mathrm{E}}} \left\{ \kappa + \alpha\, v_n \right\} & \text{on} \quad \Gamma(t), \end{cases}
$$

with ℓ and c as the latent heat and heat capacity per unit mass, K the diffusivity, ρ the density, T_E the equilibrium freezing temperature, $[s]_{\mathrm{E}}$ (energy/(volume $\cdot$ degree)) the entropy difference per volume, σ (energy/area) the surface tension, and α the strength of kinetic undercooling. The unit normal, sum of principal curvatures, and velocity of the interface are given by $\mathbf{n}$, κ, and v_n, while $[\![\cdots]\!]_-^+$ denotes the difference in the limits from the two sides of the interface.

The history of this problem dates back to 1831 when Lamé and Clapeyron [25] studied the freezing of the ground using (1.1) with $T = T_E$ replacing the third equation in (1.1). Reformulated in 1889, it became known as the classical Stefan problem [34]. It has the appealing mathematical feature that the temperature, $T(x,t)$, determines the phase at each point (x,t). By definition, $T(x,t) > T_E$ implies that the material is liquid at that point (or, more generally, in the phase with the higher internal energy), while $T(x,t) < T_E$ means that it is solid, and $T(x,t) = T_E$ defines the interface $\Gamma(t)$. Thus, the condition that $T(x,t) = T_E$ at the interface appears to be mathematically convenient. Nevertheless, the mathematical study of classical solutions to the Stefan model posed difficult challenges. Modern analysis (e.g., [26, 31]) converts (1.1) (with $\sigma = 0$) to the single equation $[e(u)]_t = D\Delta u$, where u is a scaled temperature, $e(u)$ is proportional to internal energy, and $D = K/(\rho c)$ is the heat diffusion coefficient. Since phase is assumed to be determined by temperature, one can write $e(u) = u + H(u)$ with H the Heaviside function.

Materials science research (e.g., [17]) in subsequent decades (after Lamé and Clapeyron [25]) showed that the interface temperature need not be at the equilibrium melting temperature, T_E, so that the material can be liquid well below the melting temperature, for example. In terms of mathematical modeling, there is a profound difference between the classical Stefan model ($T = T_E$ on $\Gamma(t)$) and the modern set of equations (1.1), since the temperature in the latter model can no longer retain its dual role of determining both the temperature and the phase. This means that using (1.1) directly necessitates tracking the interface in time, which is difficult but mathematically possible; see [15] for the well-posedness of the problem. Even if this is done, however, equations (1.1) are valid only so long as the interface does not self-intersect.

An alternative approach, known as the phase field or diffuse interface model, is to formulate the problem in terms of two variables, temperature and phase field (see [10] for more discussion). The mathematical problem is then to solve the following parabolic system for $(T_\varepsilon, \phi_\varepsilon)$ in its full dimensional form:

$$
(1.2) \qquad \begin{cases} (\alpha_\varepsilon\, \phi_\varepsilon)_t = \Delta \phi_\varepsilon - \varepsilon^{-2} W'(\phi_\varepsilon) + \varepsilon^{-1} [s]_E \sigma^{-1} G'(\phi_\varepsilon) [T_\varepsilon - T_E] \;, \\ (\rho c T_\varepsilon + \frac{1}{2} \rho \ell \phi_\varepsilon)_t = \operatorname{div}(K \nabla T_\varepsilon), \end{cases}
$$

where the unknowns $T_\varepsilon(x,t)$ and $\phi_\varepsilon(x,t)$ are, respectively, the temperature and the phase indicator ($\phi_\varepsilon = 1$ for liquid and -1 for solid) and ε is a small positive parameter representing the thickness of the interfacial region. Here W is a potential with double well of equal depth at ± 1 and G is a function relating microscopically how energy is relayed in the thin interfacial region. As discussed in full in [10], in order for the phase field model to approximate accurately the free boundary model (1.1), it is better to require G and α_ε to satisfy certain compatibility conditions; in particular, the following choice is sufficient:

$$W(s) = \frac{1}{2}(1-s^2)^2, \quad G(s) = s - \frac{1}{3}s^3, \quad \alpha_\varepsilon = \alpha + \frac{5}{12}\frac{\varepsilon\rho\ell[s]_{\mathrm{E}}}{(K\sigma)}. \tag{1.3}$$

The interface in this formulation is now defined as the level set

$$\Gamma_\varepsilon(t) := \{x \in \Omega(t) \mid \phi_\varepsilon(x,t) = 0\}\,; \tag{1.4}$$

thus there is no need to track it explicitly, and the practical problem is simply the computation of a smooth system of parabolic differential equations. A number of works (e.g., Caginalp and Chen [8, 9] and Soner [33]) have proved that solutions of the phase field equations converge to those of the corresponding free boundary problems as $\varepsilon \to 0$. The parameter ε represents the thickness of the interfacial region, whose true value is on an atomic scale. Computing with this true physical value would make many realistic computations unfeasible. However, it has been shown that the value of ε can be used essentially as a free parameter that can be increased by orders of magnitude without significantly altering the behavior of the interface [12]. Although the phase field approach provides a methodology for understanding the physical interface problems directly and has been used to derive the sharp interface models, one can also view it as a computational approach designed to approximate the limiting sharp interface (free boundary) problem. This is the perspective we adopt in this paper.

The use of phase field computations in realistic physical situations has led to a growing interest in developing and testing different phase field equations that better approximate the limiting free boundary problem. Let $\Gamma(t)$ and $\Gamma_\varepsilon(t)$ denote the interface of the free boundary problem (1.1) and the zero level set of the phase function $\phi_\varepsilon(x,t)$ of the phase field model (1.2), respectively. We are interested in approximating the free boundary problem with the phase field model by the following criteria: there exist positive constants C and ε_0 such that

$$\text{distance } (\Gamma(t), \Gamma_\varepsilon(t)) \le C\varepsilon^k \qquad \forall\, \varepsilon \in (0, \varepsilon_0]. \tag{1.5}$$

Established theoretical results (e.g., [8, 9]) and computations (e.g., [13, 19]) indicate that these estimates are valid for $k = 1$. Recently, in [10] we derived a phase field model that ensures a second order accuracy (namely, $k = 2$ in the bound above) for the approximation of the free boundary. Unlike the automatic [29] second order approximation of motion by mean curvature by the Allen–Cahn equation [2], the second order accuracy here is obtained by special choices of G and α_ε. In particular, by utilizing the choice (1.3), all first order terms automatically cancel out, thus leading to a second order model. Here the coefficient $\frac{5}{12}$ for the first order correction of α_ε is calculated from the special choices of W and G. The derivation and proof use a method that differs from the standard technique of matched asymptotic expansions [1, 6, 9, 11, 16]. In our recent work, the inner expansion is computed with respect to the interface $\Gamma(t)$ of the limit interface and not with respect to the level set $\Gamma_\varepsilon(t)$ of the phase field as in more traditional approaches. A key advantage of this new technique is that it permits tracking of the position of the perturbed interface by a distance function h_ε to the limit interface; see section 2.

There have been other studies attempting to derive phase field models that converge more rapidly to their sharp interface limits by an alternative procedure of finding conditions that eliminate undesired terms of first order, as done, e.g., in [3]. It is not always obvious, however, which terms should be cancelled in order to obtain robust approximation properties that are an improvement over the original models.

The rigorous theory does not establish the constant C in the estimates (1.5) or the value of the upper bound ε_0 for which the bounds are valid. Consequently, we perform numerical computations on this recently derived phase field equation to determine whether the second order accuracy described by these bounds is valid for typical parameter ranges and computational constraints. In particular, one of the tests utilizes the physical measurables for succinonitrile that is used in many of the dendritic experiments [18, 24].

2. The phase field model.

In this section, we state the phase field model (1.2) introduced in [10] in a form that is convenient for computation.

2.1. Nondimensionalization.

Using the fully physical dimensional form of equations has its advantage and convenience for practical considerations. Mathematically, however, it is awkward and numerically complicated in realizing the stiffness of the problem. From the viewpoint of scaling invariance, it is desirable to make a change of variables to transfer the fully dimensional version of the free boundary problem (1.1) and the phase field model (1.2) into their nondimensional counterparts.

To convert (1.1), introduce L, the diameter of the sample, and use the standard transformation

$$u := \frac{T - T_E}{\ell/c}, \qquad D := \frac{K}{\rho c}, \qquad d_0 := \frac{\sigma\ c}{[s]_E\ \ell}, \quad a := \alpha D,$$
$$d := \frac{d_0}{L}, \qquad \frac{x}{L} =: \tilde{x} \longrightarrow x, \qquad \frac{D}{L^2} t =: \tilde{t} \longrightarrow t.$$

The free boundary problem (1.1) then has the following dimensionless form:

$$\begin{cases} u_t^{\pm} = \Delta u^{\pm} & \text{in } \ \Omega^{\pm}(t), \\ u^{\pm} = -d\,(\kappa + a\mathrm{v}) & \text{on } \ \Gamma(t), \\ \mathrm{v} = [\![\nabla u \cdot \mathbf{n}]\!]_{+}^{-} & \text{on } \ \Gamma(t), \end{cases} \tag{2.1}$$

where $\Omega^+(t) \cup \Omega^-(t) \cup \Gamma(t) = \Omega$, $\mathbf{n}$ is the unit vector normal to $\Gamma(t)$ pointing toward $\Omega^+(t)$, κ is the sum of the principal curvatures of $\Gamma(t)$ (positive for convex solid), and v is the normal velocity of $\Gamma(t)$ (positive for solidification).

Note that the size of the sample (i.e., Ω) in the new, $\bar{x}$, units is 1 in (2.1). There are only two physical dimensionless parameters: a and d.

1. The constant a represents the strength of kinetic undercooling; it is a measurable *dimensionless material constant.*
2. While d_0 is a *material constant* that relates the surface tension or size of the nucleation radius, the dimensionless constant

$$d := \frac{d_0}{L}$$

depends on the particular experiment. For example, for a typical $d_0 = 10^{-7}$ cm, if the "sample size" or "macroscopic resolution size" is $L = 10^{-3}$ cm, then $d = 10^{-4}$.

Although d is small and the difference between the Gibbs–Thomson condition $u = -d\,(\kappa + a\mathrm{v})$ and the Stefan condition $u = 0$ on the free boundary may appear insignificant, the respective interface motion is known to be significantly different for the two conditions [28, 30].

3. We use a unit for time that matches the units of space. For example, if $D = 10^{-3}$ cm^2/s and $L = 10^{-3}$ cm, then $t = 1000$ represents a physical $L^2t/D = 1$ second. In other words, $t = 1$ represents one millisecond.
4. Here $u = 1$ represents the temperature that a liquid at melting temperature attains after absorbing an amount of energy equal to the latent heat. For water, $u = 1$ represents $T = 80°$ C $= 353$ Kelvin. In this problem there is another parameter, u_∞, the dimensionless temperature at the far field, that plays a role. When u_∞ is small, quite often it is convenient to use $u/|u_\infty|$ as the dimensionless temperature. For succinonitrile, $\ell/c = 23.13$ Kelvin, so $T_\infty = T_E - 0.2313$ Kelvin is equivalent to $u_\infty = -0.01$.

In addition to the dimensionless quantities above, it is useful to scale the extra parameter, ε, representing the thickness of interfacial region (5–100 atomic distances). Introducing dimensionless constants

$$\bar{\epsilon} := \frac{\varepsilon}{d_0}, \qquad \epsilon := \frac{\varepsilon}{L} = \frac{\varepsilon}{d_0}\,\frac{d_0}{L} = \bar{\epsilon}\, d,$$

the phase field model (1.2) has the following dimensionless form:

$$\text{(2.2)} \qquad \begin{cases} u_t + \frac{1}{2}\phi_t \;=\; \Delta u, \\ (a + \frac{5}{12}\,\bar{\epsilon})\phi_t \;=\; \Delta\phi + \epsilon^{-2}(2\phi + \bar{\epsilon}u)(1-\phi^2). \end{cases}$$

The stiffness of the phase field model comes from the largeness of the quantity ϵ^{-2} on the right-hand side of the second equation. Here the important correction term $\frac{5}{12}\bar{\epsilon}$ is an addition to the traditional phase field model. It eliminates the first order terms in the asymptotic expansion. As mentioned in the introduction, the applicability of the phase field model in numerical computations is due to the fact that one does not need to use the actual (atomic) size of ϵ. One can use ϵ that is much larger—though still small—without altering the solution significantly [12, 13].

2.2. Initial data. To obtain second order approximation, the initial value to the phase field system (2.2) has to be second order consistent with the free boundary problem (2.1). This leads to the following choice of initial data for (2.2) (see [10] for details):

$$\text{(2.3)} \quad \begin{cases} \phi(\cdot,0) \;= \tanh\frac{h}{\epsilon}, \\ u(\cdot,0) \;= \dfrac{u_0^+}{1+e^{-2h/\epsilon}} + \dfrac{u_0^-}{1+e^{2h/\epsilon}} + \dfrac{\epsilon}{2}\nabla h\cdot\nabla(u_0^+ - u_0^-)\displaystyle\int_{-\infty}^{h/\epsilon} z\, d\tanh z. \end{cases}$$

Here $h = h(x)$ is the signed distance from x to the initial interface $\Gamma(0)$, and $u_0^\pm = u_0^\pm(x)$ are smooth extensions of the initial temperature for the free boundary problem.

3. Analytic feature of the numerical example. The main purpose of this paper is to check numerically the validity of the assertion that the new phase field model (2.2) approximates the free boundary model (2.1) with second order accuracy, using physical parameters in one case. In particular, the computations can address the issue of the constants C and ε_0 in (1.5), thereby determining whether there is a computational advantage to the new phase field model in practical circumstances. The test example in our earlier paper [10] is one dimensional, so the curvature effect is not present. We would like to find a test case that has the following features:

1. Explicit solutions for the free boundary problem are available.
2. There are curvature effects.
3. There are kinetic undercooling effects.
4. The ratio between the curvature effect and kinetic undercooling effect can be adjusted.

There is an example in a three dimensional radially symmetric situation that models the solidification (growing) process of a solid ball in undercooled liquid [32]. The solution has the properties that (i) the free boundary is located at $|x| = R(t) = 2\gamma\sqrt{t}$, and (ii) the temperature is a combination of three self-similar solutions to the heat equation $u_t = \Delta u$:

$$u(x,t) = u_0 := 1,$$

$$u(x,t) = u_1(|x|,t), \qquad u_1(r,t) := \frac{\operatorname{erf}(r/\sqrt{4t})}{r}, \qquad \operatorname{erf}(z) := \frac{2}{\sqrt{\pi}}\int_0^z e^{-y^2}\,dy,$$

$$u(x,t) = u_2\Big(\frac{|x|}{\sqrt{4t}}\Big), \qquad u_2(z) := \int_z^\infty \frac{e^{-y^2}}{y^2}dy.$$

The following calculations verify that for each $\gamma > 0$, there is exactly one such solution to (2.1).

1. When the free boundary is given by $|x| = R(t) := 2\gamma\sqrt{t}$, one has

$$\kappa = \frac{2}{R(t)}, \qquad \mathrm{v} = \frac{dR(t)}{dt} = \frac{\gamma}{\sqrt{t}} = \frac{2\gamma^2}{R(t)}, \qquad \frac{a\mathrm{v}}{\kappa} = a\gamma^2.$$

The Gibbs–Thomson condition requires the temperature at the free boundary to be

$$u\Big|_{\Gamma(t)} = -d(\kappa + a\mathrm{v}) = -\frac{A}{R(t)}, \qquad A := 2d(1 + a\gamma^2).$$

2. In the ball $\{x \mid |x| < R(t)\}$, the material is in the solid phase. The only self-similar solution to the heat equation $u_t = \Delta u$ with boundary value $u(x,t) = -A/r$ at $r = |x| = R(t)$ and vanishing derivative at $r = 0$ is given by

$$u^-(x,t) = -\frac{Au_1(|x|,t)}{\operatorname{erf}(\gamma)}. \tag{3.1}$$

3. Outside the ball $\{x \mid |x| \le R(t)\}$, the material is in the liquid phase. There is a family, with parameter B, of solutions having boundary value $u(x,t) = -A/r$ at $r = |x| = R(t)$:

$$u^+(x,t) = -\frac{Au_1(|x|,t)}{\operatorname{erf}(\gamma)} + B\Big(u_2(\gamma) - u_2(|x|/\sqrt{4t})\Big). \tag{3.2}$$

The solution we need corresponds to that satisfying $\mathrm{v} = [\![u_r]\!]_+^-$. Thus, we have

$$\frac{2\gamma^2}{R(t)} = \frac{Bu_2'(\gamma)}{\sqrt{4t}}, \quad \text{i.e.,} \quad B = \frac{2\gamma}{u_2'(\gamma)} = -2\gamma^3 e^{\gamma^2}.$$

In conclusion, for each $\gamma > 0$ we have a solution to (2.1), given by

$$\text{(3.3)} \qquad \begin{cases} \Gamma(t) &= \{x \mid |x| = R(t) := 2\gamma\sqrt{t}\}, \\ u(x,t) &= -\dfrac{2d\,(1+a\gamma^2)\,\mathrm{erf}(|x|/\sqrt{4t})}{\mathrm{erf}(\gamma)\ |x|} - \displaystyle\int_\gamma^{\max\{\gamma,|x|/\sqrt{4t}\}} \frac{2\gamma^3 e^{\gamma^2-y^2}}{y^2}dy. \end{cases}$$

For such a solution, the ratio of the strength of the kinetic undercooling to the strength of the surface tension is $a\mathrm{v}/\kappa = a\gamma^2$. Also, there is an important physical quantity,

$$u_\infty := u|_{|x|=\infty} = -2\gamma^3 \int_\gamma^\infty \frac{e^{\gamma^2-y^2}}{y^2}dy.$$

Given $u_\infty < 0$, one can show that there is a unique positive γ that satisfies the above relation. Thus, the measure of the degree of undercooling u_∞ is equivalent to the measure of γ.

4. Numerical simulation. For a solution (3.3) of the free boundary problem (2.1), we solve numerically the corresponding radially symmetric solution to the phase field model (2.2) in the unit ball:

$$\Omega := \{x \in \mathbb{R}^3 \mid |x| < 1\}.$$

The system (2.2) is first discretized with respect to time by a second order scheme. Fix a time mesh size $\delta t > 0$. For every integer $k \geqslant 0$, denote by $(u_k(\cdot), \phi_k(\cdot))$ the approximation of the solution $(u(\cdot, k\delta t), \phi(\cdot, k\delta t))$ at time $t = k\delta t$. The semi-discretization in time has the form

$$\begin{aligned} \frac{u_{k+1}-u_k}{\delta t} + \frac{\phi_{k+1}-\phi_k}{2\,\delta t} &= \Delta\frac{u_{k+1}+u_k}{2}, \\ \epsilon^2 a_\epsilon \frac{\phi_{k+1}-\phi_k}{\delta t} - \epsilon^2\Delta\frac{\phi_{k+1}+\phi_k}{2} &= -W'(\phi_k) - \tfrac{1}{2}W''(\phi_k)[\phi_{k+1}-\phi_k] \\ &\quad + \frac{\bar\epsilon\,(u_k+u_{k+1})}{2}\Big\{G'(\phi_k) + \tfrac{1}{2}G''(\phi_k)[\phi_{k+1}-\phi_k]\Big\}, \end{aligned}$$

where

$$\bar\epsilon := \frac{\varepsilon}{d_0}, \quad \epsilon := \frac{\varepsilon}{L}, \quad a_\epsilon := a + \frac{5}{12}\,\bar\epsilon, \quad W(s) = \frac{1}{2}(1-s^2)^2, \quad G(s) = s - \frac{1}{3}s^3.$$

In the radial ($r = |x|$) coordinates, the Laplacian Δ is further discretized by linear finite elements on a uniform mesh of size $\delta r = 1/n$, where n is the total number of spatial mesh points. This scheme leads to a nonlinear system for each time step.

The boundary condition for temperature of the phase field model is taken as the known exact solution to the free boundary problem, whereas the boundary value for ϕ is taken as $\phi|_{|x|=1} = 1$. The solution is calculated for a time interval $[t_0, t_1]$ according to the timing of the solution (3.3) of the free boundary problem. Here $t_0 > 0$ is the initial time, and the terminal time t_1 satisfies $2\gamma\sqrt{t_1} < 1$. The initial condition (at time $t = t_0$) is taken as (2.3), where $h = h(x) = |x| - R(t_0)$, and u_0^- and u_0^+ are as in (3.1) and (3.2).

In what follows, Model 1 refers to the original phase field model where the correction term $\frac{5}{12}\,\bar\epsilon$ in the kinetic coefficient, a_ϵ, is not present; i.e., $a_\epsilon = a$. Model 2 is that with the correction added: $a_\epsilon = a + \frac{5}{12}\,\bar\epsilon$.

Remark. (1) In our actual implementation, the quantity $\frac{\bar{\epsilon}\,(u_k+u_{k+1})}{2}$ on the right-hand side of the second equation is replaced by $\bar{\epsilon}u_k$. The advantage of such a change is that the nonlinear system for (ϕ_{k+1}, u_{k+1}) is decoupled into two linear systems, one for ϕ_{k+1} and the other for u_{k+1}. Though theoretically the resulting discretization becomes first order in δt, the discretization is still stable, and in the special case when $\bar{\epsilon}$ is small, this change from $\frac{\bar{\epsilon}\,(u_k+u_{k+1})}{2}$ to $\bar{\epsilon}u_k$ can be regarded as second order.

(2) The initial condition (2.3) is derived only from model 2; namely, it may not apply to Model 1, where $a_\epsilon = a$. Thus, in numerical implementation for Model 1, the last term in (2.3) is not added in the initial value for u. This is a routine practice in traditional numerical simulations for the phase field models. Indeed, starting from any crude initial data, the phase field dynamics automatically produces a needed fine profile after a small initiation time.

Computation 1. We begin by testing the accuracy of the numerical scheme. That is, for fixed ϵ we find the rate of convergence of the numerical scheme with respect to the spatial mesh size $\delta r = 1/n$ and the time mesh size δt. This helps us to determine how fine a mesh is needed in order to compare the difference between the exact solution to (2.1) and the exact solution to (2.2).

As an illustration, we take the following values of the dimensionless quantities:

$$a = 20, \qquad d := \frac{d_0}{L} = 0.001, \quad \bar{\epsilon} = \frac{\varepsilon}{d_0} = 5, \qquad \epsilon = \frac{\varepsilon}{L} = 0.005, \qquad u_\infty = -0.0046\ .$$

The corresponding value of γ and the ratio of kinetic undercooling to curvature effect are, respectively,

$$\gamma = \frac{1}{20}, \qquad \frac{a\mathrm{v}}{\kappa} = a\gamma^2 = \frac{1}{20}.$$

We calculate the solution from time $t_0 = 1.0$ with initial radius of solid $R(t_0) = 0.1$ to time $t_1 = 9.0$ with final radius $R(t_1) = 0.3$. The numerical result is summarized in Table 1. For easy reference, errors to the exact solution of the phase field model (PFM) and differences to the solution of the free boundary problem (FBP) are calculated in their relative sizes. In calculating the relative error of the numerical scheme here, the exact solution is postulated to be the numerical solution with the finest mesh.

With this assumption we examine the previous level of refinement, namely $n = 3200$, $\delta t = 2.5 \times 10^{-5}$, and observe that the errors relative to the PFM are much smaller than those relative to the FBP. This indicates that the mesh refinement is more than adequate to test how accurately each of the two (phase field) models approximates the FBP. In particular (for $n = 3200$, $\delta t = 2.5 \times 10^{-5}$) the relative error of computation for Model 1 is 10^{-5}, while the difference between the computed and the exact values of the free boundary is 40 times larger at 4×10^{-4}. For Model 2 there is a factor of 17. Examining the prior two levels of refinement ($n = 1600$ and $n = 800$) for Model 1, we see that the relative error (computation compared with the PFM) diminishes from 2×10^{-4} ($n = 800$) to 4.7×10^{-5} ($n = 1600$) to 10^{-5} ($n = 3200$), i.e., factors of about 4, while the relative difference between computation and the FBP varies only from 2.2×10^{-4} ($n = 800$) to 3.7×10^{-4} ($n = 1600$) to 4.0×10^{-4} ($n = 3200$). Hence, the difference between the computed Model 1 and exact FBP stabilizes near 4.0×10^{-4}. The situation is similar for Model 2; i.e., the numerical error in computing the PFM is negligible compared to the difference between the Model (either 1 or 2) and the exact FBP. Note also that halving the time step has a very small effect on these errors.

Table 1

Computation 1. *The rate of convergence of the numerical scheme, shown in relative error to the partial differential equation (rel. error to PFM). Spatial and time mesh sizes are $\delta r = 1/n$ and δt. Also shown is the relative difference with respect to the solution to the free boundary problem (rel. diff. to FBP). Here $\epsilon = 0.005$.*

n	δt	Model 1			Model 2		
$(1/\delta r)$		Position	Rel. error to PFM	Rel. diff. to FBP	Position	Rel. error to PFM	Rel. diff. to FBP
200	4.0 e-4	0.284124	5.3 e-2	5.3 e-2	0.283753	5.4 e-2	5.4 e-2
200	2.0 e-4	0.284529	5.2 e-2	5.2 e-2	0.283945	5.3 e-2	5.4 e-2
200	1.0 e-4	0.284865	5.1 e-2	5.0 e-2	0.294068	5.3 e-2	5.3 e-3
400	4.0 e-4	0.299819	1.0 e-3	6.0 e-4	0.299643	1.0 e-3	1.2 e-3
400	2.0 e-4	0.299853	0.9 e-3	4.9 e-4	0.299676	0.9 e-3	1.1 e-3
400	1.0 e-4	0.299870	0.8 e-3	4.3 e-4	0.299693	0.9 e-3	1.0 e-3
800	2.0 4-4	0.300059	2.1 e-4	2.0 e-4	0.299884	2.3 e-4	3.9 e-4
800	1.0 e-4	0.300063	2.0 e-4	2.1 e-4	0.299889	2.1 e-4	3.7 e-4
800	5.0 e-5	0.300065	1.9 e-4	2.2 e-4	0.299891	2.0 e-4	3.6 e-4
1600	1.0 e-4	0.300110	4.7 e-5	3.7 e-4	0299937	5.0 e-5	2.1 e-4
1600	5.0 e-5	0.300110	4.7 e-5	3.7 e-4	0.299938	4.7 e-5	2.1 e-4
1600	2.5 e-5	0.300110	4.7 e-5	3.7 e-4	0.299938	4.7 e-5	2.1 e-4
3200	1.0 e-4	0.300121	1.0 e-5	4.0 e-4	0.299949	1.0 e-5	1.7 e-4
3200	5.0 e-5	0.300121	1.0 e-5	4.0 e-4	0.299949	1.0 e-5	1.7 e-4
3200	2.5 e-5	0.300121	1.0 e-5	4.0 e-4	0.299949	1.0 e-5	1.7 e-4
6400	1.0 e-4	0.300124		4.1 e-4	0.299952		1.6 e-4
6400	5.0 e-5	0.300124		4.1 e-4	0.299952		1.6 e-4
6400	2.5 e-5	0.300124		4.1 e-4	0.299952		1.6 e-4

Table 2

Computation of interface position at terminal time from the phase field models, in comparison with 0.300000 *from the free boundary model.*

ϵ	Model 1		Model 2	
	$R_\epsilon(t_1)$	$\|1 - R_\epsilon/R\|$	$R_\epsilon(t_1)$	$\|1 - R_\epsilon/R\|$
0.0400	0.29814	6.2 e-3	0.29685	1.0 e-2
0.0200	0.29991	3.1 e-4	0.29924	2.5 e-3
0.0100	0.30015	5.0 e-4	0.29981	6.4 e-4
0.0050	0.30012	4.0 e-4	0.29995	1.7 e-4
0.0025	0.30007	2.4 e-4	0.29999	4.4 e-5
0	0.30000	0	0.30000	0

Thus the computations summarized in Table 1 provide a guide to the error in the numerical computations of the PDEs in terms of the mesh sizes for n and δt. With these numerical errors under control, we can pursue our central goal of distinguishing the differences between the models and the free boundary problems. In what follows we will vary ϵ and examine the behavior of these differences as a function of ϵ. In particular we would like to determine if the difference between Model 2 and the free boundary problem is indeed proportional to ϵ^2, particularly when we use material parameters that are drawn from experiments of dendrites (see Computation 4 below).

In the following examples, the numerical effects are controlled so that the difference shown can be regarded as that between the solutions to the phase field model (2.2) and to the free boundary model (2.1).

Computation 2. Using a sufficiently fine mesh that eliminates significant numerical error (as discussed above), we perform a set of calculations with the material parameters above. These computations involve a spectrum of values of ϵ, as shown in Table 2, and will be compared with the hypothesized relation

$$\left|1 - \frac{R_\epsilon}{R}\right| \sim C\epsilon^2 \tag{4.1}$$

that has been proved [10] using Model 2 for sufficiently small ϵ. In this first set of computations we explore the large ϵ part of this spectrum. For each model we compute $\log|1 - R_\epsilon/R|$ and plot it against $\log\epsilon$, so that a slope of 2 indicates agreement with (4.1), while a slope of 1 suggests an $O(\epsilon)$ error that is the expected result for Model 1. The results for Model 2 are

$$\log|1 - R_\epsilon/R| = 0.554 + 1.8556\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	0.554	0.112
Log(ϵ)	1.8556	0.054

with R-Sq = 99.7% and the F-value for the analysis of variance at 1181. The R-Sq value obtained is a statistical measure (not to be confused with the position R that we have above) that indicates that essentially all of the variation in the data points is explained by the linear model above. The F-value is a measure of the squares of differences between the linear model and the mean relative to the linear model and the data points. For four, five or six data points the F-value needed for 95% statistical confidence is 12, 10, and 9, respectively. A complete discussion of these measures can be found in a basic statistical text such as [27]. The coefficient of $\log(\epsilon)$ is $0.8556/0.054 = 15.84$ or almost 16 standard deviations away from the coefficient value of 1. This yields a p-value that is essentially zero; i.e., there is essentially zero probability that the slope differs from 1 due to randomness. In other words, the null hypothesis that the relative difference between Model 2 and the free boundary problem corresponds to an exponent of 1 must be rejected overwhelmingly (16 standard deviations). The values for Model 2 are plotted using the large dots.

A similar analysis for Model 1 (plotted with small dots) shows that the relative error displays a less regular pattern, yielding a coefficient of 0.832 (i.e., slightly less than a linear relationship), but with a p-value of only 0.154 and an F-value of only 3.61:

$$\log|1 - R_\epsilon/R| = -1.5276 + 0.832\log(\epsilon).$$

In practical terms, there is a significant improvement from Model 1 to Model 2 that is evident particularly for smaller values of ϵ. For the smallest value tested in these computations, namely, $\epsilon = 0.0025$, one has a ratio of $240/44 = 5.4545$ in the relative differences (between the two models) to the exact free boundary problem. The consistency of the results for Model 2 and the coefficient computed above suggest that the difference between the models grows as ϵ is made smaller.

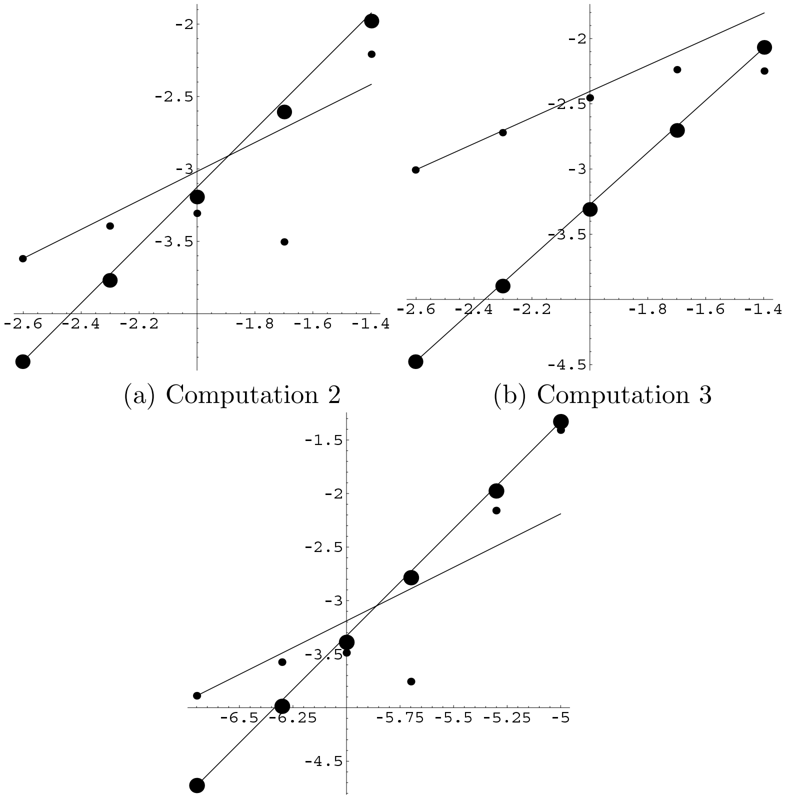
Computation 3. We solve numerically the phase field model (2.2) with the following parameter values:

$$a = 20, \qquad d := \frac{d_0}{L} = 0.001, \qquad u_\infty = -0.011, \qquad \gamma = 0.08, \qquad \frac{a\mathrm{v}}{\kappa} = a\gamma^2 = 12.8\%.$$

The parameter ϵ is taken in the following range:

$$\epsilon = \frac{\varepsilon}{L} \in [0.0025, 0.04], \qquad \bar{\epsilon} = \frac{\varepsilon}{d_0} \in [2.5, 40].$$

The time window is $[0.390625, 3.515625]$ during which the interface moves from $R(t_0) = 0.1$ to $R(t_1) = 0.3$. The numerical results are reported in Table 3 and Figure 1(b).



(a) Computation 2 (b) Computation 3

(c) Computation 4

FIG. 1. *Horizontal axis is* $\log_{10}(\epsilon)$ *in* (a) *and* (b) *and* $\log_{10}(\varepsilon/cm)$ *in* (c). *The vertical axis is* $\log_{10}|1 - R_\epsilon/R|$. *Small dots correspond to Model* 1 *and large dots to Model* 2. *Straight lines represent the hypothetical formula* $|1 - R_\epsilon/R| = C\epsilon^k$ *with* $k = 1$ *for line with slope* 1 *and* $k = 2$ *for line with slope* 2.

TABLE 3
For Computation 3.

ϵ	Model 1		Model 2	
	$R_\epsilon(t_1)$	$\lvert 1 - R_\epsilon/R\rvert$	$R_\epsilon(t_1)$	$\lvert 1 - R_\epsilon/R\rvert$
0.0400	0.30169	5.6 e-3	0.29743	8.6 e-3
0.0200	0.30173	5.8 e-3	0.29941	2.0 e-3
0.0100	0.30105	3.5 e-3	0.29985	4.9 e-4
0.0050	0.30057	1.9 e-3	0.29996	1.3 e-4
0.0025	0.30030	9.9 e-4	0.29999	3.1 e-5
0	0.30000	0	0.30000	0

Performing a least squares analysis as in the previous example for Model 2, we have the result

$$\log|1 - R_\epsilon/R| = 0.741 + 2.01753\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	0.74126	0.03563
Log(ϵ)	2.01753	0.07058

with R-Sq = 100% and an F-value of 13402. Hence, for these parameters there is overwhelming evidence confirming the hypothetical exponent of 2. The coefficient calculated above differs from 1 by 1.0175/0.07058 = 14.42 standard deviations.

By comparison the data for Model 1 leads to the result

$$\log|1 - R_\epsilon/R| = -1.21 + 0.661\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	-1.21	0.2306
Log(ϵ)	0.661	0.1128

so that the relative difference between Model 1 and the exact free boundary solution behaves as $\epsilon^{0.661}$. A similar regression without the largest value of ϵ leads to $\epsilon^{0.85}$ power behavior.

Thus, one can conclude from this range of computations that Model 2 is within $O(\varepsilon^2)$, while Model 1 is even slightly worse than $O(\varepsilon)$ in these computations. Note that theorems establishing that Model 1 is $O(\varepsilon)$ are also of the form "there exists $\varepsilon_0 > 0$ such that for $\varepsilon < \varepsilon_0$ one has ...." Hence, the data in this range of parameters shows a significant practical improvement by using Model 2 in place of Model 1 that is analogous to the rigorous result.

Computation 4. Finally we provide an example using material data from succinonitrile. We take

$$D = 1.134 \times 10^{-3}\text{cm}^2/\text{s}, \qquad d_0 = 2.821 \times 10^{-7}\text{cm}, \qquad \frac{\ell}{c} = 23.13 \text{ Kelvin}, \qquad \alpha = 10^4\text{s}/\text{cm}^2.$$

Here D, d_0, and ℓ/c are from [24]. Note that there are no direct measurements on α and we choose α as in [4].

We focus on the part of the sample of size $L = 10^{-4}$ cm with undercooling $T_E - T_\infty = 0.2521$ Kelvin in a solidification process during which the solid ball grows from radius $R_0 = 10^{-5}$ cm to $R_1 = 4 \times 10^{-5}$ cm.

These dimensional numbers translate to the following dimensionless quantities:

$$a = \alpha D = 11.34, \qquad d := \frac{d_0}{L} = 0.002821, \qquad u_\infty = -\frac{0.2521}{23.13} = -0.0109,$$

$$\gamma = 0.079, \qquad \frac{a\text{v}}{\kappa} = a\gamma^2 = 7.2\%, \qquad t_0 = \left(\frac{R_0}{2\gamma L}\right)^2 = 0.40, \qquad t_1 = \left(\frac{R_1}{2\gamma L}\right)^2 = 6.40.$$

The amount of real time for such a solidification process takes $\frac{(t_1 - t_0)L^2}{D} = 5.2 \times 10^{-5}$ seconds.

To treat such a scenario within the capacity of computer power, we take ε in the range 10^{-5} cm to 2×10^{-7} cm. As we said earlier, the true size of ε is much smaller, but the interfacial motion is not very sensitive to the size of ε provided it is not very large. In dimensionless quantities, this translates to

$$\epsilon = \frac{\varepsilon}{L} \in [0.002,\ 0.1], \qquad \bar{\epsilon} = \frac{\varepsilon}{d_0} \in [1,\ 35], \qquad \bar{\epsilon}\,|u_\infty| \in [0.01,\ 0.35].$$

In Table 4 we list the relative differences between the solution of the phase field model and that of the free boundary problem, with ε given in cm.

Table 4

For Computation 4, the relative difference between solutions of the free boundary model and solutions of the phase field models 1 and 2.

ε	Model 1		Model 2	
	Position	$\vert 1 - R_\epsilon/R\vert$	Position	$\vert 1 - R_\epsilon/R\vert$
(cm)	(cm)		(cm)	
1 e-5	3.8435 e-5	3.9 e-2	3.8122 e-5	4.7 e-2
5 e-6	3.9723 e-5	6.9 e-3	3.9577 e-5	1.1 e-2
2 e-6	3.9993 e-5	1.8 e-4	3.9935 e-5	1.6 e-3
1 e-6	4.0013 e-5	3.3 e-4	3.9984 e-5	4.1 e-4
5 e-7	4.0011 e-5	2.7 e-4	3.9996 e-5	1.0 e-4
2 e-7	4.0005 e-5	1.3 e-4	3.9999 e-5	1.9 e-5
0	4.0000 e-5	0	4.0000 e-5	0

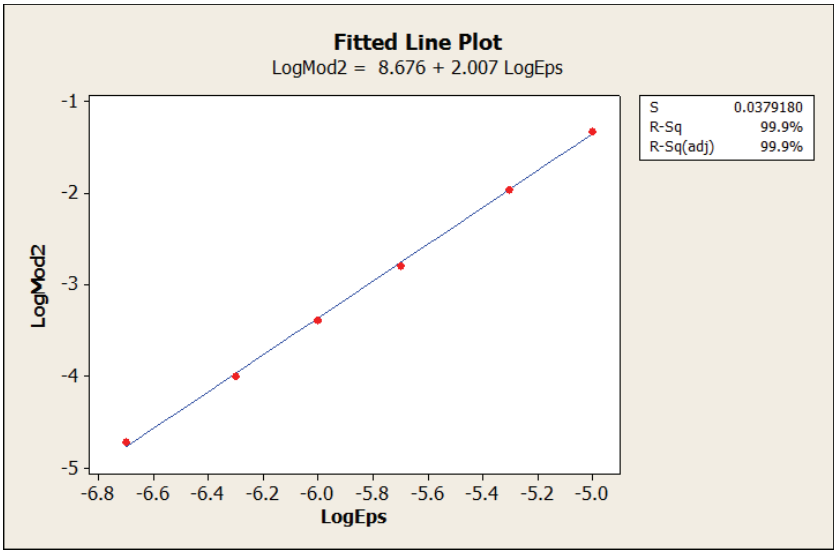


Fig. 2. *Computation 4, Model 2.*

Using the above procedure on Model 2, we obtain from the least squares analysis the result

$$\log|1 - R_\epsilon/R| = 8.6759 + 2.007\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	8.6759	0.1575
$\mathrm{Log}(\epsilon)$	2.007	0.02687

Hence, the exponent 2.007 differs from 1 by $1.007/0.02687 = 37.48$ standard deviations, establishing overwhelming evidence that the relative difference between Model 2 and the exact free boundary solution is better than linear in terms of ϵ. One also has that R-Sq $= 99.9\%$ and $F = 5580$ in the analysis of variance. As shown in Figure 2 the data points are indistinguishable from the straight line with slope 2.007. For this key set of physical parameters, the standard error of 0.02687 shows that the exponent is 2.007 ± 0.02687 so that the computational results are in agreement with the theoretical exponent of 2 in (4.1).

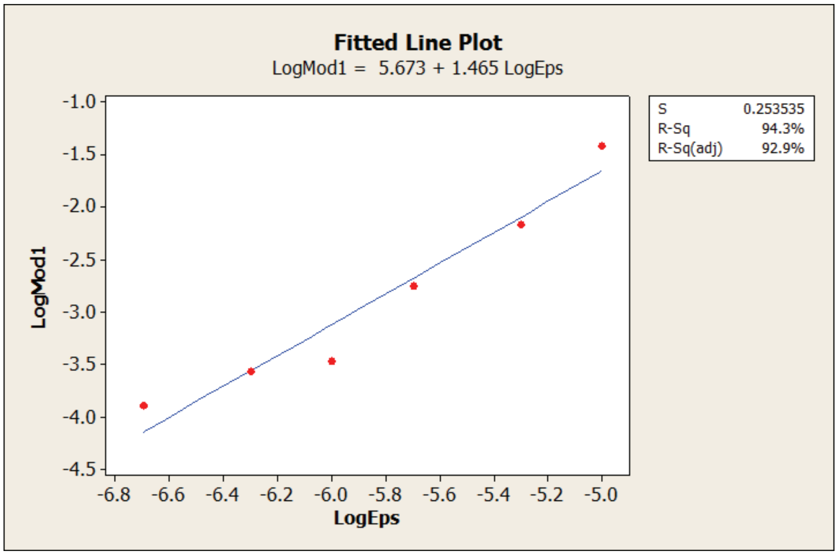


FIG. 3. *Computation 4, Model 1.*

The same analysis for Model 1 leads to the linear regression

$$\log|1 - R_\epsilon/R| = 5.673 + 1.465\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	5.673	1.053
Log(ϵ)	1.465	0.1796

The exponent of 1.47 ± 0.18 appears to be better than the theoretical expectation of 1. Note that the standard error of 0.18 is much larger than the corresponding 0.026 for Model 2 computed above. In Figure 3 one can observe that the slope appears to diminish for smaller ϵ. In particular, for the four smallest values of ϵ, one has the result

$$\log|1 - R_\epsilon/R| = 3.08 + 1.05\log(\epsilon)$$

Predictor	Coef	Std Error of Coef
Constant	3.08	1.793
Log(ϵ)	1.05	0.2899

Examining the practical differences between the exact solution and those rendered by Models 1 and 2 for the smallest ϵ in Table 4, one observes that the ratio of the error in Model 1 to the error in Model 2 is given by

$$\frac{1.3 \times 10^{-4}}{1.9 \times 10^{-5}} = 6.8421$$

so that a factor of almost seven is attained using Model 1. Note also that the improvement accuracy due to refining ε from 5×10^{-7} to 2×10^{-7} is $2.7/1.3 = 2.08$ for Model 1 but $10^{-4}/(1.9 \times 10^{-5}) = 5.2632$ for Model 2. Thus, one would expect that a calculation with $\varepsilon = 0.8 \times 10^{-7}$ would lead to a factor of

$$6.8421 \times \frac{5.2632}{2.08} = 17.313.$$

In other words, our analysis shows that for computing capacity that is capable of resolving the phase field model with $\varepsilon = 0.8 \times 10^{-7}$ the error (in approximating the free boundary) in Model 2 would be only $(17.313)^{-1} = 0.05776$, or less than 6% of the error of Model 1. Similarly, for $\varepsilon = 0.32 \times 10^{-7}$ the corresponding ratio would be

$$6.8421 \times \left(\frac{5.2632}{2.08}\right)^2 = 43.809,$$

leading to about 2% of the error.

The rigorous proof of second order convergence [10] is valid for $\varepsilon < \varepsilon_0$ for some positive ε_0. In any proof of this type one has no assurance that the ε_0 will be large enough to be of any practical significance. In the computations discussed above, particularly the last one in which we utilized material parameters of experiments, it is evident that one obtains this second order convergence using values of ε that are feasible with current computing capacity. Furthermore, there is the issue of the constant in (4.1). Although the constant in (4.1) is larger for Model 2 than for the corresponding expression for Model 1, the factor of ε^2 is small enough to render a much more accurate interface location (relative to the free boundary problem) as discussed above.

Hence, for computations using ε that is about half of the value we have used, one may conclude that our new phase field model (i.e., Model 2) can reduce the error in approximating a free boundary by a factor of 50.

5. Conclusion. We have presented numerical results for a classical phase field model and a new phase field model, demonstrating their asymptotic agreement with a free boundary (sharp interface) model using the Gibbs–Thomson condition and dynamical undercooling at the liquid-solid interface. For both phase field models, the interface, defined as the zero level set of the phase function, is compared with the free boundary of the sharp interface model which is the asymptotic limit of the phase field models. The results confirm the theoretical prediction that the distance between interface and free boundary is of order ε for the classical phase field model and of order ε^2 for the new model. Indeed, these asymptotic behaviors are seen more clearly in the new model than in the classical model. A well-behaved second order accuracy asymptotic behavior of the new model starts from a small ε which is much larger than that of the classical model, which is first order. While the classical model shows considerable deviations from its first order asymptotic behavior of approximating the free boundary model for ε that is not very small, the new model already demonstrates its second order approximation behavior. When ε is small, the new model always leads to a substantially better approximation than the classical one.

The theoretical assertion that the new phase field model is a second order accurate approximation of the free boundary model is derived in [10] from formal expansions in which $1/d := L/d_0$ is regarded as an order one constant and solutions are expanded in $\epsilon := \varepsilon/L$ power series. Here we omit the details of the formal asymptotic expansions and their rigorous verifications; we refer interested readers to the original formal expansions of Caginalp [5, 7] and rigorous verifications of Caginalp and Chen [9]. In reality it is true that $\epsilon = \varepsilon/L$ is smaller than $d = d_0/L$, but in numerical simulations such as those demonstrated in this paper, ε is taken as large as d_0; i.e., d is as small as ϵ.

In such a scenario, one can indeed assume that $\bar{\epsilon} := \varepsilon/d_0 = \epsilon/d$ is a fixed positive constant, expand the solution in ϵ or d power series, and demonstrate the following:

1. The leading (zeroth) order expansions of both the new and the classical phase

field models correspond to solutions of the classical Stefan problem, e.g., the solution (3.3) with $d = 0$.

2. The first order expansion of the solution of the new phase field model corresponds to a solution of the free boundary problem (2.1), e.g., the solution (3.3) with $0 < d \ll 1$.
3. The zero level set Γ_ϵ of the phase indicator function ϕ of the new phase field model is $O(\epsilon) = O(\bar{\epsilon}d)$ distance away from the free boundary of the classical Stefan model and is $O(\epsilon^2) = O(\bar{\epsilon}^2 d^2)$ distance away from that of the free boundary problem (2.1) (assuming that both free boundary problems admit smooth solutions).
4. On the other hand, the zero level set of the phase function of the classical phase field model is $O(\epsilon) = O(\bar{\epsilon}d)$ distance away from the free boundaries of both the Stefan problem and (2.1).

For the numerics of our current paper, which involve the mathematical limit of ε approaching zero, the computations are very close to the exact solutions even if ε/d_0 is not small. When $\bar{\epsilon} := \varepsilon/d_0$ is large in numerical simulation, the addition of $5\bar{\epsilon}/12$ to the kinetic undercooling coefficient from a to $a_\epsilon = a + 5\bar{\epsilon}/12$ can become significant.

To use the new phase field model (2.2) to approximate (2.1), one needs a resolution of order $o(d)$ at the interface. Since theoretical predication and numerical validation of this paper indicate that the error of this approximation at interface is $O(\epsilon^2)$, what we need is $\epsilon^2 = \bar{\epsilon}^2 d^2 = o(d)$, that is,

$$0 < \varepsilon \ll \sqrt{Ld_0}.$$

For example, in a dendritic growth experiment [22] with $d_0 = 8 \times 10^{-7}$ cm and $L = 0.8$ cm, the above criterion means that in numerical simulations using the new phase field model (2.2) to capture the Gibbs–Thomson condition, the parameter ε used should be smaller than $\sqrt{Ld_0} = 8 \times 10^{-4}$ cm, i.e., $\epsilon = \varepsilon/L < 0.001$. This amounts to thousands of grid points in each space dimension and millions of time steps for simulations of real experiments. Hence for values such as $\epsilon = 0.001$, yielding $\bar{\epsilon} = \varepsilon/d_0 = 1000$, there is a huge computational advantage in using the new phase field model with $a_\epsilon = a + 5\bar{\epsilon}/12$ replacing a of the traditional phase field model.

REFERENCES

[1] A. D. Alikakos, P. W. Bates, and X. Chen, *Convergence of the Cahn-Hilliard equation to the Hele-Shaw model*, Arch. Rational Mech. Anal., 128 (1994), pp. 165–205.

[2] S. Allen and J. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall., 27 (1979), pp. 1085–1095.

[3] R. F. Almgren, *Second-order phase field asymptotics for unequal conductivities*, SIAM J. Appl. Math., 59 (1999), pp. 2086–2107.

[4] Y. B. Altundas and G. Caginalp, *Computations of dendrites in 3-D and comparison with microgravity experiments*, J. Stat. Phys., 110 (2003), pp. 1055–1067.

[5] G. Caginalp, *An analysis of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 887–896.

[6] G. Caginalp, *The role of microscopic anisotropy in the macroscopic behavior of a phase boundary*, Ann. Physics, 172 (1986), pp. 136–155.

[7] G. Caginalp, *Limiting Behavior of a Free Boundary in the Phase Field Model*, CMU report 82-5, Carnegie Mellon University, Pittsburgh, PA, 1982.

[8] G. Caginalp and X. Chen, *Phase field equations in the singular limit of sharp interface problems*, in On the Evolution of Phase Boundaries, IMA Vol. Math. Appl. 43, M. Gurtin, ed., Springer, New York, 1992, pp. 1–27.

[9] G. Caginalp and X. Chen, *Convergence of the phase field model to its sharp interface limits*, European J. Appl. Math., 9 (1998), pp. 417–445.

[10] G. Caginalp, X. Chen, and C. Eck, *A rapidly converging phase field model*, Discrete Contin. Dyn. Syst., 15 (2006), pp. 1017–1034.

[11] G. Caginalp and C. Eck, *Rapidly converging phase field models via second order asymptotics*, Discrete Contin. Dyn. Syst., (2005), pp. 142–152.

[12] G. Caginalp and E. A. Socolovsky, *Efficient computation of a sharp interface by spreading via a phase field method*, Appl. Math. Lett., 2 (1989), p. 117.

[13] G. Caginalp and E. A. Socolovsky, *Computation of sharp phase boundaries by spreading: The planar and spherically symmetric cases*, J. Comput. Phys., 95 (1991), pp. 85–100.

[14] X. Chen, *Spectrums of the Allen–Cahn, Cahn–Hilliard, and phase field equations for generic interfaces*, Comm. Partial Differential Equations, 19 (1994), pp. 1371–1395.

[15] X. Chen and F. Reitich, *Local existence and uniqueness of solutions of the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[16] H. Garcke and B. Stinner, *Second order phase field asymptotics for multi-component systems*, Interfaces Free Bound., 8 (2006), pp. 131–157.

[17] J. W. Gibbs, *Collected Works*, Yale University Press, New Haven, CT, 1948.

[18] M. E. Glicksman, R. J. Schaefer, and J. D. Ayers, *Dendritic growth—A test of theory*, Met. Trans. A, 7A (1976), pp. 1747–1759.

[19] S. I. Hariharan and G. W. Young, *Comparison of asymptotic solutions of a phase-field model to a sharp-interface model*, SIAM J. Appl. Math., 62 (2001), pp. 244–263.

[20] A. Karma and W.-J. Rappel, *Quantitative phase-field modeling of dendritic growth in two and three dimensions*, Phys. Rev. E (3), 57 (1998), pp. 4323–4349.

[21] Y. Kim, N. Provatas, N. Goldenfeld, and J. Dantzig, *Universal dynamics of phase field models for dendritic growth*, Phys. Rev. E (3), 59 (1999), pp. 2546–2549.

[22] R. Kobayashi, *Modelling and numerical simulation of dendritic crystal growth*, Phys. D, 63 (1993), pp. 410–423.

[23] R. Kobayashi, J. A. Warren, and W. C. Carter, *Vector valued phase field model for crystallization and grain boundary formation*, Phys. D, 119 (1998), pp. 415–423.

[24] M. B. Koss, J. C. LaCombe, L. A. Tennenhouse, M. E. Glicksman, and E. A. Winsa, *Dendritic growth tip velocities and radii of curvature in microgravity*, Met. Mat. Trans. A, 30A (1999), pp. 3177–3190.

[25] G. Lamé and B. P. Clapeyron, *Memoire sur la solidification par refroidissement d'un globe solide*, Ann. Chem. Phys., 47 (1831), pp. 250–256.

[26] A. M. Meirmanov, *On a classical solution of the multidimensional Stefan problem for quasilinear parabolic equations*, Mat. Sb., 112 (1980), pp. 170–192.

[27] W. Mendenhall, *Introduction to Probability and Statistics*, PWS Publishers, Boston, MA, 1987.

[28] W. Mullins and R. F. Sekerka, *Stability of planar interface during solidification of a dilute binary alloy*, J. Appl. Phys., 35 (1964), pp. 444–451.

[29] R. H. Nochetto and C. Verdi, *Convergence past singularities for a fully discrete approximation of curvature-driven interfaces*, SIAM J. Numer. Anal., 34 (1997), pp. 490–512.

[30] J. Ockendon, *Linear and nonlinear stability of a class of moving boundary problems*, in Free Boundary Problems (Pavia, 1979), E. Magenes, ed., Ist. Naz. Alta Mat. Francesco Severi, Rome, 1980, pp. 443–447.

[31] L. I. Rubinstein, *The Stefan Problem*, Transl. Math. Monogr. 27, AMS, Providence, RI, 1971.

[32] R. F. Sekerka, P. W. Vorhees, S. R. Coriell, and G. B. McFadden, *Initial conditions implied by $t^{1/2}$ solidification of a sphere with capillarity and interfacial kinetics*, J. Crystal Growth, 87 (1988), pp. 415–420.

[33] H. M. Soner, *Convergence of the phase-field equations to the Mullins-Sekerka problem with kinetic undercooling*, Arch. Rational Mech. Anal., 131 (1995), pp. 139–197.

[34] J. Stefan, *Über einige Probleme der Theorie der Wärmeleitung*, S.-B. Wien Akad. Mat. Natur., 98 (1889), pp. 473–484.

[35] S.-L. Wang, R. F. Sekerka, A. A. Wheeler, B. T. Murray, S. R. Coriell, R. J. Braun, and G. B. McFadden, *Thermodynamically consistent phase-field models for solidification*, Phys. D, 69 (1993), pp. 189–200.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1535–1556

TIME REVERSAL FOCUSING OF THE INITIAL STATE FOR KIRCHHOFF PLATE*

KIM DANG PHUNG† AND XU ZHANG‡

Abstract. Consider a Kirchhoff plate $\partial_t^2 u + \Delta^2 u - \partial_t^2 \Delta u = 0$ in $\Omega \times (0,T)$, with boundary data $u = \Delta u = 0$ on $\partial\Omega \times (0,T)$ and unknown initial data $u(\cdot,0) = u_0$ and $\partial_t u(\cdot,0) = u_1$ in Ω. We study an inverse problem of determining (u_0,u_1) from an interior observation $u|_{\omega\times(0,T)}$. Here Ω is a bounded domain, ω a nonempty open subset of Ω, and $T > 0$ a suitable time duration. By means of an iterative time reversal technique, we derive an asymptotic formula of reconstructing (u_0,u_1) approximately with a logarithmical convergence rate for smooth initial data. The convergence becomes uniform and exponential when (Ω,ω,T) satisfies the geometric control condition introduced by Bardos, Lebeau, and Rauch.

Key words. Kirchhoff plate, inverse problem, quantitative unique continuation, observability estimate, time reversal technique

AMS subject classifications. Primary, 35R30; Secondary, 74K20, 93B07, 35B37

DOI. 10.1137/070684823

1. Introduction and main results.

Let $\Omega \subset \mathbb{R}^d$ ($d \in \mathbb{N}$) be a bounded open set with sufficiently smooth boundary $\partial\Omega$, ω a nonempty open subset of Ω, $T > 0$ a suitable time duration, and $\beta \in (0,1)$ any fixed parameter. Let $\mathcal{M} = \left(\alpha^{ij}\right)_{1\le i,j\le d} \in C^\infty\left(\overline{\Omega};\mathbb{R}^{d\times d}\right)$ be a symmetric and uniformly positive definite matrix (hence $(\beta^{ij})_{1\le i,j\le d} = \mathcal{M}^{1/2}$ is well defined). Denote by $1_{|\omega}$ the characteristic function of ω in Ω. Let $Q = \Omega \times (0,T)$ and $\Sigma = \partial\Omega \times (0,T)$. Throughout this paper, we shall use $C = C(\Omega,\omega,T,d,\beta,\mathcal{M})$ to denote a generic positive constant, which may change from line to line.

Denote by $\Delta = \sum_{i,j=1}^d \partial_{x_i}(\alpha^{ij}\partial_{x_j})$ the "Laplacian" associated to the matrix $\mathcal{M}$. We consider the following Kirchhoff plate equation in an inhomogeneous media:

$$\begin{cases} \partial_t^2 u + \Delta^2 u - \partial_t^2 \Delta u = 0 & \text{in } \Omega \times \mathbb{R}, \\ u = \Delta u = 0 & \text{on } \partial\Omega \times \mathbb{R}, \\ u(\cdot,0) = u_0, \quad \partial_t u(\cdot,0) = u_1 & \text{in } \Omega. \end{cases} \tag{1.1}$$

Let

$$\mathcal{H} \overset{\triangle}{=} \left\{ z \in H^3(\Omega) \;\middle|\; z = \Delta z = 0 \text{ on } \partial\Omega \right\} \times \left(H^2(\Omega) \cap H_0^1(\Omega) \right).$$

Clearly, $\mathcal{H}$ is a Hilbert space with the norm

$$||(u_0,u_1)||_{\mathcal{H}} \overset{\triangle}{=} \|(\nabla\Delta u_0, u_1, \Delta u_1)\|_{(L^2(\Omega))^d \times H_0^1(\Omega) \times L^2(\Omega)} .$$

*Received by the editors March 9, 2007; accepted for publication (in revised form) February 8, 2008; published electronically June 6, 2008. This work was supported by the NSF of China under grants 10525105 and 10771149, grant MTM2005-00714 of the Spanish MEC, and the Chinese Postdoctoral Science Foundation.

http://www.siam.org/journals/siap/68-6/68482.html

†Yangtze Center of Mathematics, Sichuan University, Chengdu 610064, China (kim_dang_phung@yahoo.fr).

‡Key Laboratory of Systems and Control, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China, and Yangtze Center of Mathematics, Sichuan University, Chengdu 610064, China (xuzhang@amss.ac.cn).

Here and henceforth, $\nabla = (\sum_{j=1}^{d}\beta^{1j}\partial_{x_j}, \ldots, \sum_{j=1}^{d}\beta^{dj}\partial_{x_j})$. It is easy to rewrite (1.1) as an abstract Cauchy problem in $\mathcal{H}$, with an unbounded operator $\mathcal{A} : D(\mathcal{A}) \subset \mathcal{H} \to \mathcal{H}$ as the generator of the underlying C_0-group. Hence, for any initial data $(u_0, u_1) \in \mathcal{H}$, system (1.1) is well-posed in $\mathcal{H}$. From the standard operator semigroup theory, $D(\mathcal{A}^k)$ ($k \in \mathbb{N}$) are themselves Hilbert spaces with the graph norms.

For any $z \in C(\mathbb{R}; \mathcal{H})$, we denote by $E(z, t)$ the functional

$$E(z,t) \overset{\triangle}{=} \frac{1}{2}\int_{\Omega}\left[|\nabla\Delta z(x,t)|^2 + |\nabla\partial_t z(x,t)|^2 + |\Delta\partial_t z(x,t)|^2\right] dx. \tag{1.2}$$

It is clear that $\mathcal{H}$ is the *finite energy space* of system (1.1), and its energy $E(\cdot, t)$ is conservative in the sense that for any u solution of (1.1) and all $t \in \mathbb{R}$,

$$E(u,t) = \frac{1}{2}||(u_0, u_1)||^2_{\mathcal{H}}. \tag{1.3}$$

The main purpose of this paper is to investigate the state-observation problem for system (1.1), which is formulated as follows: To determine the initial data (u_0, u_1) of a solution u of (1.1) from the single interior measurement $u|_{\omega\times(0,T)}$. It is well known that the state-observation problem is closely related to the inverse source problem, i.e., to determine the source term which causes the evolution process from the boundary and/or interior measurement. Inverse source problems of PDEs have been the object of numerous studies in recent years. Extensive related references can be found, say, in [18, 24, 25, 26] for the hyperbolic equations, in [22] for the Euler–Bernoulli plate equation, and in other works cited therein. Most of the references on inverse source problems cited above are addressed to global uniqueness and stability; here we give a constructive strategy to recover the initial data from a partial measurement of the solution. Our strategy for identification of source is inspired by the time reversal method and may be more practical than the formal tools of control theory (e.g., [23]). By means of an iterative time reversal technique, we further establish an asymptotic formula to reconstruct the desired initial state (u_0, u_1) of (1.1) by superposing different solutions of some Kirchhoff plates depending only on the measurement $u|_{\omega\times(0,T)}$.

More precisely, the knowledge of u on $\omega \times (0, T)$ allows us to consider a sequence of solutions $\{v^{(j)}\}_{j\geq 0}$ given as follows. First, let $U^{(-1)} = \frac{1}{2}u$ in $\omega \times (0, T)$. Next, define $v^{(j)} = v^{(j)}(x, t)$ ($j = 0, 1, 2, \ldots$) inductively to be the solution of the following system:

$$\begin{cases} \partial_t^2 v^{(j)} + \Delta^2 v^{(j)} - \partial_t^2\Delta v^{(j)} + \partial_t\Delta v^{(j)} \cdot 1_{|\omega} = -2\partial_t\Delta U^{(j-1)}(\cdot, T-t)\cdot 1_{|\omega} & \text{in } Q, \\ v^{(j)} = \Delta v^{(j)} = 0 & \text{on } \Sigma, \\ v^{(j)}(\cdot,0) = \partial_t v^{(j)}(\cdot,0) = 0 & \text{in } \Omega, \end{cases} \tag{1.4}$$

where $U^{(j)} = U^{(j)}(x, t)$ is given by

$$U^{(j)}(x,t) = \begin{cases} v^{(0)}(x,t) - u(x,T-t), & j = 0, \\ v^{(j)}(x,t) - U^{(j-1)}(x,T-t), & j > 0, \end{cases} \quad \text{for } (x,t) \in \omega \times (0,T). \tag{1.5}$$

Note that the values of functions $U^{(j)}$ are defined only in $\omega \times (0, T)$. Nevertheless, by system (1.4), it suffices to determine the values of $v^{(j)}$ in the whole domain Q

from $\partial_t \Delta U^{(j-1)}\big|_{\omega\times(0,T)}$. It is easy to see that the functions $v^{(j)}$ depend only on $\partial_t\Delta u$ restricted to $\omega\times(0,T)$.

We say that (Ω,ω,T_0) satisfies the classical geometric control condition (GCC), introduced in [2, 3], if $\partial\Omega$ is C^∞ with no contact of infinite order with its tangent, and any generalized bicharacteristic ray $(x(\rho),t(\rho))$ of $\partial_t^2-\Delta$ starting at $\rho=0$ with $t(0)=0$ meets $\omega\times(0,T_0)$ (see also [4] for an improvement on the regularity of $\partial\Omega$ and of $\mathcal{M}$). Notice that GCC can be rephrased by a geodesic condition (see [17]).

The main results of this paper are stated as follows.

Theorem 1.1. *Under GCC, for any $T\geq T_0$ there exists a constant $\sigma>0$ such that for any initial data $(u_0,u_1)\in\mathcal{H}$ and any $N>0$, it holds that*

$$\left\|\left(\sum_{k=0}^{N} v^{(2k)}(\cdot,T)-u_0,\ \sum_{k=0}^{N}\partial_t v^{(2k)}(\cdot,T)+u_1\right)\right\|_{\mathcal{H}} \leq Ce^{-\sigma N}\left\|(u_0,u_1)\right\|_{\mathcal{H}}. \tag{1.6}$$

Theorem 1.2. *Suppose*

$$J_3 \overset{\triangle}{=} \sup_{j>0}\left\|\left(\sum_{k=0}^{j} v^{(2k)}(\cdot,T),\ \sum_{k=0}^{j}\partial_t v^{(2k)}(\cdot,T)\right)\right\|_{D(\mathcal{A}^3)} < +\infty$$

and that Ω is connected. Then, for any nonempty open subset ω of Ω and any $\beta\in(0,1)$, there exists a time $T>0$ such that for any initial data $(u_0,u_1)\in D(\mathcal{A}^3)$ and any $N>0$, it holds that

$$\left\|\left(\sum_{k=0}^{N} v^{(2k)}(\cdot,T)-u_0,\ \sum_{k=0}^{N}\partial_t v^{(2k)}(\cdot,T)+u_1\right)\right\|_{\mathcal{H}} \leq \frac{C}{\ln^\beta(1+N)}[J_3+\|(u_0,u_1)\|_{D(\mathcal{A}^3)}]. \tag{1.7}$$

The above results say that $(\sum_{k=0}^N v^{(2k)}(\cdot,T),\ -\sum_{k=0}^N \partial_t v^{(2k)}(\cdot,T))$ can be employed to serve as an asymptotic formula to recover the initial state (u_0,u_1) of system (1.1). The key point to do this is the time reversibility of Kirchhoff plate. Fink (see [6, 7]) experimented with the time reversal mirror and succeeded in generating many applications (e.g., in biomedical engineering and telecommunication). Next, many mathematicians were also interested in this phenomenon (e.g., [1, 8, 19]). Thanks to the refocusing properties of the time-reversed waves, the time reversal technique has been successfully used to solve inverse problems for acoustic waves or electromagnetic waves (see, e.g., [5, 12]). Nevertheless, the main novelty in Theorems 1.1 and 1.2 is, respectively, the explicit exponential and logarithmical convergence rates for $(\sum_{k=0}^N v^{(2k)}(\cdot,T),\ -\sum_{k=0}^N \partial_t v^{(2k)}(\cdot,T))$ to approximate (u_0,u_1) in the strong topology of $\mathcal{H}$. Note also that Theorem 1.2 is for the case without GCC on (Ω,ω,T), for which one can usually expect a much weaker result than the case with GCC (we refer the reader to [14, 20] for a different yet related topic for the hyperbolic equations).

Technically, the proofs of Theorems 1.1 and 1.2 are reduced to suitable observability estimates for system (1.1). Under GCC, the desired observability estimate follows from the known result in [2, 3] for the wave equation. For the treatment in the case without GCC, by the Fourier–Bros–Iagolnitzer transformation given in [14], the obtaining of the desired observability estimate for the evolution system (1.1) depends on some quantitative unique continuation property for a fourth order elliptic-like equation with multiple characteristics (see (3.1)), which, in turn, will be established by means of global Carleman estimate. Although global Carleman estimates are well understood for many PDEs with single characteristics or without characteristics, it seems that there is no reference for the multiple-characteristic PDEs. The crucial point for the possibility of applying the Carleman estimate to the above-mentioned multiple-characteristic equation is that this equation can be rewritten equivalently as two coupled elliptic equations of second order, and that, based on a useful pointwise estimate for second order differential operators with symmetric coefficients (without any sign condition), we are successful in using Carleman estimates with a common weight function for these equations.

To end this section, we remark that, if the first equation in (1.4) is replaced by

$$\partial_t^2 v^{(j)} + \Delta^2 v^{(j)} - \partial_t^2 \Delta v^{(j)} - (-\Delta)^{-1}\big(\partial_t v^{(j)} \cdot 1_{|\omega}\big) = 2(-\Delta)^{-1}\big(\partial_t U^{(j-1)}(\cdot, T-t)\cdot 1_{|\omega}\big),$$

while (u_0, u_1) is assumed only to belong to $D(\mathcal{A})$, then, based on inequality (5.7) in Theorem 5.2, the estimate (1.7) in Theorem 1.2 becomes

$$\left\| \left(\sum_{k=0}^{N} v^{(2k)}(\cdot, T) - u_0,\ \sum_{k=0}^{N} \partial_t v^{(2k)}(\cdot, T) + u_1 \right) \right\|_{\mathcal{H}} \le \frac{C}{\ln^{\beta}(1+N)} [J_1 + \|(u_0, u_1)\|_{D(\mathcal{A})}].$$

The rest of this paper is organized as follows. In section 2, we derive the desired pointwise estimate for second order differential operators with symmetric coefficients. Section 3 shows an interpolation inequality for the fourth order elliptic-like equation with multiple characteristics mentioned above. Section 4 is devoted to a quantitative unique continuation property for system (1.1). In section 5, we establish two observability estimates for solutions of (1.1). The proofs of Theorems 1.1 and 1.2 are given in section 6.

2. Pointwise estimate for second order differential operators with symmetric coefficients.

In this section, we will establish a pointwise estimate for second order differential operators with symmetric coefficients (without any sign condition), which will play a key role in what follows.

Let $m \in \mathbb{N}$. For simplicity, for a function u, we will use the notation $u_i = \frac{\partial u}{\partial x_i}$, where x_i is the ith coordinate of a generic point $(x_1, \dots, x_m)$ in $\mathbb{R}^m$.

For any

$$a^{ij} = a^{ji} \in C^1(\mathbb{R}^m), \qquad i, j = 1, 2, \dots, m, \tag{2.1}$$

we recall the following known identity (see [10, Theorem 4.1], and also [9, Theorem 1.1] for a variant version).

Lemma 2.1. *Assume $u, \ell, \Psi \in C^2(\mathbb{R}^m)$. Let $\theta = e^{\ell}$ and $v = \theta u$. Then*

$$\begin{aligned} \theta^2 \left| \sum_{i,j=1}^{m} (a^{ij} u_i)_j \right|^2 &+ 2\sum_{j=1}^{m} \Bigg\{ 2 \sum_{i,i',j'=1}^{m} a^{ij} a^{i'j'} \ell_{i'} v_i v_{j'} - \sum_{i,i',j'=1}^{m} a^{ij} a^{i'j'} \ell_i v_{i'} v_{j'} \\ &+ \Psi \sum_{i=1}^{m} a^{ij} v_i v - \sum_{i=1}^{m} a^{ij} \left[(A+\Psi)\ell_i + \frac{\Psi_i}{2} \right] v^2 \Bigg\}_j \\ = 2 \sum_{i,j=1}^{m} c^{ij} v_i v_j + B v^2 &+ \left| \sum_{i,j=1}^{m} (a^{ij} v_i)_j - A v \right|^2 + 4 \left| \sum_{i,j=1}^{m} a^{ij} \ell_i v_j \right|^2, \end{aligned} \tag{2.2}$$

where

$$\left\{\begin{aligned} &A \stackrel{\triangle}{=} -\sum_{i,j=1}^m \Big(a^{ij}\ell_i\ell_j - a^{ij}_j\ell_i - a^{ij}\ell_{ij}\Big) - \Psi,\\ &B \stackrel{\triangle}{=} 2\left\{A\Psi - \sum_{i,j=1}^m \big[(A+\Psi)a^{ij}\ell_i\big]_j\right\} + \Psi^2 - \sum_{i,j=1}^m \big(a^{ij}\Psi_j\big)_i,\\ &c^{ij} \stackrel{\triangle}{=} \sum_{i',j'=1}^m \Big[2a^{ij'}(a^{i'j}\ell_{i'})_{j'} - (a^{ij}a^{i'j'}\ell_{i'})_{j'}\Big] + \Psi a^{ij}. \end{aligned}\right. \tag{2.3}$$

In what follows, for any function $\psi \in C^4(\mathbb{R}^m)$, and any (large) parameters $\varsigma > 1$ and $\kappa > 1$, we choose the function ℓ in Lemma 2.1 as follows:

$$\ell = \varsigma\varphi, \qquad \varphi = e^{\kappa\psi}. \tag{2.4}$$

It is easy to check that

$$\ell_i = \varsigma\kappa\varphi\psi_i, \qquad \ell_{ij} = \varsigma\kappa^2\varphi\psi_i\psi_j + \varsigma\kappa\varphi\psi_{ij}, \qquad i,j = 1,2,\dots,m. \tag{2.5}$$

For $n \in \mathbb{N}$, we denote by $O(\kappa^n)$ a function of order κ^n for large κ (which is independent of ς); by $O_\kappa(\varsigma^n)$ a function of order ς^n for fixed κ and for large ς. The desired pointwise estimate for the operator "$\sum_{i,j=1}^m \frac{\partial}{\partial_{x_j}}\big(a^{ij}\frac{\partial}{\partial_{x_i}}\big)$" is stated as follows.

THEOREM 2.2. *Assume* (2.1) *holds, and* $u \in C^2(\mathbb{R}^m)$. *Let*

$$\theta = e^\ell, \quad v = \theta u, \quad \Psi = 2\sum_{i,j=1}^m a^{ij}\ell_{ij}. \tag{2.6}$$

Then

$$\begin{aligned} &\theta^2\left|\sum_{i,j=1}^m (a^{ij}u_i)_j\right|^2 + 2\sum_{i,j=1}^m \Bigg\{\sum_{i',j'=1}^m \Big[2a^{ij}a^{i'j'}\ell_{i'}v_iv_{j'} - a^{ij}a^{i'j'}\ell_iv_{i'}v_{j'}\Big]\\ &\qquad + \Psi a^{ij}v_iv - a^{ij}\left[(A+\Psi)\ell_i + \frac{\Psi_i}{2}\right]v^2\Bigg\}_j\\ &\geq 2\sum_{i,j=1}^m c^{ij}v_iv_j + Bv^2, \end{aligned} \tag{2.7}$$

where A, B, *and* c^{ij} *are given in* (2.3). *Moreover, for* ς *and* κ *large enough, the following estimates hold uniformly in any bounded set of* $\mathbb{R}^m$:

(2.8)

$$\begin{aligned} \sum_{i,j=1}^m c^{ij}v_iv_j \geq \varsigma\kappa\varphi\Bigg\{&\kappa\left(\sum_{i,j=1}^m a^{ij}\psi_i\psi_j\right)\left(\sum_{i,j=1}^m a^{ij}v_iv_j\right) + \sum_{i,j,i',j'=1}^m \Big[2a^{ij'}a^{i'j}\psi_{i'j'}\\ &+ a^{ij}a^{i'j'}\psi_{i'j'} + 2a^{ij'}a^{i'j}_{j'}\psi_{i'} - (a^{ij}a^{i'j'})_{j'}\psi_{i'}\Big]v_iv_j\Bigg\}, \end{aligned}$$

$$B = 2\varsigma^3\kappa^4\varphi^3\left(\sum_{i,j=1}^m a^{ij}\psi_i\psi_j\right)^2 + \varsigma^3\varphi^3O(\kappa^3) + O_\kappa(\varsigma^2).$$

Proof. Clearly, (2.7) is a direct consequence of Lemma 2.1. Recalling (2.3) for c^{ij}, and noting (2.6) and (2.5), we have

$$\begin{aligned}
&\sum_{i,j=1}^{m} c^{ij} v_i v_j \\
&= \sum_{i,j,i',j'=1}^{m} \Big[2a^{ij'}a^{i'j}\ell_{i'j'} + a^{ij}a^{i'j'}\ell_{i'j'} + 2a^{ij'}a^{i'j}_{j'}\ell_{i'} - (a^{ij}a^{i'j'})_{j'}\ell_{i'}\Big] v_i v_j \\
&= 2\varsigma\kappa^2\varphi\left(\sum_{i,j=1}^{m} a^{ij}\psi_i v_j\right)^2 + \varsigma\kappa^2\varphi\left(\sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j\right)\left(\sum_{i,j=1}^{m} a^{ij} v_i v_j\right) \\
&\quad + \varsigma\kappa\varphi \sum_{i,j,i',j'=1}^{m} \Big[2a^{ij'}a^{i'j}\psi_{i'j'} + a^{ij}a^{i'j'}\psi_{i'j'} + 2a^{ij'}a^{i'j}_{j'}\psi_{i'} - (a^{ij}a^{i'j'})_{j'}\psi_{i'}\Big] v_i v_j \\
&\geq \varsigma\kappa^2\varphi\left(\sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j\right)\left(\sum_{i,j=1}^{m} a^{ij} v_i v_j\right) \\
&\quad + \varsigma\kappa\varphi \sum_{i,j,i',j'=1}^{m} \Big[2a^{ij'}a^{i'j}\psi_{i'j'} + a^{ij}a^{i'j'}\psi_{i'j'} + 2a^{ij'}a^{i'j}_{j'}\psi_{i'} - (a^{ij}a^{i'j'})_{j'}\psi_{i'}\Big] v_i v_j,
\end{aligned}$$

which gives the first inequality in (2.8).

On the other hand, by (2.5), recalling the definitions of Ψ and A, we see that

$$\Psi = 2\varsigma\kappa^2\varphi \sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j + \varsigma\varphi O(\kappa), \qquad A = -\varsigma^2\kappa^2\varphi^2 \sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j + O_\kappa(\varsigma).$$

Hence, from the definition of B, we have

$$\begin{aligned}
B \;&= 2\Bigg\{ -2\varsigma^3\kappa^4\varphi^3\left(\sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j\right)^2 + \varsigma^3\varphi^3 O(\kappa^3) + O_\kappa(\varsigma^2) \\
&\qquad + \varsigma\kappa \sum_{i,j=1}^{m}\left[\left(\varsigma^2\kappa^2\varphi^3 \sum_{i',j'=1}^{m} a^{i'j'}\psi_{i'}\psi_{j'} + O_\kappa(\varsigma)\right) a^{ij}\psi_i\right]_j \Bigg\} + O_\kappa(\varsigma^2) \\
&= 2\Bigg\{ -2\varsigma^3\kappa^4\varphi^3\left(\sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j\right)^2 + \varsigma^3\varphi^3 O(\kappa^3) + O_\kappa(\varsigma^2) \\
&\qquad + \varsigma\kappa \sum_{i,j=1}^{m}\left(3\varsigma^2\kappa^3\varphi^3 \sum_{i',j'=1}^{m} a^{i'j'}\psi_{i'}\psi_{j'} + \varsigma^2\varphi^3 O(\kappa^2) + O_\kappa(\varsigma)\right) a^{ij}\psi_i\psi_j \Bigg\} \\
&= 2\varsigma^3\kappa^4\varphi^3\left(\sum_{i,j=1}^{m} a^{ij}\psi_i\psi_j\right)^2 + \varsigma^3\varphi^3 O(\kappa^3) + O_\kappa(\varsigma^2),
\end{aligned}$$

which yields the second inequality in (2.8). □

3. Interpolation inequality for a fourth order elliptic-like equation with multiple characteristics.

As a crucial preliminary, we derive in this section the following a priori estimate for a fourth order elliptic-like equation with multiple characteristics.

THEOREM 3.1. *Suppose that Ω is connected. Then, for any nonempty open subset ω of Ω, there exists a constant $C_0 = C_0(\Omega, \omega, d, \mathcal{M}) > 0$ such that for any $w = w(x,s) \in H^2(\Omega \times (-2,2))$ and $f = f(x,s) \in L^2(\Omega \times (-2,2))$ with*

$$\begin{cases} -\partial_s^2 w + \Delta^2 w + \partial_s^2 \Delta w = f & \text{in } \Omega \times (-2,2)\,, \\ w = \Delta w = 0 & \text{on } \partial\Omega \times (-2,2) \end{cases} \tag{3.1}$$

we have

$$\begin{aligned} \int_{-1}^{1}\int_{\Omega} |\Delta w|^2 \, dxds \le \; & C_0 e^{C_0/\varepsilon} \left[\int_{-2}^{2}\int_{\omega} \left(|w|^2 + |\Delta w|^2 \right) dxds + \int_{-2}^{2}\int_{\Omega} |f|^2 \, dxds \right] \\ & + e^{-2/\varepsilon} \int_{-2}^{2}\int_{\Omega} \left(|\Delta w|^2 + |\partial_s \Delta w|^2 \right) dxds \qquad \forall\, \varepsilon > 0. \end{aligned} \tag{3.2}$$

Notice that this interpolation estimate (3.2) or Hölder dependence continuous inequality has already appeared in [13] for second order elliptic operators in the framework of null controllability for the heat equation.

Before proving Theorem 3.1, we remark that inequality (3.2) is a kind of quantitative unique continuation of (3.1) in the following sense: If $w \in H^2(\Omega \times (-2,2))$ solves (3.1) with $f = 0$ in $\Omega \times (-2,2)$, and $w = 0$ in $\omega \times (-2,2)$, then, by Theorem 3.1, $w = 0$ in $\Omega \times (-1,1)$. On the other hand, it is easy to verify that any solution w to (3.1) with $f = 0$ in $\Omega \times (-2,2)$ is of the form

$$w(x,s) = \sum_{k=1}^{\infty} \left(a_k e^{s\sqrt{1+\lambda_k - \frac{1}{1+\lambda_k}}} + b_k e^{-s\sqrt{1+\lambda_k - \frac{1}{1+\lambda_k}}} \right) \varphi_k(x), \qquad a_k,\, b_k \in \mathbb{C},$$

where $\{\lambda_k\}_{k\ge 1}$ are the eigenvalues of $-\Delta$ with homogeneous Dirichlet boundary condition and $\{\varphi_k\}_{k=1}^{\infty}$ the corresponding eigenvectors (constituting an orthonormal basis of $L^2(\Omega)$). Therefore, $w(\cdot, s)$ is analytic with respect to s, which, in turn, implies $w = 0$ in $\Omega \times (-2,2)$.

Note also that (3.1) is not elliptic in the classical sense. Indeed, the symbol of its principal operator reads $\xi^4 + \xi^2\eta^2$, which vanishes for $\xi = 0$ and any $\eta \in \mathbb{R}$. As mentioned in the introduction, we use global Carleman estimates to establish (3.2). To do this, a key observation is the possibility of decomposing the operator $-\partial_s^2 + \Delta^2 + \partial_s^2\Delta$ as follows:

$$-\partial_s^2 + \Delta^2 + \partial_s^2\Delta = (\partial_s^2 + \Delta)(-I + \Delta) + \Delta, \tag{3.3}$$

where I is the identity. Consequently, in order to derive the desired inequality (3.2), it is natural to proceed in cascade by applying the global Carleman estimates to the second order elliptic operators $\partial_s^2 + \Delta$ and Δ. Thanks to Theorem 2.2, this is possible because only the symmetry of the matrix $(a^{ij})_{1\le i,j\le m}$ is required. Therefore, Theorem 2.2 applies to both the operators $\partial_s^2 + \Delta$ and Δ. We remark that, due to the necessity of using same weight function for these two different elliptic operators of second order, there seems no existing Carleman estimate in the literature for our purpose.

Proof of Theorem 3.1. The proof is divided into four steps.

Step 1. *Choice of the weight function.* In order to apply Theorem 2.2 in cascade to the operators $\partial_s^2 + \Delta$ and Δ, it is important to choose a common weight function $\theta = \theta(x, s)$ for these two different operators.

It is well known that (see [11] or [21], for example) there is a function $\widehat{\psi} \in C^4\left(\overline{\Omega}\right)$ such that $\widehat{\psi} > 0$ in Ω, $\widehat{\psi} = 0$ on $\partial\Omega$, and

$$0 < \sum_{i=1}^{d} \left|\partial_{x_i}\widehat{\psi}(x)\right|^2 \le C\left|\nabla\widehat{\psi}(x)\right|^2 \qquad \forall\, x \in \overline{\Omega\setminus\omega_0} \tag{3.4}$$

where $\omega_0 \subset \omega$ is an arbitrary fixed nonempty open subset of Ω such that $\overline{\omega_0} \subset \omega$. Therefore,

$$h \overset{\triangle}{=} \frac{1}{||\widehat{\psi}\,||_{L^\infty(\Omega)}} \min_{x\in\overline{\Omega\setminus\omega_0}} |\nabla\widehat{\psi}(x)| > 0. \tag{3.5}$$

Let us introduce

$$b = \sqrt{1 + \frac{1}{\kappa}\ln\left(2 + e^\kappa\right)}, \qquad b_0 = \sqrt{b^2 - \frac{1}{\kappa}\ln\left(\frac{1+e^\kappa}{e^\kappa}\right)}, \tag{3.6}$$

where $\kappa > \ln 2$ is the parameter that appeared in Theorem 2.2 and is chosen large enough. It is easy to see that

$$1 < b_0 < b \le 2.$$

Further, we choose

$$\psi(x, s) = \frac{\widehat{\psi}(x)}{||\widehat{\psi}\,||_{L^\infty(\Omega)}} + b^2 - s^2. \tag{3.7}$$

By (2.4) and (2.6), this gives the function $\varphi\left(x, s\right) = e^{\kappa\psi(x,s)}$ and the desired weight function $\theta\left(x, s\right) = e^{\varsigma e^{\kappa\psi(x,s)}}$ (recall Theorem 2.2 for the parameter ς). It is easy to check that

$$\begin{cases} \varphi\left(\cdot, s\right) \ge 2 + e^\kappa & \text{for any } s \text{ satisfying } |s| \le 1, \\ \varphi\left(\cdot, s\right) \le 1 + e^\kappa & \text{for any } s \text{ satisfying } b_0 \le |s| \le b. \end{cases} \tag{3.8}$$

Step 2. *Reduction of* (3.1) *to a cascade system.* Let

$$z = -w + \Delta w. \tag{3.9}$$

Then, in view of (3.3), system (3.1) can be written equivalently as the following elliptic system of second order in cascade:

$$\begin{cases} \Delta w = z + w & \text{in } \Omega\times\left(-2, 2\right), \\ \partial_s^2 z + \Delta z = f - z - w & \text{in } \Omega\times\left(-2, 2\right), \\ w = z = 0 & \text{on } \partial\Omega\times\left(-2, 2\right). \end{cases} \tag{3.10}$$

Note, however, that there is no (homogeneous) boundary condition for w (and hence z) at $s = \pm 2$. Now, we introduce a cut-off function $\phi = \phi(s) \in C_0^\infty(-b, b) \subset C_0^\infty(\mathbb{R})$ such that

$$\begin{cases} 0 \le \phi(s) \le 1, & |s| < b, \\ \phi(s) \equiv 1, & |s| \le b_0. \end{cases} \tag{3.11}$$

Let

$$\tilde{w} = \phi w, \qquad \tilde{z} = \phi z. \tag{3.12}$$

Then, noticing that ϕ does not depend on x, it follows by (3.10) that

$$\begin{cases} \Delta \tilde{w} = \tilde{z} + \tilde{w} & \text{in } \Omega \times (-2, 2), \\ \partial_s^2 \tilde{z} + \Delta \tilde{z} = \phi f + 2\partial_s \phi \partial_s z + z \partial_s^2 \phi - \tilde{z} - \tilde{w} & \text{in } \Omega \times (-2, 2), \\ \tilde{w} = \tilde{z} = 0 & \text{on } \partial\Omega \times (-2, 2), \\ \operatorname{supp} \tilde{w}(x, \cdot) \bigcup \operatorname{supp} \tilde{z}(x, \cdot) \subset (-b, b), & x \in \Omega. \end{cases} \tag{3.13}$$

Step 3. *Carleman estimates.* First, we apply Theorem 2.2 with $m = d + 1$, $x_{d+1} = s$, $(a^{ij})_{1 \le i,j \le d+1} = \left(\begin{smallmatrix} \mathcal{M} & 0 \\ 0 & 0 \end{smallmatrix}\right)$, u replaced by $\tilde{w}$, and the weight function θ given above. In this case, by the definition of c^{ij} in (2.3), it is easy to check that

$$c^{ij} = 0 \qquad \text{whenever one of } i \text{ and } j \text{ is equal to } d + 1. \tag{3.14}$$

Moreover, by (2.8), recalling (3.7) for the definition of ψ and (3.5) for the positive constant h, we conclude that there is a constant $\kappa_0 > 1$ such that for any $\kappa \ge \kappa_0$, one can find a constant $\varsigma_0 > 1$ so that for any $\varsigma \ge \varsigma_0$, the following estimates hold uniformly for $(x, s) \in \overline{\Omega \setminus \omega_0} \times [-b, b]$:

$$\begin{aligned} &\sum_{i,j=1}^{d} c^{ij} v_i v_j \ge \varsigma\varphi \Big\{ \kappa^2 |\nabla\psi|^2 + O(\kappa) \Big\} |\nabla v|^2 \ge \frac{h^2}{2} \varsigma \kappa^2 \varphi |\nabla v|^2, \\ &B = 2\varsigma^3 \kappa^4 \varphi^3 |\nabla\psi|^4 + \varsigma^3 \varphi^3 O(\kappa^3) + O_\kappa(\varsigma^2) \ge h^4 \varsigma^3 \kappa^4 \varphi^3, \end{aligned} \tag{3.15}$$

where $v = \theta\tilde{w}$. Now, integrating inequality (2.7) (with u replaced by $\tilde{w}$) of Theorem 2.2 in $\Omega \times (-b, b)$, recalling that ϕ vanishes near $s = \pm b$, $\widehat{\psi} = 0$ on $\partial\Omega$, and $v = 0$ on $\partial\Omega \times (-b, b)$, noting (3.15) and the first equation in (3.13), one arrives at

$$\begin{aligned} &\frac{1}{C} \left\{ \varsigma\kappa^2 \int_{-b}^{b} \int_\Omega \varphi |\nabla v|^2 dx ds + \varsigma^3 \kappa^4 \int_{-b}^{b} \int_\Omega \varphi^3 |v|^2 dx ds \right\} \\ &\le \int_{-b}^{b} \int_\Omega \theta^2 \left| \tilde{z} + \tilde{w} \right|^2 dx ds + \frac{\varsigma\kappa}{||\widehat{\psi}||_{L^\infty(\Omega)}} \int_{-b}^{b} \int_{\partial\Omega} \varphi \frac{\partial \widehat{\psi}}{\partial \nu} \left| \frac{\partial v}{\partial \nu_{\mathcal{M}}} \right|^2 d(\partial\Omega) ds \\ &\quad + C \left\{ \varsigma\kappa^2 \int_{-b}^{b} \int_{\omega_0} \theta^2 \varphi |\nabla \tilde{w}|^2 dx ds + \varsigma^3 \kappa^4 \int_{-b}^{b} \int_{\omega_0} \theta^2 \varphi^3 |\tilde{w}|^2 dx ds \right\}, \end{aligned} \tag{3.16}$$

where $\frac{\partial \widehat{\psi}}{\partial \nu} = \sum_{i=1}^{d} \widehat{\psi}_i \nu^i$, $\frac{\partial v}{\partial \nu_{\mathcal{M}}} = \sum_{i,j=1}^{d} \alpha^{ij} v_i \nu^j$, and $\nu = \left(\nu^1, \ldots, \nu^d\right) = \nu(x)$ is the unit outward normal vector of Ω at $x \in \partial\Omega$. For the boundary term in (3.16), we have used that

$$\sum_{i,j=1}^{d+1}\left\{\sum_{i',j'=1}^{d+1}\left[2a^{ij}a^{i'j'}\ell_{i'}v_iv_{j'}-a^{ij}a^{i'j'}\ell_iv_{i'}v_{j'}\right]\right\}\nu^j$$

$$=\frac{\varsigma\kappa\varphi}{||\widehat{\psi}\,||_{L^\infty(\Omega)}}\left(\sum_{i=1}^{d}\widehat{\psi}_i\nu^i\right)\left|\sum_{i,j=1}^{d}a^{ij}v_i\nu^j\right|^2=\frac{\varsigma\kappa\varphi}{||\widehat{\psi}\,||_{L^\infty(\Omega)}}\frac{\partial\widehat{\psi}}{\partial\nu}\left|\frac{\partial v}{\partial\nu_{\mathcal{M}}}\right|^2,$$

which follows from the fact that on $\partial\Omega\times(-b,b)$, we have for $j=1,\dots,d$,

$$v_j=\left(\sum_{i=1}^{d}v_i\nu^i\right)\nu^j,\, \ell_j=\varsigma\kappa\varphi\psi_j=\frac{\varsigma\kappa\varphi}{||\widehat{\psi}\,||_{L^\infty(\Omega)}}\widehat{\psi}_j=\frac{\varsigma\kappa\varphi}{||\widehat{\psi}\,||_{L^\infty(\Omega)}}\left(\sum_{i=1}^{d}\widehat{\psi}_i\nu^i\right)\nu^j,$$

and $\nu^{d+1}=0$.

Choose a cut-off function $g\in C_0^\infty(\omega)$ with $g\equiv1$ in ω_0 and $0\le g\le1$ in ω. Multiplying the first equation in (3.13) by $g\theta^2\varphi\tilde{w}$ and integrating it in $\Omega\times(-b,b)$, using integration by parts, one obtains

$$\int_{-b}^{b}\int_{\omega_0}\theta^2\varphi|\nabla\tilde{w}|^2dxds\le C\left[\varsigma\kappa^2\int_{-b}^{b}\int_{\omega}\theta^2\varphi^2|\tilde{w}|^2dxds+\int_{-b}^{b}\int_{\omega}\theta^2|\tilde{z}|^2dxds\right]. \tag{3.17}$$

Recalling that $v=\theta\tilde{w}$, by (2.5), we get

$$\frac{1}{C}\theta^2(|\nabla\tilde{w}|^2+\varsigma^2\kappa^2\varphi^2|\tilde{w}|^2)\le|\nabla v|^2+\varsigma^2\kappa^2\varphi^2v^2\le C\theta^2(|\nabla\tilde{w}|^2+\varsigma^2\kappa^2\varphi^2|\tilde{w}|^2). \tag{3.18}$$

By the choice of $\widehat{\psi}$, one can check that $\frac{\partial\widehat{\psi}}{\partial\nu}<0$ on $\partial\Omega$. Therefore, by (3.16) and noting (3.17)–(3.18), we end up with

$$\begin{aligned}&\varsigma\kappa^2\int_{-b}^{b}\int_{\Omega}\theta^2\varphi|\nabla\tilde{w}|^2dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_{\Omega}\theta^2\varphi^3|\tilde{w}|^2dxds\\&\le C\left(\int_{-b}^{b}\int_{\Omega}\theta^2\left|\tilde{z}\right|^2dxds+\varsigma\kappa^2\int_{-b}^{b}\int_{\omega}\theta^2\left|\tilde{z}\right|^2dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_{\omega}\theta^2\varphi^3|\tilde{w}|^2dxds\right).\end{aligned}\tag{3.19}$$

Next, we apply Theorem 2.2 with $m=d+1$, $x_{d+1}=s$, $(a^{ij})_{1\le i,j\le d+1}=\left(\begin{smallmatrix}\mathcal{M}&0\\0&1\end{smallmatrix}\right)$, u replaced by $\tilde{z}$, and the weight function θ as the above. In this case, for any fixed $b_1\in(0,b)$, by (2.8), recalling again (3.7) for the definition of ψ and (3.5) for the positive constant h, we conclude that there is a constant $\kappa_1\ge\kappa_0$ such that for any $\kappa\ge\kappa_1$, one can find a constant $\varsigma_1\ge\varsigma_0$ so that, for any $\varsigma\ge\varsigma_1$, the following estimates hold uniformly for $(x,s)\in(\Omega\times(-b,b))\setminus(\omega_0\times(-b_1,b_1))$:

$$\begin{aligned}\sum_{i,j=1}^{d+1}c^{ij}p_ip_j&\ge\varsigma\varphi\Big[\kappa^2(|\nabla\psi|^2+|\partial_s\psi|^2)+O(\kappa)\Big](|\nabla p|^2+|\partial_sp|^2)\\&\ge\frac{h^2}{2}\varsigma\kappa^2\varphi(|\nabla p|^2+|\partial_sp|^2),\end{aligned}\tag{3.20}$$

$$B=2\varsigma^3\kappa^4\varphi^3(|\nabla\psi|^2+|\partial_s\psi|^2)^2+\varsigma^3\varphi^3O(\kappa^3)+O_\kappa(\varsigma^2)\ge h^4\varsigma^3\kappa^4\varphi^3,$$

where $p = \theta\tilde{z}$. Using (3.20) and the second equation in (3.13), similar to the proof of (3.16), one obtains

(3.21)
$$\begin{aligned}
&\frac{1}{C}\left\{\varsigma\kappa^2\int_{-b}^{b}\int_\Omega \varphi(|\nabla p|^2+|\partial_s p|^2)dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_\Omega \varphi^3|p|^2dxds\right\}\\
&\leq \int_{-b}^{b}\int_\Omega \theta^2\left|\phi f+2\partial_s\phi\partial_s z+z\partial_s^2\phi-\tilde{z}-\tilde{w}\right|^2dxds\\
&\quad+\frac{\varsigma\kappa}{||\widehat{\psi}\,||_{L^\infty(\Omega)}}\int_{-b}^{b}\int_{\partial\Omega}\varphi\frac{\partial\widehat{\psi}}{\partial\nu}\left|\frac{\partial p}{\partial\nu_{\mathcal{M}}}\right|^2d(\partial\Omega)ds\\
&\quad+C\left\{\varsigma\kappa^2\int_{-b_1}^{b_1}\int_{\omega_0}\theta^2\varphi(|\nabla\tilde{z}|^2+|\partial_s\tilde{z}|^2)dxds+\varsigma^3\kappa^4\int_{-b_1}^{b_1}\int_{\omega_0}\theta^2\varphi^3|\tilde{z}|^2dxds\right\}.
\end{aligned}$$

By the second equation in (3.13), similar to (3.17), one has

(3.22)
$$\begin{aligned}
&\int_{-b_1}^{b_1}\int_{\omega_0}\theta^2\varphi(|\nabla\tilde{z}|^2+|\partial_s\tilde{z}|^2)dxds\\
&\leq C\left[\varsigma\kappa^2\int_{-b}^{b}\int_\omega\theta^2\varphi^2|\tilde{z}|^2dxds+\int_{-b}^{b}\int_\omega\theta^2\left|\phi f+2\partial_s\phi\partial_s z+z\partial_s^2\phi-\tilde{w}\right|^2dxds\right].
\end{aligned}$$

Now, similar to (3.19), from (3.21) and (3.22), it follows that

$$\begin{aligned}
&\varsigma\kappa^2\int_{-b}^{b}\int_\Omega\theta^2\varphi(|\nabla\tilde{z}|^2+|\partial_s\tilde{z}|^2)dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_\Omega\theta^2\varphi^3|\tilde{z}|^2dxds\\
&\leq C\Bigg\{\varsigma\kappa^2\int_{-b}^{b}\int_\Omega\theta^2\left|\phi f+2\partial_s\phi\partial_s z+z\partial_s^2\phi\right|^2dxds+\int_{-b}^{b}\int_\Omega\theta^2\left|\tilde{w}\right|^2dxds\\
&\quad+\varsigma\kappa^2\int_{-b}^{b}\int_\omega\theta^2\left|\tilde{w}\right|^2dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_\omega\theta^2\varphi^3|\tilde{z}|^2dxds\Bigg\}.
\end{aligned}\tag{3.23}$$

Combining (3.19) and (3.23), we find that for any ς and κ large enough,

$$\begin{aligned}
&\int_{-b}^{b}\int_\Omega\theta^2\varphi(|\nabla\tilde{z}|^2+|\partial_s\tilde{z}|^2)dxds+\varsigma^2\kappa^2\int_{-b}^{b}\int_\Omega\theta^2\varphi^3|\tilde{z}|^2dxds\\
&\quad+\varsigma^3\kappa^4\int_{-b}^{b}\int_\Omega\theta^2\varphi|\nabla\tilde{w}|^2dxds+\varsigma^5\kappa^6\int_{-b}^{b}\int_\Omega\theta^2\varphi^3|\tilde{w}|^2dxds\\
&\leq C\Bigg(\int_{-b}^{b}\int_\Omega\theta^2\left|\phi f+2\partial_s\phi\partial_s z+z\partial_s^2\phi\right|^2dxds+\varsigma^3\kappa^4\int_{-b}^{b}\int_\omega\theta^2\left|\tilde{z}\right|^2dxds\\
&\quad+\varsigma^5\kappa^6\int_{-b}^{b}\int_\omega\theta^2\varphi^3|\tilde{w}|^2dxds\Bigg).
\end{aligned}\tag{3.24}$$

Recall that $\tilde{z} = \Delta\tilde{w} - \tilde{w}$. Therefore, (3.24) leads to

$$\begin{aligned} & \varsigma^2\kappa^2 \int_{-b}^{b}\int_{\Omega} \theta^2\varphi^3|\Delta\tilde{w}|^2 dxds \\ & \leq C\Bigg[\int_{-b}^{b}\int_{\Omega} \theta^2(|f|^2 + \left|2\partial_s\phi\partial_s z + z\partial_s^2\phi\right|^2) dxds \\ & \qquad + \varsigma^3\kappa^4 \int_{-b}^{b}\int_{\omega} \theta^2\varphi^3|\Delta\tilde{w}|^2 dxds + \varsigma^5\kappa^6 \int_{-b}^{b}\int_{\omega} \theta^2\varphi^3|\tilde{w}|^2 dxds \Bigg]. \end{aligned} \tag{3.25}$$

Step 4. *Completion of the proof.* Denote $c_0 = 2 + e^{\kappa} > 1$, and recall (3.6) for $b_0 \in (1, b)$. Fixing the parameter κ in (3.25), by (3.9), using (3.8) and (3.11), one finds

$$\begin{aligned} & \varsigma^2 e^{2\varsigma c_0} \int_{-1}^{1}\int_{\Omega} |\Delta w|^2\, dxds \\ & \leq Ce^{C\varsigma}\left[\int_{-b}^{b}\int_{\Omega} |f|^2 dxds + \int_{-b}^{b}\int_{\omega} (|w|^2 + |\Delta w|^2) dxds \right] \\ & \quad + Ce^{2\varsigma(c_0-1)} \int_{(-b,-b_0)\cup(b_0,b)} \int_{\Omega} \left|2\partial_s\phi\partial_s(\Delta w - w) + (\Delta w - w)\partial_s^2\phi\right|^2 dxds. \end{aligned} \tag{3.26}$$

From (3.26), one concludes that there exists $\varepsilon_0 > 0$ such that the desired inequality (3.2) holds for $\varepsilon \in (0, \varepsilon_0]$, which, in turn, implies that it holds for any $\varepsilon > 0$. This completes the proof of Theorem 3.1. □

4. Quantitative unique continuation for Kirchhoff plate. This section shows the following quantitative unique continuation for solutions of (1.1).

Theorem 4.1. *Suppose that* Ω *is connected. Then, for any nonempty open subset* ω *of* Ω *and any* $\beta \in (0,1)$, *there exists a time* $T > 0$ *such that the solution* u *of* (1.1) *satisfies*

$$\|(u_0,u_1)\|_{H^2(\Omega)\times H^1(\Omega)}^2 \leq e^{C\left[\frac{\|(u_0,u_1)\|_{\mathcal{H}}}{\|(u_0,u_1)\|_{H^2(\Omega)\times H^1(\Omega)}}\right]^{1/\beta}} \int_0^T \int_{\omega} |u\,(x,t)|^2\, dxdt \qquad \forall\ (u_0,u_1) \in \mathcal{H}\setminus\{0\}. \tag{4.1}$$

Remark 4.1. *Estimate* (4.1) *is equivalent to*

$$\|(u_0,u_1)\|_{H^2(\Omega)\times H^1(\Omega)}^2 \leq \frac{C}{\ln^{2\beta}\left(1 + \frac{\|(u_0,u_1)\|_{\mathcal{H}}^2}{\|u\|_{L^2(\omega\times(0,T))}^2}\right)} \|(u_0,u_1)\|_{\mathcal{H}}^2 \tag{4.2}$$

or, equivalently,

$$\|(u_0,u_1)\|_{H^2(\Omega)\times H^1(\Omega)}^2 \leq Ce^{C/\mu} \int_0^T\int_{\omega} |u\,(x,t)|^2\, dxdt + \mu^{2\beta}\, \|(u_0,u_1)\|_{\mathcal{H}}^2 \quad \forall\, \mu > 0. \tag{4.3}$$

This kind of interpolation estimate or logarithmic dependence continuous inequality has already appeared in [14, 20] in the framework of boundary control and stabilization for hyperbolic equations.

In order to prove Theorem 4.1, we need the following known result from [14, p. 473].

PROPOSITION 4.2. *For any $N \in \mathbb{N}$, the function*

$$F(z) \triangleq \frac{1}{2\pi} \int_{\mathbb{R}} e^{iz\tau} e^{-\tau^{2N}} d\tau \tag{4.4}$$

is holomorphic in $\mathbb{C}$ and there exist four positive constants A, c_0, c_1, and c_2 such that

$$\begin{cases} |F(z)| \leq A e^{c_0 |\operatorname{Im} z|^{\alpha}}, \\ |\operatorname{Im} z| \leq c_2 |\operatorname{Re} z| \Longrightarrow |F(z)| \leq A e^{-c_1 |z|^{\alpha}}, \end{cases}$$

where $\alpha = \frac{2N}{2N-1}$.

Proof of Theorem 4.1. Fix any nonempty open subset ω_1 of ω such that $\overline{\omega_1} \subset \omega$. We claim that it suffices to show that

$$\begin{aligned} \|(u_0, u_1)\|^2_{H^2(\Omega)\times H^1(\Omega)} &\leq C e^{C/\mu} \int_0^T \int_{\omega_1} \left[|u(x,t)|^2 + |\Delta u(x,t)|^2 \right] dxdt \\ &\quad + \mu^{2\beta} \|(u_0, u_1)\|^2_{\mathcal{H}} \qquad \forall \mu > 0. \end{aligned} \tag{4.5}$$

To see this, we choose a cut-off function $\varrho \in C_0^{\infty}(\omega)$ such that $\varrho = 1$ in ω_1 and $0 \leq \varrho \leq 1$ in ω. Then, for any $\mu > 0$,

$$\begin{aligned} &\int_0^T \int_{\omega_1} \left[|u(x,t)|^2 + |\Delta u(x,t)|^2 \right] dxdt \leq \int_0^T \|\varrho u(\cdot,t)\|^2_{H^2(\mathbb{R}^d)} dt \\ &\leq C \left[\frac{1}{\mu^2} \int_0^T \|\varrho u(\cdot,t)\|^2_{L^2(\mathbb{R}^d)} dt + \mu \int_0^T \|\varrho u(\cdot,t)\|^2_{H^3(\mathbb{R}^d)} dt \right] \\ &\leq C \left[\frac{1}{\mu^2} \int_0^T \int_{\omega} |u(x,t)|^2 dxdt + \mu \|(u_0, u_1)\|^2_{\mathcal{H}} \right]. \end{aligned} \tag{4.6}$$

Combining (4.5) and (4.6), we arrive at (4.3). By Remark 4.1, this yields the desired inequality (4.1).

We now prove (4.5) and divide the proof into three steps.

Step 1. *Reducing the problem to a fourth order elliptic-like equation.* Let $\beta \in (0,1)$, and choose $N \in \mathbb{N}$ such that $0 < \beta + \frac{1}{2N} < 1$. Let $\gamma = \frac{1}{\alpha} = 1 - \frac{1}{2N} > \beta$. Recall the definition of $F(z)$ in (4.4) and the constant $C_0 > 0$ in Theorem 3.1 (with ω replaced by ω_1). By Proposition 4.2, for any $\lambda \geq 1$, the holomorphic function

$$F_{\lambda}(z) \triangleq \lambda^{\gamma} F(\lambda^{\gamma} z) \equiv \frac{1}{2\pi} \int_{\mathbb{R}} e^{iz\tau} e^{-\left(\frac{\tau}{\lambda^{\gamma}}\right)^{2N}} d\tau \tag{4.7}$$

satisfies

$$\begin{cases} |\operatorname{Im} z| \leq 2 \Longrightarrow |F_{\lambda}(z)| \leq A \lambda^{\gamma} e^{2^{\alpha} c_0 \lambda}, \\ |\operatorname{Im} z| \leq 2 \text{ and } \frac{2}{c_2} \leq |\operatorname{Re} z| \Longrightarrow |F_{\lambda}(z)| \leq A \lambda^{\gamma} e^{-c_1 \lambda |\operatorname{Re} z|^{\alpha}}. \end{cases} \tag{4.8}$$

In what follows, we fix a time T satisfying

$$T > 8 \max\left(2, \frac{2}{c_2}, \sqrt[\alpha]{\frac{1 + 2^{\alpha} c_0 C_0}{c_1}} \right). \tag{4.9}$$

For any $\Phi = \Phi(t) \in C_0^\infty(0,T) \subset C_0^\infty(\mathbb{R})$ and any solution u of (1.1), following [14], we introduce the following *Fourier–Bros–Iagolnitzer transformation* of u:

$$W_{\ell_0,\lambda}(x,s) = \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\Phi(\ell)u(x,\ell)d\ell, \qquad s, \ell_0 \in \mathbb{R}. \tag{4.10}$$

Clearly, $\partial_s^2 F_\lambda(\ell_0 + is - \ell) = -\partial_\ell^2 F_\lambda(\ell_0 + is - \ell)$. Hence

$$\begin{aligned}
&\partial_s^2 (I-\Delta) W_{\ell_0,\lambda}(x,s)\\
&= -\int_{\mathbb{R}} \partial_\ell^2 F_\lambda(\ell_0 + is - \ell)\Phi(\ell)(I-\Delta)u(x,\ell)d\ell\\
&= -\int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\left[\Phi''(\ell)(I-\Delta)u(x,\ell) + 2\Phi'(\ell)\partial_t (I-\Delta)u(x,\ell)\right]d\ell\\
&\quad -\int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\Phi(\ell)\partial_t^2 (I-\Delta)u(x,\ell)d\ell.
\end{aligned} \tag{4.11}$$

Since u is a solution of (1.1), it follows from (4.11) that $W_{\ell_0,\lambda}$ satisfies the following fourth order elliptic-like equation with multiple characteristics:

$$\begin{cases}
-\partial_s^2 W_{\ell_0,\lambda}(x,s) + \Delta^2 W_{\ell_0,\lambda}(x,s) + \partial_s^2 \Delta W_{\ell_0,\lambda}(x,s)\\
\quad = \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell)\left[\Phi''(\ell)(I-\Delta)u(x,\ell) + 2\Phi'(\ell)\partial_t (I-\Delta)u(x,\ell)\right]d\ell,\\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (x,s) \in \Omega \times \mathbb{R},\\
W_{\ell_0,\lambda}(x,s) = \Delta W_{\ell_0,\lambda}(x,s) = 0, \qquad (x,s) \in \partial\Omega \times \mathbb{R},\\
W_{\ell_0,\lambda}(x,0) = (F_\lambda * \Phi u(x,\cdot))(\ell_0), \qquad x \in \Omega.
\end{cases} \tag{4.12}$$

On the other hand, we have

$$\begin{aligned}
&\|\Phi\Delta u(x,\cdot)\|_{L^2\left(\frac{T}{2}-1,\frac{T}{2}+1\right)}\\
&\le \|\Phi\Delta u(x,\cdot) - F_\lambda * \Phi\Delta u(x,\cdot)\|_{L^2\left(\frac{T}{2}-1,\frac{T}{2}+1\right)} + \|F_\lambda * \Phi\Delta u(x,\cdot)\|_{L^2\left(\frac{T}{2}-1,\frac{T}{2}+1\right)}\\
&\le \|\Phi\Delta u(x,\cdot) - F_\lambda * \Phi\Delta u(x,\cdot)\|_{L^2(\mathbb{R})} + \left(\int_{T/2-1}^{T/2+1} |\Delta W_{t,\lambda}(x,0)|^2 dt\right)^{1/2}.
\end{aligned}$$

Denote by $\mathcal{F}(f)$ the Fourier transform of f. Therefore, using Parseval's equality and the fact that $\mathcal{F}(F_\lambda)(\tau) = e^{-\left(\frac{\tau}{\lambda^\gamma}\right)^{2N}}$, one finds

$$\begin{aligned}
&\|\Phi\Delta u(x,\cdot) - F_\lambda * \Phi\Delta u(x,\cdot)\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}}\left\|\mathcal{F}\Big(\Phi\Delta u(x,\cdot) - F_\lambda * \Phi\Delta u(x,\cdot)\Big)\right\|_{L^2(\mathbb{R})}\\
&= \frac{1}{\sqrt{2\pi}}\left(\int_{\mathbb{R}}\left|\left(1 - e^{-\left(\frac{\tau}{\lambda^\gamma}\right)^{2N}}\right)\mathcal{F}\Big(\Phi\Delta u(x,\cdot)\Big)(\tau)\right|^2 d\tau\right)^{1/2}\\
&\le C\left(\int_{\mathbb{R}}\left|\frac{\tau}{\lambda^\gamma}\mathcal{F}\Big(\Phi\Delta u(x,\cdot)\Big)(\tau)\right|^2 d\tau\right)^{1/2} \le \frac{C}{\lambda^\gamma}\|\partial_t(\Phi\Delta u(x,\cdot))\|_{L^2(\mathbb{R})}.
\end{aligned}$$

Hence,

$$\begin{aligned}&\int_{T/2-1}^{T/2+1} |\Phi(t)\Delta u(x,t)|^2\,dt \\ &\le C\left[\frac{1}{\lambda^{2\gamma}}\left\|\partial_t\left(\Phi\Delta u(x,\cdot)\right)\right\|_{L^2(\mathbb{R})}^2 + \int_{T/2-1}^{T/2+1}|\Delta W_{t,\lambda}(x,0)|^2\,dt\right].\end{aligned} \tag{4.13}$$

For any $t\in\left(\frac{T}{2}-1,\frac{T}{2}+1\right)\subset\mathbb{R}$, by Cauchy's integral theorem (in the theory of complex variable functions) and Hölder's inequality, we deduce that

$$\begin{aligned}|\Delta W_{t,\lambda}(x,0)| &\le \frac{1}{\pi}\int_{|\ell_0-t|\le 1}\int_{|s|\le 1}|\Delta W_{\ell_0,\lambda}(x,s)|\,dsd\ell_0 \\ &\le \frac{2}{\pi}\left(\int_{|\ell_0-t|\le 1}\int_{|s|\le 1}|\Delta W_{\ell_0,\lambda}(x,s)|^2\,dsd\ell_0\right)^{1/2}.\end{aligned}$$

Hence

$$\begin{aligned}\int_{T/2-1}^{T/2+1}|\Delta W_{t,\lambda}(x,0)|^2\,dt &\le \frac{4}{\pi^2}\int_{T/2-1}^{T/2+1}dt\int_{|\ell_0-t|\le 1}\int_{|s|\le 1}|\Delta W_{\ell_0,\lambda}(x,s)|^2\,dsd\ell_0 \\ &\le \frac{4}{\pi^2}\int_{T/2-1}^{T/2+1}dt\int_{T/2-2}^{T/2+2}d\ell_0\int_{|s|\le 1}|\Delta W_{\ell_0,\lambda}(x,s)|^2\,ds \\ &\le \frac{8}{\pi^2}\int_{T/2-2}^{T/2+2}d\ell_0\int_{|s|\le 1}|\Delta W_{\ell_0,\lambda}(x,s)|^2\,ds.\end{aligned} \tag{4.14}$$

Noticing that

$$\int_\Omega \left\|\partial_t\left(\Phi\Delta u(x,\cdot)\right)\right\|_{L^2(\mathbb{R})}^2 dx \le C\left\|(u_0,u_1)\right\|_{\mathcal{H}}^2,$$

combining (4.13) and (4.14), we get

$$\begin{aligned}&\int_{T/2-1}^{T/2+1}\int_\Omega |\Phi(t)\Delta u(x,t)|^2\,dxdt \\ &\le C\left[\frac{1}{\lambda^{2\gamma}}\left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 + \int_{T/2-2}^{T/2+2}d\ell_0\int_{|s|\le 1}\int_\Omega|\Delta W_{\ell_0,\lambda}(x,s)|^2\,dxds\right].\end{aligned} \tag{4.15}$$

Step 2. *The estimate on* $\int_{T/2-2}^{T/2+2}d\ell_0\int_{|s|\le 1}\int_\Omega|\Delta W_{\ell_0,\lambda}(x,s)|^2\,dxds$. We now fix any $\Phi=\Phi(t)\in C_0^\infty(0,T)$ satisfying $0\le\Phi\le 1$ in $(0,T)$ and $\Phi\equiv 1$ on $\left[\frac{T}{4},\frac{3T}{4}\right]$.

Applying Theorem 3.1 (with ω replaced by ω_1) to $W_{\ell_0,\lambda}$, we obtain that for all $\varepsilon>0$,

$$\begin{aligned}&\int_{|s|\le 1}\int_\Omega|\Delta W_{\ell_0,\lambda}(x,s)|^2\,dxds \\ &\le e^{-2/\varepsilon}\int_{|s|\le 2}\int_\Omega\left[|\Delta W_{\ell_0,\lambda}(x,s)|^2+|\partial_s\Delta W_{\ell_0,\lambda}(x,s)|^2\right]dxds \\ &+\ C_0e^{C_0/\varepsilon}\int_{|s|\le 2}\int_{\omega_1}\left[|W_{\ell_0,\lambda}(x,s)|^2+|\Delta W_{\ell_0,\lambda}(x,s)|^2\right]dxds\end{aligned} \tag{4.16}$$

$$+ C_0 e^{C_0/\varepsilon} \int_{|s|\leq 2} \int_\Omega \Big| \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell) \Big[\Phi''(\ell) \left(I - \Delta\right) u(x,\ell) + 2\Phi'(\ell) \partial_t \left(I-\Delta\right) u(x,\ell) \Big] d\ell \Big|^2 dxds.$$

Using the first conclusion in (4.8), we deduce that

$$\begin{aligned}
&\int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 2} \int_\Omega \left[\left|\Delta W_{\ell_0,\lambda}(x,s)\right|^2 + \left|\partial_s \Delta W_{\ell_0,\lambda}(x,s)\right|^2 \right] dxds \\
&= \int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 2} \int_\Omega \left[\left| \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell) \Phi(\ell) \Delta u(x,\ell) d\ell \right|^2 \right. \\
&\qquad \left. + \left| \partial_s \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell) \Phi(\ell) \Delta u(x,\ell) d\ell \right|^2 \right] dxds \\
&\leq \int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 2} \int_\Omega \left[\left| \int_0^T \left(A\lambda^\gamma e^{2^\alpha c_0 \lambda} \right) \left|\Delta u(x,\ell)\right| d\ell \right|^2 \right. \\
&\qquad \left. + \left| \int_0^T \left(A\lambda^\gamma e^{2^\alpha c_0 \lambda} \right) \left|\Delta \partial_\ell u(x,\ell) + \Phi'(\ell) \Delta u(x,\ell)\right| d\ell \right|^2 \right] dxds \\
&\leq C\lambda^{2\gamma} e^{2^{\alpha+1} c_0 \lambda} \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 .
\end{aligned} \tag{4.17}$$

Similarly,

$$\begin{aligned}
&\int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 2} \int_{\omega_1} \left[\left|W_{\ell_0,\lambda}(x,s)\right|^2 + \left|\Delta W_{\ell_0,\lambda}(x,s)\right|^2 \right] dxds \\
&\leq C\lambda^{2\gamma} e^{2^{\alpha+1} c_0 \lambda} \int_0^T \int_{\omega_1} \left[\left|u(x,t)\right|^2 + \left|\Delta u(x,t)\right|^2 \right] dxdt.
\end{aligned} \tag{4.18}$$

Further, by the choice of Φ, it is obvious that

$$\operatorname{supp}\left(\partial_t^2 \Phi\right) \subset \operatorname{supp}\left(\partial_t \Phi\right) \subset K \triangleq \left[0, \frac{T}{4}\right] \bigcup \left[\frac{3T}{4}, T\right].$$

Let $K_0 = \left[\frac{3T}{8}, \frac{5T}{8}\right]$. Then dist $(K, K_0) = \frac{T}{8}$. The choice of T in (4.9) implies $T > 16$ and $T > 16/c_2$. Hence, $\left(\frac{T}{2} - 2, \frac{T}{2} + 2\right) \subset K_0$, and

$$|\ell_0 - \ell| \geq \frac{T}{8} \geq \frac{2}{c_2} \qquad \forall\ (\ell_0, \ell) \in K_0 \times K.$$

Therefore, using the second conclusion in (4.8), we deduce that

$$\begin{aligned}
&\int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 2} \int_\Omega \left| \int_{\mathbb{R}} F_\lambda(\ell_0 + is - \ell) \Big[\Phi''(\ell) \left(I-\Delta\right) u(x,\ell) \right. \\
&\qquad \left. +2\Phi'(\ell) \partial_t \left(I-\Delta\right) u(x,\ell) \Big] d\ell \right|^2 dxds \\
&\leq C \int_{K_0} d\ell_0 \int_{|s|\leq 2} \int_\Omega \left| \int_K \left(A\lambda^\gamma e^{-c_1 \lambda |\ell_0 - \ell|^\alpha} \right) \Big(\left|\left(I-\Delta\right) u(x,\ell)\right| \right.
\end{aligned} \tag{4.19}$$

$$+ \left|\partial_t \left(I-\Delta\right) u(x,\ell)\right|\Big) d\ell\Big|^2 dxds$$

$$\leq C \left|A\lambda^{\gamma} e^{-c_1\lambda(\mathrm{dist}(K,K_0))^{\alpha}}\right|^2 \int_\Omega \left|\int_K [\left|\left(I-\Delta\right) u(x,\ell)\right| + \left|\partial_t \left(I-\Delta\right) u(x,\ell)\right|] \, d\ell\right|^2 dx$$

$$\leq C\lambda^{2\gamma} e^{-2c_1\lambda\left(\frac{T}{8}\right)^{\alpha}} \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 .$$

Combining (4.17)–(4.19), we arrive at

$$\begin{aligned}&\int_{T/2-2}^{T/2+2} d\ell_0 \int_{|s|\leq 1} \int_\Omega \left|\Delta W_{\ell_0,\lambda}(x,s)\right|^2 dxds \\ &\leq C\lambda^{2\gamma} \Bigg\{ e^{2^{\alpha+1}c_0\lambda} e^{-2/\varepsilon} \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 + e^{-2c_1\lambda\left(\frac{T}{8}\right)^{\alpha}} e^{C_0/\varepsilon} \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 \\ &\qquad + e^{2^{\alpha+1}c_0\lambda} e^{C_0/\varepsilon} \int_0^T \int_{\omega_1} \left[\left|u(x,t)\right|^2 + \left|\Delta u(x,t)\right|^2\right] dxdt \Bigg\}.\end{aligned} \tag{4.20}$$

Step 3. *Choice of* ε *and completion of the proof.* We deduce from (4.15) and (4.20) that

$$\begin{aligned}&\int_{T/2-1}^{T/2+1} \int_\Omega \left|\Phi(t)\Delta u(x,t)\right|^2 dxdt \\ &\leq C \Bigg\{ \left[\frac{1}{\lambda^{2\gamma}} + \lambda^{2\gamma} e^{2^{\alpha+1}c_0\lambda} e^{-2/\varepsilon} + \lambda^{2\gamma} e^{-2c_1\lambda\left(\frac{T}{8}\right)^{\alpha}} e^{C_0/\varepsilon}\right] \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 \\ &\qquad + \lambda^{2\gamma} e^{2^{\alpha+1}c_0\lambda} e^{C_0/\varepsilon} \int_0^T \int_{\omega_1} \left[\left|u(x,t)\right|^2 + \left|\Delta u(x,t)\right|^2\right] dxdt \Bigg\}.\end{aligned} \tag{4.21}$$

We now choose

$$\varepsilon = \frac{1}{2^{\alpha+1}c_0\lambda}. \tag{4.22}$$

Recall the choice of T in (4.9), which gives $-c_1\left(\frac{T}{8}\right)^{\alpha} + 2^{\alpha}c_0C_0 \leq -1$. Hence

$$e^{-2c_1\lambda\left(\frac{T}{8}\right)^{\alpha}} e^{C_0/\varepsilon} = \exp\left\{2\left[-c_1\left(\frac{T}{8}\right)^{\alpha} + 2^{\alpha}c_0C_0\right]\lambda\right\} \leq e^{-2\lambda}. \tag{4.23}$$

Combining (4.21)–(4.23), we deduce that, for all $\lambda \geq 1$,

$$\begin{aligned}&\int_{T/2-1}^{T/2+1} \int_\Omega \left|\Delta u(x,t)\right|^2 dxdt = \int_{T/2-1}^{T/2+1} \int_\Omega \left|\Phi(t)\Delta u(x,t)\right|^2 dxdt \\ &\leq \frac{C}{\lambda^{2\gamma}} \left\|(u_0,u_1)\right\|_{\mathcal{H}}^2 + Ce^{C\lambda} \int_0^T \int_{\omega_1} \left[\left|u(x,t)\right|^2 + \left|\Delta u(x,t)\right|^2\right] dxdt.\end{aligned} \tag{4.24}$$

Finally, by means of the energy method, it is easy to show that

$$\left\|(u_0,u_1)\right\|_{H^2(\Omega)\times H^1(\Omega)}^2 \leq C \int_{T/2-1}^{T/2+1} \int_\Omega \left|\Delta u(x,t)\right|^2 dxdt. \tag{4.25}$$

Combining (4.24) and (4.25), it is easy to conclude that, for any $\mu \in (0,1)$,

$$\begin{aligned} &\|(u_0,u_1)\|^2_{H^2(\Omega)\times H^1(\Omega)} \\ &\leq Ce^{C/\mu}\int_0^T\int_{\omega_1}\left[|u(x,t)|^2+|\Delta u(x,t)|^2\right]dxdt + C\mu^{2\beta}\|(u_0,u_1)\|^2_{\mathcal{H}}. \end{aligned} \tag{4.26}$$

Note that this inequality is trivially true for any $\mu \geq 1$ and for some $C > 1$. Consequently, inequality (4.5) holds true for all $\mu > 0$. This completes the proof of Theorem 4.1. □

5. Observability estimates for Kirchhoff plate. The purpose of this section is to establish two observability estimates which quantify the unique continuation property of (1.1) saying that if its solution u satisfies $\partial_t\Delta u = 0$ in $\omega \times (0,T)$, then $(u_0,u_1) = 0$. Due to the finite speed of propagation, the time $T > 0$ has to be chosen large enough.

First, when GCC is assumed, we have the following estimate.

THEOREM 5.1. *Under GCC, for any $T \geq T_0$, the solution u of* (1.1) *satisfies*

$$E(u,0) \leq C\int_0^T\int_\omega |\partial_t\Delta u(x,t)|^2\,dxdt \qquad \forall\,(u_0,u_1)\in\mathcal{H}. \tag{5.1}$$

Proof. Under GCC, it follows from [2, 3] that for any $g \in L^2(\Omega\times(0,T_0))$ and any $(\chi_0,\chi_1)\in H_0^1(\Omega)\times L^2(\Omega)$, the solution χ of the wave equation

$$\begin{cases} \partial_t^2\chi - \Delta\chi = g & \text{in } \Omega\times(0,T_0), \\ \chi = 0 & \text{on } \partial\Omega\times(0,T_0), \\ \chi(\cdot,0)=\chi_0, \quad \partial_t\chi(\cdot,0)=\chi_1 & \text{in } \Omega \end{cases} \tag{5.2}$$

satisfies the following observability estimate:

$$\|(\chi_0,\chi_1)\|^2_{H_0^1(\Omega)\times L^2(\Omega)} \leq C\left[\int_0^{T_0}\int_\omega|\partial_t\chi(x,t)|^2\,dxdt + \int_0^{T_0}\int_\Omega|g(x,t)|^2\,dxdt\right]. \tag{5.3}$$

Note that the solution u of (1.1) solves

$$\begin{cases} \partial_t^2(I-\Delta)u - \Delta(I-\Delta)u = -\Delta u & \text{in } \Omega\times\mathbb{R}, \\ (I-\Delta)u = 0 & \text{on } \partial\Omega\times\mathbb{R}, \\ (I-\Delta)u(\cdot,0) = (I-\Delta)u_0, \quad \partial_t(I-\Delta)u(\cdot,0) = (I-\Delta)u_1 & \text{in } \Omega. \end{cases} \tag{5.4}$$

Applying estimate (5.3) to system (5.4) (with $\chi = (I-\Delta)u$ and $g = -\Delta u$), we conclude that

$$\begin{aligned} E(u,0) &\leq C\left[\int_0^{T_0}\int_\omega|\partial_t(I-\Delta)u(x,t)|^2\,dxdt + \int_0^{T_0}\int_\Omega|\Delta u(x,t)|^2\,dxdt\right] \\ &\leq C\left[\int_0^{T_0}\int_\omega|\partial_t\Delta u(x,t)|^2\,dxdt + \|(u_0,u_1)\|^2_{H^2(\Omega)\times H^1(\Omega)}\right]. \end{aligned} \tag{5.5}$$

However, by (5.5) and using the classical uniqueness-compactness argument (e.g., [15]), it follows that

$$\|(u_0,u_1)\|^2_{H^2(\Omega)\times H^1(\Omega)} \leq C\int_0^{T_0}\int_\omega |\partial_t \Delta u\,(x,t)|^2\, dxdt. \tag{5.6}$$

Therefore, combining (5.5) and (5.6), we deduce the desired estimate (5.1). This completes the proof of Theorem 5.1. □

Next, when GCC is not assumed, we have the following weaker estimate.

THEOREM 5.2. *Suppose that Ω is connected. Then, for any nonempty open subset ω of Ω, any $\beta \in (0,1)$, and the time $T > 0$ given in Theorem 4.1, the solution u of (1.1) satisfies*

$$\begin{aligned} E(u,0) \leq Ce^{C/\mu}\int_0^T\int_\omega |\partial_t u\,(x,t)|^2\, dxdt + \mu^{2\beta}\,\|(u_0,u_1)\|^2_{D(\mathcal{A})} \\ \forall\,(u_0,u_1)\in D(\mathcal{A})\setminus\{0\}, \quad \forall\,\mu>0, \end{aligned} \tag{5.7}$$

and

$$\begin{aligned} E(u,0) \leq Ce^{C/\mu}\int_0^T\int_\omega |\partial_t \Delta u\,(x,t)|^2\, dxdt + \mu^{2\beta}\,\|(u_0,u_1)\|^2_{D(\mathcal{A}^3)} \\ \forall\,(u_0,u_1)\in D(\mathcal{A}^3)\setminus\{0\}, \quad \forall\,\mu>0. \end{aligned} \tag{5.8}$$

Proof. By (4.3) in Remark 4.1, one sees that the solution u of system (1.1) satisfies

$$\int_0^T\int_\Omega |\Delta u\,(x,t)|^2\, dxdt \leq Ce^{C/\mu}\int_0^T\int_\omega |u\,(x,t)|^2\, dxdt + \mu^{2\beta}\,\|(u_0,u_1)\|^2_{\mathcal{H}} \quad \forall\,\mu>0, \tag{5.9}$$

and

$$\int_0^T\int_\Omega |u\,(x,t)|^2\, dxdt \leq Ce^{C/\mu}\int_0^T\int_\omega |u\,(x,t)|^2\, dxdt + \mu^{2\beta}\,\|(u_0,u_1)\|^2_{\mathcal{H}} \quad \forall\,\mu>0. \tag{5.10}$$

It remains to apply (5.9) (resp., (5.10)) with u replaced by $\partial_t u$ (resp., $\partial_t\Delta u$) to get the desired estimate (5.7) (resp., (5.8)) by using the following inequality:

$$E\,(u,0) \leq C\int_0^T\int_\Omega |\partial_t \Delta u\,(x,t)|^2\, dxdt,$$

which, in turn, follows from the usual energy method. □

6. Proof of Theorems 1.1 and 1.2. We begin with the proof of Theorem 1.2. Recall that the functions $v^{(j)}$ and $U^{(j)}$, defined, respectively, in (1.4) and (1.5), depend only on $\partial_t\Delta u\,(x,T-t)\cdot 1_{|\omega}$. Let

$$w^{(j)}(x,t) = \begin{cases} v^{(0)}(x,t) - u(x,T-t), & j=0, \\ v^{(j)}(x,t) - w^{(j-1)}(x,T-t), & j>0, \end{cases} \quad \text{for } (x,t)\in Q. \tag{6.1}$$

Clearly, $w^{(j)} = U^{(j)}$ in $\omega\times(0,T)$. Also, it is easy to check that $w^{(j)}$ solves

$$\begin{cases} \partial_t^2 w^{(j)} + \Delta^2 w^{(j)} - \partial_t^2\Delta w^{(j)} + \partial_t\Delta w^{(j)}\cdot 1_{|\omega} = 0 & \text{in } Q, \\ w^{(j)} = \Delta w^{(j)} = 0 & \text{on } \Sigma. \end{cases} \tag{6.2}$$

By (6.1) and the third equation in system (1.4), noticing the conservative law (1.3), we deduce that

$$\left\{\begin{array}{ll} E(w^{(j)},0)=E(w^{(j-1)},T), & j>0,\\ E(w^{(0)},0)=E(u,T)=E(u,0). & \end{array}\right. \tag{6.3}$$

First, applying the standard energy method to system (6.2), it follows that

$$E(w^{(j)},T)-E(w^{(j)},0)+\int_0^T\int_\omega\left|\partial_t\Delta w^{(j)}\,(x,t)\right|^2dxdt=0 \qquad \forall\, T>0. \tag{6.4}$$

By (5.8) in Theorem 5.2 and using a well-known perturbation argument (e.g., [16, section 5]), we conclude that the solution $w^{(j)}$ of (6.2) satisfies, for any $\mu>0$,

$$E(w^{(j)},0)\le Ce^{C/\mu}\int_0^T\int_\omega\left|\partial_t\Delta w^{(j)}(x,t)\right|^2dxdt+\mu^{2\beta}\left\|\left(w^{(j)}(\cdot,0),\ \partial_t w^{(j)}(\cdot,0)\right)\right\|^2_{D(\mathcal{A}^3)}. \tag{6.5}$$

Combining (6.4)–(6.5) and the first line in (6.3), it follows that, for any $\mu>0$,

$$E(w^{(j)},0)\le Ce^{C/\mu}\Big[E(w^{(j)},0)-E(w^{(j+1)},0)\Big]+\mu^{2\beta}\left\|\left(w^{(j)}(\cdot,0),\ \partial_t w^{(j)}(\cdot,0)\right)\right\|^2_{D(\mathcal{A}^3)}. \tag{6.6}$$

In view of the dissipation law for system (6.2), noticing again (6.1) and the third equation in system (1.4), one has

$$\left\|\left(w^{(2j+1)}(\cdot,T),\ \partial_t w^{(2j+1)}(\cdot,T)\right)\right\|_{D(\mathcal{A}^3)}\le C\left\|\left(w^{(2j)}(\cdot,T),\ \partial_t w^{(2j)}(\cdot,T)\right)\right\|_{D(\mathcal{A}^3)}. \tag{6.7}$$

Therefore, by (6.6)–(6.7) and denoting $M\overset{\triangle}{=}\sup_{j>0}\left\|\left(w^{(2j)}(\cdot,T),\ \partial_t w^{(2j)}(\cdot,T)\right)\right\|^2_{D(\mathcal{A}^3)}$, we conclude that

$$E(w^{(j)},0)\le Ce^{C/\mu}\Big[E(w^{(j)},0)-E(w^{(j+1)},0)\Big]+\mu^{2\beta}M \quad \forall\,\mu>0. \tag{6.8}$$

Now, by (6.8) and similar to Remark 4.1, one deduces that the solution $w^{(j)}$ of (6.2) satisfies

$$\frac{E\left(w^{(j)},0\right)}{M}\le C\ln^{-2\beta}\left(1+\frac{M}{E\left(w^{(j)},0\right)-E\left(w^{(j+1)},0\right)}\right). \tag{6.9}$$

Let

$$\alpha_n=\frac{E\left(w^{(n)},0\right)}{M}.$$

Then,

$$\alpha_{n+1}=\frac{E\left(w^{(n+1)},0\right)}{M}=\frac{E\left(w^{(n)},T\right)}{M}\le\alpha_n. \tag{6.10}$$

Combining (6.9) and (6.10), we obtain

$$\alpha_{n+1}\le C\ln^{-2\beta}\left(1+\frac{1}{\alpha_n-\alpha_{n+1}}\right) \qquad \forall\, n\in\mathbb{N}. \tag{6.11}$$

Similar to [14, 27], starting from (6.11), one deduces that

$$\alpha_{n+1} \leq C \ln^{-2\beta}(1+n) \qquad \forall\, n \in \mathbb{N}, \tag{6.12}$$

which gives

$$\frac{E\left(w^{(2N)}, T\right)}{M} \leq C \ln^{-2\beta}(1+2N). \tag{6.13}$$

Now it remains to compute $w^{(2N)}(\cdot, T)$. By induction, it is easy to verify that, for any $N \geq 1$,

$$w^{(2N)}(\cdot, t) = \sum_{k=1}^{N} \left[v^{(2k)}(\cdot, t) - v^{(2k-1)}(\cdot, T-t) \right] + v^{(0)}(\cdot, t) - u(\cdot, T-t).$$

Therefore,

$$w^{(2N)}(\cdot, T) = \sum_{k=0}^{N} v^{(2k)}(\cdot, T) - u_0, \quad \partial_t w^{(2N)}(\cdot, T) = \sum_{k=0}^{N} \partial_t v^{(2k)}(\cdot, T) + u_1. \tag{6.14}$$

Finally, by (6.13)–(6.14) and $J_3 < +\infty$, one arrives at the desired estimate (1.7).

Similarly, the proof of Theorem 1.1 follows from (6.3), (6.14), and Theorem 5.1. □

REFERENCES

[1] C. Bardos and M. Fink, *Mathematical foundations of the time reversal mirror*, Asymptot. Anal., 29 (2002), pp. 157–182.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Un exemple d'utilisation des notions de propagation pour le contrôle et la stabilisation des problèmes hyperboliques*, in Nonlinear Hyperbolic Equations in Applied Mathematics, Rend. Sem. Mat. Univ. Politec. Torino 1988, (Special Issue) (1989), pp. 11–32.

[3] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[4] N. Burq, *Contrôlabilité exacte de l'équation des ondes dans des ouverts peu réguliers*, Asymptot. Anal., 14 (1997), pp. 157–191.

[5] O. Dorn, *Time-Reversal and the Adjoint Method with an Application in Telecommunication*, preprint, http://www.arxiv.org/abs/math.OC/0412379, 2004.

[6] M. Fink, *Time reversal of ultrasonic fields—basic principles*, IEEE Trans. Ultrasonics Ferroelectric and Frequency Control, 39 (1992), pp. 555–556.

[7] M. Fink and C. Prada, *Acoustic time reversal mirrors*, Inverse Problems, 17 (2001), pp. R1–R38.

[8] J.-P. Fouque, J. Garnier, and A. Nachbin, *Time reversal for dispersive waves in random media*, SIAM J. Appl. Math., 64 (2004), pp. 1810–1838.

[9] X. Fu, *A weighted identity for partial differential operators of second order and its applications*, C. R. Math. Acad. Sci. Paris, 342 (2006), pp. 579–584.

[10] X. Fu, J. Yong, and X. Zhang, *Exact controllability for multidimensional semilinear hyperbolic equations*, SIAM J. Control Optim., 46 (2007), pp. 1578–1614.

[11] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 1996.

[12] B. L. G. Jonsson, M. Gustafsson, V. H. Weston, and M. V. de Hoop, *Retrofocusing of acoustic wave fields by iterated time reversal*, SIAM J. Appl. Math., 64 (2004), pp. 1954–1986.

[13] G. Lebeau and L. Robbiano, *Contrôle exacte de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[14] G. Lebeau and L. Robbiano, *Stabilisation de l'équation des ondes par le bord*, Duke Math. J., 86 (1997), pp. 465–491.

[15] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome 1: Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.

[16] K. Liu, B. Rao, and X. Zhang, *Stabilization of the wave equations with potential and indefinite damping*, J. Math. Anal. Appl., 269 (2002), pp. 747–769.

[17] L. Miller, *Escape function conditions for the observation, control, and stabilization of the wave equation*, SIAM J. Control Optim., 41 (2003), pp. 1554–1566.

[18] L. Pan, K. L. Teo, and X. Zhang, *State-observation problem for a class of semi-linear hyperbolic systems*, Chinese J. Contemp. Math., 25 (2004), pp. 163–172.

[19] G. Papanicolaou, L. Ryzhik, and K. Sølna, *Statistical stability in time reversal*, SIAM J. Appl. Math., 64 (2004), pp. 1133–1155.

[20] L. Robbiano, *Fonction de coût et contrôle des solutions des équations hyperboliques*, Asymptot. Anal., 10 (1995), pp. 95–115.

[21] G. Wang and L. Wang, *The Carleman inequality and its application to periodic optimal control governed by semilinear parabolic differential equations*, J. Optim. Theory Appl., 118 (2003), pp. 249–461.

[22] Y. Wang, *Global uniqueness and stability for an inverse plate problem*, J. Optim. Theory Appl., 132 (2007), pp. 161–173.

[23] M. Yamamoto, *Stability, reconstruction formula and regularization for an inverse source hyperbolic problem by a control method*, Inverse Problems, 11 (1995), pp. 481–496.

[24] M. Yamamoto, *Uniqueness and stability in multidimensional hyperbolic inverse problems*, J. Math. Pures Appl., 78 (1999), pp. 65–98.

[25] M. Yamamoto and X. Zhang, *Global uniqueness and stability for an inverse wave source problem for less regular data*, J. Math. Anal. Appl., 263 (2001), pp. 479–500.

[26] M. Yamamoto and X. Zhang, *Global uniqueness and stability for a class of multidimensional inverse hyperbolic problems with two unknowns*, Appl. Math. Optim., 48 (2003), pp. 211–228.

[27] X. Zhang and E. Zuazua, *Long time behavior of a coupled heat-wave system arising in fluid-structure interaction*, Arch. Ration. Mech. Anal., 184 (2007), pp. 49–120.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1557–1573

ELECTRICAL IMPEDANCE TOMOGRAPHY BY ELASTIC DEFORMATION*

H. AMMARI[†], E. BONNETIER[‡], Y. CAPDEBOSCQ[§], M. TANTER[†], AND M. FINK[†]

Abstract. This paper presents a new algorithm for conductivity imaging. Our idea is to extract more information about the conductivity distribution from data that have been enriched by coupling impedance electrical measurements to localized elastic perturbations. Using asymptotics of the fields in the presence of small volume inclusions, we relate the pointwise values of the energy density to the measured data through a nonlinear PDE. Our algorithm is based on this PDE and takes full advantage of the enriched data. We give numerical examples that illustrate the performance and the accuracy of our approach.

Key words. electrical impedance tomography, elastic perturbation, asymptotic formula, reconstruction, substitution algorithm, 0-Laplacian

AMS subject classification. 35R30

DOI. 10.1137/070686408

1. Introduction.

The electrical impedance tomography (EIT) technique has been an active research topic since the early 1980s. In EIT, one measures the boundary voltages due to multiple injection currents to reconstruct images of the conductivity distribution. This problem is known to be ill-posed [1] due to the fact that boundary voltages are not very sensitive to local changes of the conductivity distribution.

Medical imaging has been one of the important application areas of EIT. Indeed, biological tissues have different electrical properties that change with cell concentration, cellular structure, and molecular composition. Such changes of electrical properties are the manifestations of structural, functional, metabolic, and pathological conditions of tissues and thus provide valuable diagnostic information.

For practitioners, the practicality of EIT is of great interest: It is a low cost and portable technology which can be used for real time monitoring. However, it suffers from poor spatial resolution and accuracy, a well-known feature of inverse problems. This motivated us to look for a new way of incorporating more information in EIT data, without altering the cost and portability of the data acquisition, which would yet improve the resolution of the reconstructed images.

The classical image reconstruction algorithms view EIT as an optimization problem. An initial conductivity distribution is iteratively updated so as to minimize the difference between measured and computed boundary voltages. This kind of method was first introduced in EIT by Yorkey, Webster, and Tompkins [43]. Numerous variations and improvements followed, which include utilization of a priori information,

[*] Received by the editors March 26, 2007; accepted for publication (in revised form) February 27, 2008; published electronically June 6, 2008. The partial support of the ANR project EchoScan (AN-06-Blan-0089) is acknowledged.

http://www.siam.org/journals/siap/68-6/68640.html

[†] Laboratoire Ondes et Acoustique, CNRS UMR 7587, ESPCI & Paris VII, 10 rue Vauquelin, 75231 Paris Cedex 05, France (habib.ammari@polytechnique.fr, mikael.tanter@espci.fr, mathias.fink@espci.fr).

[‡] Laboratoire de Modélisation et Calcul, Université Joseph Fourier, BP 53, 38041, Grenoble Cedex 9, France (Eric.Bonnetier@imag.fr).

[§] LMV, Université de Versailles-Saint Quentin, CNRS, Versailles, France (yves.capdeboscq@uvsq.fr).

and various forms of regularization [42, 19]. This approach is quite greedy in computational time yet produces images with poor accuracy and spatial resolution.

In the 1980s, Barber and Brown [14] introduced a back-projection algorithm that was the first fast and efficient algorithm for EIT, although it provides images with very low resolution. Since this algorithm is inspired from computed tomography, it can be viewed as a generalized Radon transform method [39].

A third technique is dynamical electrical impedance imaging, developed by the Rensselaer impedance tomography group [24] to produce images of changes in conductivity due to cardiac or respiratory functions. Its main idea consists in viewing the conductivity as the sum of a static term (the background conductivity of the human body) plus a perturbation (the change of conductivity caused by respiration or by heart beats). The mathematical problem here is to visualize the perturbation term by an EIT system. Although this algorithm provides accurate images if the initial guess of the background conductivity is good, its resolution does not completely satisfy practitioners especially when screening for breast cancer (see also [25]).

Recently, a commercial system called TransScan TS2000 (TransScan Medical, Ltd, Migdal Ha'Emek, Israel) was released for adjunctive clinical uses with X-ray mammography in the diagnosis of breast cancer. Interestingly, the TransScan system is similar to the frontal plane impedance camera that initiated EIT research early in 1978. The mathematical model of the TransScan can be viewed as a realistic or practical version of the general EIT system, so any theory developed for this model can be applied to other areas in EIT, especially to detection of anomalies. In the TransScan, a patient holds a metallic cylindrical reference electrode, through which a constant voltage of 1 to 2.5 V, with frequencies spanning 100 Hz–100 KHz, is applied. A scanning probe with a planar array of electrodes, kept at ground potential, is placed on the breast. The voltage difference between the hand and the probe induces a current flow through the breast, from which information about the impedance distribution in the breast can be extracted.

Using a simplified dipole method, Assenheimer et al. [13] and Scholz [40] gave a physical interpretation of the white spots in TransScan images.

More recently, taking advantage of the smallness of the anomalies to be detected, Ammari et al. [11] analyzed trans-admittance data for the detection of breast cancer using the TransScan system. Their model assumes that breast tissues can be considered homogeneous, at least near the surface, where the planar array of electrodes is attached, and that the lesion to be detected is located near the surface. In [11], the authors developed better ways of interpreting TransScan images which improve accuracy. They also derived a multifrequency approach to handle the case where the background conductivity is inhomogeneous and not known a priori.

This latter work relies on asymptotic expansions of the fields when the medium contains inclusions of small volume, a technique that has proven useful in many other contexts. Such asymptotics have been investigated in the case of the conduction equation [23, 22, 17, 6, 20, 18, 21], the operator of elasticity [5, 10], and the Helmholtz equation or the Maxwell system [44, 12, 9]. See the books [7, 8] and their lists of references. The remarkable feature of this technique is that it allows a stable and accurate reconstruction of the location and of the geometric features of the inclusions, even for moderately noisy data.

Since all the present EIT technologies are practically applicable only in feature extraction of anomalies, improving EIT calls for innovative measurement techniques that incorporate structural information. A very promising direction of research is the recent magnetic resonance imaging technique, called current density imaging, which measures the internal current density distribution. See the breakthrough work by Seo and his group [36, 37, 32]. However, this technique has a number of disadvantages, among which are the lack of portability and a potentially long imaging time. Moreover, it uses an expensive magnetic resonance imaging scanner.

The aim of this paper is to propose another mathematical direction for future EIT research in view of biomedical applications, without eliminating the most important merits of EIT (real time imaging, low cost, and portability). Our method is based on the simultaneous measurement of an electric current and of acoustic vibrations induced by ultrasound waves. Its intrinsic resolution depends on the size of the focal spot of the acoustic perturbation, and thus our method should provide high resolution images.

Let us now formulate our problem. We first recall that, in mathematical terms, EIT consists in recovering the conductivity map of a d dimensional body Ω, where d is the dimension of the ambient space, from measuring the voltage response to one or several currents applied on the boundary. In practice, a set of electrodes is attached to the body. One or several currents ϕ_i, $1 \leq i \leq I$, are applied to one or several electrodes, and the corresponding voltage potentials f_i, $1 \leq i \leq I$, are recorded on the others. Denoting by $\gamma(x)$ the unknown conductivity, the voltage potential u_i solves the conduction problem

$$(1.1) \qquad \left\{ \begin{aligned} \nabla_x \cdot (\gamma(x) \nabla_x u_i) &= 0 \quad \text{in } \Omega, \\ \gamma(x) \frac{\partial u_i}{\partial n} &= \phi_i \quad \text{on } \partial\Omega. \end{aligned} \right.$$

The problem of impedance tomography is the inverse problem of recovering the coefficients γ of the elliptic conduction PDE, knowing one or more current-to-voltage pairs $(\phi_i,\, f_i := u_i|_{\partial\Omega})$.

The core idea of our approach is to extract more information about the conductivity from data that has been enriched by coupling the electric measurements to localized elastic perturbations. More precisely, we propose to perturb the medium during the electric measurements by focusing ultrasonic waves on regions of small diameter inside the body. Using a simple model for the mechanical effects of the ultrasound waves, we show that the difference between the measurements in the unperturbed and perturbed configurations is asymptotically equal to the pointwise value of the energy density at the center of the perturbed zone. In practice, the ultrasounds impact a spherical or ellipsoidal zone of a few millimeters in diameter. The perturbation should thus be sensitive to conductivity variations at the millimeter scale, which is the precision required for breast cancer diagnosis.

By scanning the interior of the body with ultrasound waves, given an applied current ϕ_i, we obtain data from which we can compute $\mathcal{S}_i(x) := \gamma(x)|\nabla u_i(x)|^2$ in an interior subregion of Ω. The new inverse problem is now to reconstruct γ knowing $\mathcal{S}_i$ for $i = 1, \ldots, I$.

The goal of this work is threefold: First, we show that taking measurements while perturbing the medium with ultrasound waves is asymptotically equivalent to measuring $\mathcal{S}_i$. To this end, we consider the zone ω deformed by the ultrasound wave as a small volume perturbation of the background potential γ. We then relate the difference between the perturbed and unperturbed potentials on the boundary to the conductivity at the center of ω asymptotically as $|\omega| \to 0$. This is our main idea: the ultrasound waves create localized perturbations that allow us, using the method of asymptotic expansions of small volume inclusions, to probe *within* the medium.

Second, noting that the potential u_i satisfies the nonlinear PDE (the 0-Laplacian)

$$\left\{ \begin{aligned} \nabla_x \cdot \left(\frac{\mathcal{S}_i(x)}{|\nabla u_i|^2} \nabla u_i \right) &= 0 \quad \text{in } \Omega, \\ \frac{\mathcal{S}_i(x)}{|\nabla u_i|^2} \frac{\partial u_i}{\partial n} &= \phi_i \quad \text{on } \partial\Omega, \end{aligned} \right. \tag{1.2}$$

we propose a numerical method to compute solutions $\tilde{u}_i$ to (1.2) and then an approximate conductivity $\tilde{\gamma} = \mathcal{S}_i/|\nabla \tilde{u}_i|^2$, using two currents (i.e., $I = 2$). Recall that an appropriate choice of ϕ_i ensures that $\nabla u_i \neq 0$ for all $x \in \Omega$. See [2, 3, 41, 26].

Third, our algorithm, as the one originally developed in [33] for current density imaging, requires data measured with two boundary currents for which the flux densities (the gradient of the voltage potentials) are locally orthogonal (or try to be). We show numerically that this algorithm is able to capture details of the conductivity map up to the precision of the underlying finite element mesh and thus proves very effective.

The paper is organized as follows. In the next section, we describe the physical model and the collection of experimental data on the boundary. Section 3 explains how, given a current ϕ_i, the values of $\mathcal{S}_i(x)$ can be approximated using this data. In section 4, we describe the numerical method for reconstruction of γ using two applied currents. Numerical examples that illustrate the performance and the accuracy of this method are presented in that section. The paper ends with a short discussion.

2. Impedance tomography perturbed by ultrasound waves.

2.1. Description of the experiment. The goal of the experiment is to obtain an impedance map inside a solid with millimetric precision.

An object (a domain Ω) is electrically probed: one or several currents are imposed on the surface and the induced potentials are measured on the boundary (see Figure 1). At the same time, a spherical region of a few millimeters in the interior of Ω is mechanically excited by focused acoustic waves.

The measurements are made as the focus of the ultrasounds scans the entire domain. Several sets of measurements can be obtained by varying the ultrasound waves amplitudes and the applied currents. Several teams have been able to obtain ultrasonic waves focusing in very small regions deep inside the tissues [16, 31]. The support of the focal spot is better represented by an ellipsoid, but the locus of the most intense area can be, in a first approximation, represented by a sphere. The experiment is successful if, for each focal point, a difference in the boundary voltage potential can be measured between the potential corresponding to an unperturbed medium and the potential corresponding to the perturbed one.

2.2. Physical modeling of the effect of ultrasonic waves on the conductivity. We chose to model the effect of the pressure wave in the simplest way. For what follows, there are two crucial points. The first one is that the local conductivity is affected by the pressure wave. The second one is that, at least for acoustic waves of moderate amplitude, the conductivity perturbation depends continuously on the amplitude of the wave—and therefore, in a first approximation, linearly. The fact that acoustic waves affect the conductivity has been known for a long time [35]. In the context of biomedical imaging, the idea of exploiting this property dates from the early 1970s [27]. For focused waves, this is much more recent. In [30, 31] it is established experimentally for moderate and high intensity waves. The experiments show that only the *local* conductivity is affected.

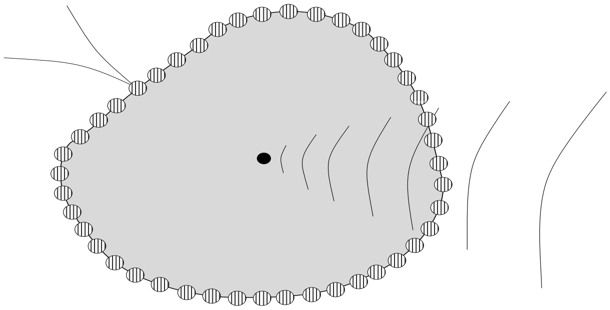


FIG. 1. *The experimental setup.*

Below is an attempt to justify these experimental findings. Within each (small) spherical volume, the conductivity is assumed to be constant per volume unit. At a point $x \in \Omega$, within a ball B of volume V_B, the electric conductivity γ is defined in terms of a density ρ as

$$\gamma(x) = \rho(x) V_B.$$

The ultrasonic waves induce a small elastic deformation of the sphere B. If this deformation is isotropic, the material points of B occupy a volume V_B^p in the perturbed configuration, which at first order is equal to

$$V_B^p = V_B \left(1 + 3\frac{\Delta r}{r}\right),$$

where r is the radius of the ball B and Δr is the variation of the radius due to the elastic perturbation. As Δr is proportional to the amplitude of the ultrasonic wave, we obtain a proportional change of the deformation. Using two different ultrasonic waves with different amplitudes but with the same focal spot, it is therefore easy to compute the ratio V_B^p/V_B for a given perturbation. We are merely pointing out that if $f(x) = a(1 + x * b)$, one can evaluate a and b in terms of $f(x_1)$, $f(x_2)$, x_1, and x_2. As a consequence, the perturbed electrical conductivity γ^p satisfies

$$\forall\, x \in \Omega, \quad \gamma^p(x) = \rho(x) V_B^p \;=\; \gamma(x)\nu(x), \tag{2.1}$$

where $\nu(x) = V_B^p/V_B$ is a known function.

2.3. Mathematical modeling of the effect of ultrasonic waves on the conductivity. We denote by u the voltage potential induced by a current ϕ, in the absence of ultrasonic perturbations. It is given by

$$\left\{ \begin{array}{r} \nabla_x \cdot (\gamma(x) \nabla_x u) = 0 \quad \text{in } \Omega, \\ \gamma(x) \frac{\partial u}{\partial n} = \phi \quad \text{on } \partial\Omega, \end{array} \right. \tag{2.2}$$

with the normalization condition $\int_{\partial\Omega} u = 0$. We assume that the conductivity γ is bounded above and below by positive constants

$$0 < c < \gamma(x) < C < +\infty \quad \text{a.e. } x \in \Omega.$$

Further, we suppose that the conductivity γ is known close to the boundary of the domain so that ultrasonic probing is limited to interior points x such that

$$\text{dist}(x, \partial\Omega) \geq d_0,$$

where d_0 is very large compared to the radius of the focal spot of the ultrasonic perturbation. We denote the corresponding open set Ω_1. We denote by $u_\omega(x)$, $x \in \Omega$, the voltage potential induced by a current ϕ, in the presence of ultrasonic perturbations localized in a domain ω of volume $|\omega|$. The voltage potential u_ω is a solution to

$$\left\{ \begin{array}{c} \nabla_x \cdot (\gamma_\omega(x) \nabla_x u_\omega(x)) = 0 \quad \text{in } \Omega, \\ \gamma(x) \frac{\partial u_\omega}{\partial n} = \phi \quad \text{on } \partial\Omega, \end{array} \right. \tag{2.3}$$

with the notation

$$\gamma_\omega(x) = \gamma(x) \left[1 + \mathbf{1}_\omega(x) \left(\nu(x) - 1 \right) \right],$$

where $\mathbf{1}_\omega$ is the characteristic function of the domain ω. In the next section, we show how comparing u_ω and u on $\partial\Omega$ provides information about the conductivity.

3. Asymptotic recovery of the conductivity. As the zone deformed by the ultrasound wave is small, we can view it as a small volume perturbation of the background conductivity γ, and we seek an asymptotic expansion of the boundary values of $u_\omega - u$.

For $x \in \mathbb{R}^d$, we note that $x = (x_1, \dots, x_d)$. For each $i = 1, \dots, d$, let ζ_ω^i be the solution to

$$\left\{ \begin{array}{l} \nabla_x \cdot \left(\gamma_\omega(x) \nabla_x \zeta_\omega^i \right) = \nabla_x \cdot (\gamma(x) \nabla_x x_i) \quad \text{in } \Omega, \\ \gamma(x) \frac{\partial \zeta_\omega^i}{\partial n} = \gamma(x) \frac{\partial x_i}{\partial n} \quad \text{on } \partial\Omega, \quad \text{with } \int_\Omega \zeta_\omega^i = 0. \end{array} \right.$$

Corresponding to ζ_ω^i, we define $\zeta^i = x_i - c_i$, where c_i is a constant, in the unperturbed case.

The following proposition is a variant of a compactness result proved in [20]. In contrast to previous work, the proof we give here requires only boundedness of the conductivity γ_ω.

PROPOSITION 3.1. *Consider a sequence of sets* $\omega \subset\subset \Omega$, *such that* $\frac{1}{|\omega|} \mathbf{1}_\omega$ *converges in the sense of measures to a probability measure* $d\mu$ *as* $|\omega|$ *tends to zero. Then, the correctors* $\frac{1}{|\omega|} \mathbf{1}_\omega \frac{\partial \zeta_\omega^i}{\partial x_j}$ *converge in the sense of measures to* M, *where* $M \in L^2(\Omega, d\mu)$ *is a matrix-valued function.*

Furthermore, the correctors $\left(\zeta_\omega^i \right)$ *satisfy*

$$\|\nabla(\zeta_\omega^i - \zeta^i)\|_{L^2(\Omega)^d} \le C|\omega|^{1/2} \text{ and } \|\zeta_\omega^i - \zeta^i\|_{L^2(\Omega)} \le C|\omega|^{\frac{1}{2}+\kappa},$$

where the constants $\kappa > 0$ *and* $C > 0$ *depend only on* Ω_1, $\sup_\Omega |\gamma_\omega|$, *and* $\inf_\Omega |\gamma_\omega|$.

Proof. The bounds on $\nabla \left(\zeta_\omega^i - \zeta^i \right)$ and $\left(\zeta_\omega^i - \zeta^i \right)$ are a direct consequence of Lemma A.1 if we remark that, inside the domain Ω, $\zeta_\omega^i - \zeta^i$ satisfies

$$\nabla_x \cdot \left(\gamma_\omega(x) \nabla_x \left(\zeta_\omega^i - \zeta^i \right) \right) = -\nabla_x \cdot \left(\mathbf{1}_\omega \left(\gamma_\omega - \gamma \right) \nabla_x x_i \right) \quad \text{in } \Omega.$$

As for the existence of a limit (and its additional properties) we refer the reader to [20]. □

One of the key elements of our method is the following representation formula.

PROPOSITION 3.2. *Assume that* $u \in W^{2,\infty}(\omega)$. *Then,*

$$\int_{\partial\Omega} (u_\omega - u) \, \phi \, d\sigma = |\omega| \int_\Omega (\gamma_\omega(x) - \gamma(x)) \, M_\omega \nabla u \cdot \nabla u \, dx + O(|\omega|^{1+\kappa}).$$

The exponent κ depends only on Ω_1, $\sup_\Omega |\gamma_\omega|$, and $\inf_\Omega |\gamma_\omega|$. The remainder term has the form

$$\left|O(|\omega|^{1+\kappa})\right| \leq C\,|\omega|^{1+\kappa} \|\nabla u\|_{L^\infty(\omega)^d} \|\nabla^2 u\|_{L^\infty(\omega)^{d\times d}},$$

where C depends only on Ω_1, $\sup_\Omega |\gamma_\omega|$, and $\inf_\Omega |\gamma_\omega|$. Finally, the matrix-valued function M_ω is given by

$$\left(M_\omega\right)_{ij}(x) = \frac{1}{|\omega|}\mathbf{1}_\omega(x)\frac{\partial}{\partial x_j}\zeta^i_\omega(x) \qquad a.e.\ x \in \Omega_1.$$

This is, globally, not a new result. This representation formula and the proof presented were already obtained by Capdeboscq and Vogelius in [20, Theorem 1]. Compared to Theorem 1 in [20], the regularity required on u is investigated more in depth. Note that, globally, u satisfies the minimal requirement $u \in H^1(\Omega)$. Additional regularity on u is required only within ω (in particular, the quality of the representation formula is not affected). We also note that if ω is a disk in the two-dimensional case, then (see, for instance, [22])

$$M_\omega = \frac{1}{|\omega|}\mathbf{1}_\omega(x)\frac{\nu-1}{\nu+1}I_2,$$

where I_2 is the unit matrix. The following corollary holds.

Corollary 3.3. *Assume that the dimension $d = 2$, that the perturbed area ω is a disk centered at z, and that $u \in W^{2,\infty}(\omega)$. Then, we have*

$$\begin{aligned}\int_{\partial\Omega}(u_\omega - u)\phi\,d\sigma &= \int_\omega \gamma(x)\frac{\left(\nu(x)-1\right)^2}{\nu(x)+1}\nabla u\cdot\nabla u\,dx + O(|\omega|^{1+\kappa})\\ &= |\nabla u(z)|^2\int_\omega \gamma(x)\frac{\left(\nu(x)-1\right)^2}{\nu(x)+1}\,dx + O(|\omega|^{1+\kappa}).\end{aligned}$$

Therefore, if γ is $\mathcal{C}^{0,2\alpha}(\omega)$, with $0 \leq \alpha \leq \kappa \leq \frac{1}{2}$, we have

$$\gamma(z)\,|\nabla u(z)|^2 = \mathcal{S}(z) + O(|\omega|^\alpha) \qquad (or\ o(1)\ if\ \alpha = 0), \tag{3.1}$$

where the function $\mathcal{S}(z)$ is defined by

$$\mathcal{S}(z) = \left(\int_\omega \frac{\left(\nu(x)-1\right)^2}{\nu(x)+1}\,dx\right)^{-1}\int_{\partial\Omega}(u_\omega - u)\phi\,d\sigma. \tag{3.2}$$

We emphasize that $\mathcal{S}(z)$ represents a known function, as the second term on the right-hand side of (3.2) is exactly the measured data.

Proof of Proposition 3.2. This proof follows the proof of Lemma 2 in [20]. First, in (3.3) we establish that the boundary data is related to an integral on ω, involving $(\gamma_1 - \gamma_0)\nabla u_\omega$. Then, in (3.7) we show that almost everywhere $(\gamma_1 - \gamma_0)\nabla u_\omega$ can be estimated by a quantity involving ∇u and the polarization tensor. Finally, in (3.8) we show that by approximation this representation formula leads to the desired result.

Integrating (2.2) against U_ω, the solution of (2.3), and vice-versa, we obtain after an integration by parts

$$\int_{\partial\Omega} u_\omega\phi\,d\sigma = \int_\Omega \gamma\nabla u_\omega\cdot\nabla u\,dx \text{ and } \int_{\partial\Omega} u\phi\,d\sigma = \int_\Omega \gamma_\omega\nabla u_\omega\cdot\nabla u\,dx.$$

Consequently,

$$\int_{\partial\Omega} (u_\omega - u)\, \phi \, d\sigma = \int_\Omega (\gamma - \gamma_\omega)\, \nabla u_\omega \cdot \nabla u \, dx \\ = \int_\omega (\gamma - \gamma_\omega)\, \nabla u_\omega \cdot \nabla u \, dx. \tag{3.3}$$

Notice that $u_\omega - u$ satisfies a homogeneous Neumann boundary condition and verifies

$$\nabla \cdot (\gamma_\omega \nabla (u_\omega - u)) = -\nabla \cdot (\mathbf{1}_\omega (\gamma_\omega - \gamma) \nabla u) \text{ in } \Omega.$$

Since $u \in W^{1,\infty}(\omega)$, we can again invoke Lemma A.1 to obtain that

$$\|\nabla(u_\omega - u)\|_{L^2(\Omega)^d} \leq C|\omega|^{1/2} \|\nabla u\|_{L^\infty(\omega)^d} \text{ and } \|u_\omega - u\|_{L^2(\Omega)} \leq C|\omega|^{1/2+\kappa} \|\nabla u\|_{L^\infty(\omega)^d},$$

where the constants C, κ depend only on Ω_1, $\sup_\Omega |\gamma_\omega|$, and $\inf_\Omega |\gamma_\omega|$. For all $\theta \in W^{1,\infty}(\Omega)$, we now compute

$$\begin{aligned} \int_\Omega \gamma_\omega \nabla (u_\omega - u) \cdot \nabla \zeta_\omega^i \, \theta \, dx &= \int_\Omega \gamma_\omega \nabla \left((u_\omega - u)\theta \right) \cdot \nabla \zeta_\omega^i \, dx \\ &\quad - \int_\Omega \gamma_\omega (u_\omega - u) \nabla \theta \cdot \nabla \zeta_\omega^i \, dx \\ &= \int_\Omega \gamma \nabla \left((u_\omega - u)\theta \right) \nabla \zeta^i \, dx + r_1 \\ &= \int_\Omega \gamma \nabla (u_\omega - u) \nabla \zeta^i \, \theta \, dx + r_2. \end{aligned} \tag{3.4}$$

The remainder term is given by

$$\begin{aligned} r_2 &= \int_\Omega \gamma (u_\omega - u) \nabla \theta \nabla \zeta^i \, dx - \int_\Omega \gamma_\omega (u_\omega - u) \nabla \theta \nabla \zeta_\omega^i \, dx \\ &= O(\|u_\omega - u\|_{L^2(\Omega)} \|\nabla \zeta - \nabla \zeta_\omega\|_{L^2(\Omega)^d}) \\ &= O(|\omega|^{1+\kappa}). \end{aligned}$$

Here, by $O(|\omega|^{1+\kappa})$ we denote a quantity that is bounded by $C\|\nabla u\|_{L^\infty(\omega)^d} \|\nabla \theta\|_{L^\infty(\Omega)^d} \cdot |\omega|^{1+\kappa}$, where C depends only on Ω_1, $\sup_\Omega |\gamma_\omega|$, and $\inf_\Omega |\gamma_\omega|$.

We shall consider both terms of identity (3.4) independently. On one hand, we have

$$\begin{aligned} \int_\Omega \gamma_\omega \nabla (u_\omega - u) \cdot \nabla \zeta_\omega^i \, \theta \, dx &= \int_\Omega \gamma_\omega \nabla (u_\omega - u) \cdot \nabla \left(\zeta_\omega^i \theta \right) dx \\ &\quad - \int_\Omega \gamma_\omega \nabla (u_\omega - u) \cdot \nabla \theta \, \zeta_\omega^i \, dx \\ &= \int_\omega (\gamma - \gamma_\omega) \nabla u \cdot \nabla \left(\zeta_\omega^i \theta \right) dx \\ &\quad - \int_\Omega \gamma_\omega \nabla (u_\omega - u) \cdot \nabla \theta \, \zeta^i \, dx + O\left(|\omega|^{1+\kappa} \right) \\ &= \int_\omega (\gamma - \gamma_\omega) \nabla u \cdot \nabla \zeta_\omega^i \, \theta \, dx \\ &\quad + \int_\Omega (\gamma \nabla u - \gamma_\omega \nabla u_\omega) \cdot \nabla \theta \, \zeta^i \, dx + O\left(|\omega|^{1+\kappa} \right). \end{aligned} \tag{3.5}$$

On the other hand, we have

$$\begin{aligned}\int_\Omega \gamma\nabla(u_\omega - u)\cdot\nabla\zeta^i\,\theta\,dx &= \int_\Omega \gamma\nabla(u_\omega - u)\cdot\nabla\left(\zeta^i\theta\right)dx\\ &\quad - \int_\Omega \gamma\nabla(u_\omega - u)\cdot\nabla\theta\,\zeta^i\,dx\\ &= \int_\omega (\gamma-\gamma_\omega)\nabla u_\omega\cdot\nabla\zeta^i\,\theta\,dx\\ &\quad + \int_\Omega (\gamma\nabla u - \gamma_\omega\nabla u_\omega)\cdot\nabla\theta\,\zeta^i\,dx.\end{aligned}\tag{3.6}$$

Inserting identities (3.5) and (3.6) into (3.4), we have obtained that, for all $i = 1,\dots,d$,

$$\int_\omega (\gamma-\gamma_\omega)\frac{\partial u_\omega}{\partial x_i}\,\theta\,dx = \sum_{j=1}^d \int_\omega (\gamma-\gamma_\omega)\,\frac{\partial u}{\partial x_j}\cdot\frac{\partial}{\partial x_j}\zeta^i_\omega\,\theta\,dx + O(|\omega|^{1+\kappa}),\tag{3.7}$$

with

$$\left|O(|\omega|^{1+\kappa})\right| \le C\,|\omega|^{1+\kappa}\|\nabla u\|_{L^\infty(\omega)^d}\|\nabla\theta\|_{L^\infty(\Omega)^d},$$

where C is a constant that depends only on Ω_1, $\sup_\Omega|\gamma_\omega|$, and $\inf_\Omega|\gamma_\omega|$. Let us now conclude the proof of Proposition 3.2. For each $i = 1,\dots,d$, choose $\theta_i = \frac{\partial}{\partial x_i}u * \eta_\epsilon$ in $\omega_\epsilon = \{x \in \omega \text{ s.t. } \operatorname{dist}(x,\partial\omega) > \epsilon\}$, where η is the standard mollifier. Let θ_ϵ be defined as

$$\theta_\epsilon = \left(\frac{\partial}{\partial x_1}u*\eta_\epsilon,\dots,\frac{\partial}{\partial x_d}u*\eta_\epsilon\right).$$

Using for each $i = 1,\dots,d$ the test function θ_i in (3.7) and summing over i, we obtain

$$\int_\omega (\gamma-\gamma_\omega)\nabla u_\omega\cdot\theta_\epsilon\,dx = |\omega|\int_\omega (\gamma-\gamma_\omega)\nabla u\cdot(M_\omega\theta_\epsilon)\,dx + O(|\omega|^{1+\kappa}),\tag{3.8}$$

where

$$O(|\omega|^{1+\kappa}) \le C\|\nabla u\|_{L^\infty(\omega)^d}\|\nabla\theta_\epsilon\|_{L^\infty(\omega)^d} \le C\|\nabla u\|_{L^\infty(\omega)^d}\|\nabla^2 u\|_{L^\infty(\omega)^d}.$$

By passing to the limit in ϵ and using (3.3), the proof of the proposition is complete. □

4. Reconstruction using the 0-Laplacian formulation. In view of deriving an approximation for the conductivity γ inside Ω_1, we introduce the following equation:

$$\begin{cases}\nabla\cdot\left(\dfrac{\mathcal{S}(x)}{|\nabla u|^2}\nabla u\right) = 0 & \text{in } \Omega,\\[2ex] \dfrac{\mathcal{S}(x)}{|\nabla u|^2}\dfrac{\partial u}{\partial n} = \phi & \text{on } \partial\Omega.\end{cases}\tag{4.1}$$

We emphasize that $\mathcal{S}$ is a known function, constructed from the measured data (3.2). Consequently, all the parameters entering (4.1) are known.

Our approach uses measurements $\mathcal{S}_1$ and $\mathcal{S}_2$ obtained using two distinct currents, ϕ_1 and ϕ_2. We choose this pair of current patterns to have $\nabla u_1 \times \nabla u_2 \neq 0$ for all $x \in \Omega$, where u_i, $i = 1, 2$, is the solution to (1.1). See [41, 26] for numerical evidence of the possibility of such a choice and [4] for a rigorous proof. The key question of course is whether such a choice of currents can be made simply by imposing appropriate boundary voltages, independently of the unknown conductivity. This issue will be discussed at length in a future work [19].

We start from an initial guess for the conductivity γ and solve the corresponding Dirichlet conductivity problem

$$\begin{cases} \nabla \cdot (\gamma \nabla u_0) = 0 & \text{in } \Omega, \\ u_0 = \psi & \text{on } \partial\Omega. \end{cases}$$

The data ψ is the Dirichlet data measured as a response to the current ϕ (say, $\phi = \phi_1$) in absence of elastic deformation.

The discrepancy between the data and our guessed solution is

$$\epsilon_0 := \frac{\mathcal{S}(x)}{|\nabla u_0|^2} - \gamma. \tag{4.2}$$

We then introduce a corrector, u_c, computed as the solution to

$$\begin{cases} \nabla \cdot (\gamma \nabla u_c) = -\nabla \cdot (\varepsilon_0 \nabla u_0) & \text{in } \Omega, \\ u_c = 0 & \text{on } \partial\Omega \end{cases}$$

and update the conductivity

$$\gamma := \frac{\mathcal{S}(x) - 2\gamma \nabla u_c \cdot \nabla u_0}{|\nabla u_0|^2}.$$

We iteratively update the conductivity, alternating directions (i.e., with $\phi = \phi_2$).

To study the efficiency of this approach, we have tested this method on various problems and domains, using the PDE solver FreeFem++ [28]. We present here one such test. The domain Ω is a disk of radius 8 centered at the origin, which contains three inclusions, an ellipse, an L-shaped domain, and a triangle, so as to image a convex object, a nonconvex object, and an object with a smooth boundary.

The background conductivity is equal to 0.5; the conductivity takes the values 2 in the triangle, 0.75 in the ellipse, and 2.55 in the L-shaped domain (see Figure 2). We purposely chose values corresponding to small and large contrasts with the background. Note that our approach is perturbative; thus the smaller the contrast, the easier the detection. If the contrast is small, both a small volume fraction approximation and a small amplitude approximation are valid: then the accuracy of a first order approximation such as the one performed here is increased. The choice of a significant contrast was not made to highlight the objects but rather to make the reconstruction more challenging.

Figure 3 shows the result of the reconstruction when perfect measures (with "infinite" precision) are available. In that case, the size of the spot is infinitesimal, that is, at least as small as the mesh size. We use two different boundary potentials, $\psi = x/|x|$ and $\psi = y/|y|$. The initial guess is depicted on the left: it is equal to 1 inside the disk of radius 6 centered at the origin, and it is equal to the supposedly

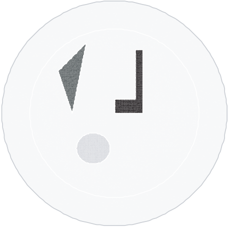


FIG. 2. *Conductivity distribution.*

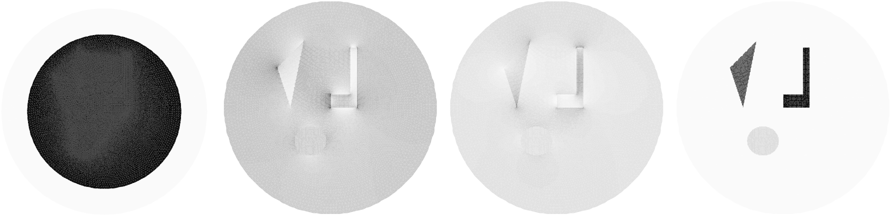


FIG. 3. *Reconstruction test. From left to right, the initial guess, the collected data* $\mathcal{S}$ *(*$x/|x|$ *and* $y/|y|$*), and the reconstructed conductivity.*

known conductivity $\gamma = 0.5$ near the boundary (outside the disk of radius 6). The two central pictures represent the collected data, $\mathcal{S}(x)$ for $\psi = x/|x|$ on the left and $\mathcal{S}(x)$ for $\psi = y/|y|$ on the right. Given the values of the contrast, we remark that although one can "see" the triangle and the L-shaped inclusions on these plots, the circle is hardly noticeable. On the far right, the reconstructed conductivity is represented: it perfectly matches the target.

In Figure 4, the error is represented as a function of the number of iterations. The dotted curve is a plot of the (intrinsic) error estimator $\max_{x\in\Omega} \epsilon_n(x)$, given by (4.2). The curve with diamond symbols depicts the L^1-norm $a_1(n)$ of the true error between the reconstructed conductivity $\tilde{\gamma}$ and the original one:

$$a_1(n) := \int_\Omega |\tilde{\gamma} - \gamma|\, dx.$$

Note that the abscissa is represented in logarithmic scale; thus the convergence seems exponential.

The other two curves correspond to the same computations, but four directions are used instead of two: $x/|x|$, $y/|y|$, $(x+y)/|x+y|$, $(x-y)/|x-y|$. Note that this does not require more measurements because of the linear dependence on the boundary condition; it is merely a change in the algorithm. The same level of error for ϵ_n is reached in 45 iterations instead of 222. The same experiment, with a contrast 5 times smaller, converges in less than ten iterations.

We also considered imperfect data. In Figure 5 we follow the same procedure but now assume that the data was measured at the nodes of a regular mesh on the disk, with 50, 100, 200, and 400 boundary points. To give an idea of the scale of the mesh compared to the objects, the projections of the conductivity that we wish to recover are represented in the left column. The two central columns depict the collected data. The column on the far right shows the obtained reconstructions.

To accelerate the computations, we used the four direction variant of the algorithm. The error as a function of the iterations is represented in Figure 6. The curves

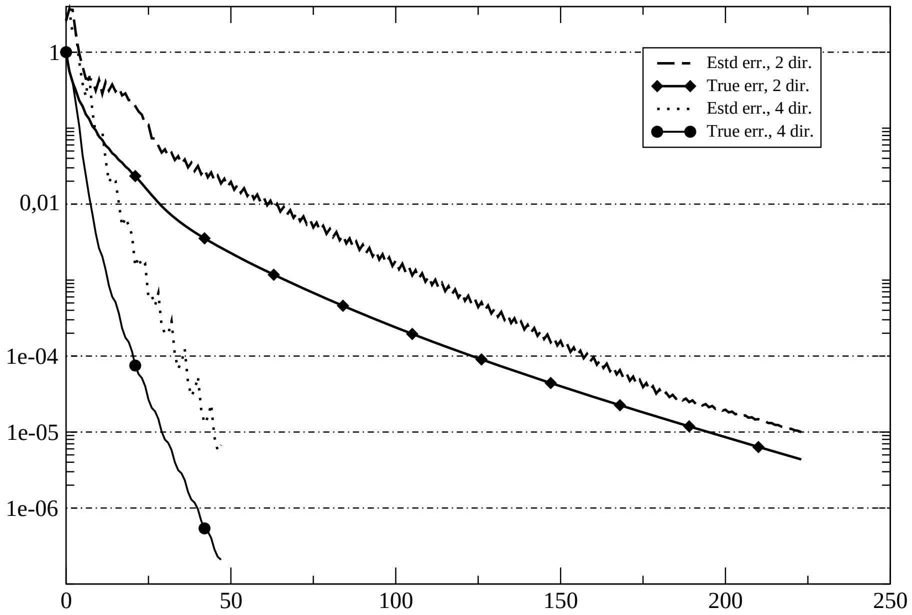


FIG. 4. *Convergence results. The curve labeled "Estd err., 2 dir." (resp., "True err., 2 dir.") corresponds to the estimated error (resp., true error) when two directions are used. The curves "Estd err., 4 dir." and "True err., 4 dir." are the errors computed when four directions are used.*

with symbols represent the L^1-norm of the true error, whereas the dashed line is the L^2-norm of the estimated error ϵ for the most precise mesh. Although the estimated error does not decrease noticeably, the true error does. The reconstructed image is obtained after ten iterations and does not change noticeably henceforth. This calls for further refinements of the algorithm, such as adapted stepsizes and adapted meshes. Improvements are indeed possible and will be the subject of a future publication [19]. As it stands, the algorithm already provides reconstructions comparable in accuracy to those of projected conductivity. Naturally, the sharp corners are easily localized, but the smooth elliptic shape is also accurately reconstructed, even at the coarsest scale.

5. Concluding remarks. We have proposed a new technique for conductivity imaging, which consists in perturbing the medium during the electric measurements, by focusing ultrasonic waves on regions of small diameter inside the body. We derived an approximation of the conductivity using small volume asymptotics and obtained a nonlinear PDE for the potential, in terms of the measured data. Based on this PDE, we proposed a new algorithm for the reconstruction of the conductivity distribution which proves remarkably accurate.

Motivated by the practical limitations of EIT, we intend to pursue the present investigation in the following directions:

(i) Study the reconstruction capabilities of this method when only partial data, measured on a small portion Γ of the boundary, is available.

(ii) Study the dependence of the algorithm on the global geometry of Ω.

(iii) Study the sensitivity of the method to limitations on the intensities of the applied voltages, as electrical safety regulations limit the amount of the total current that patients can sustain.

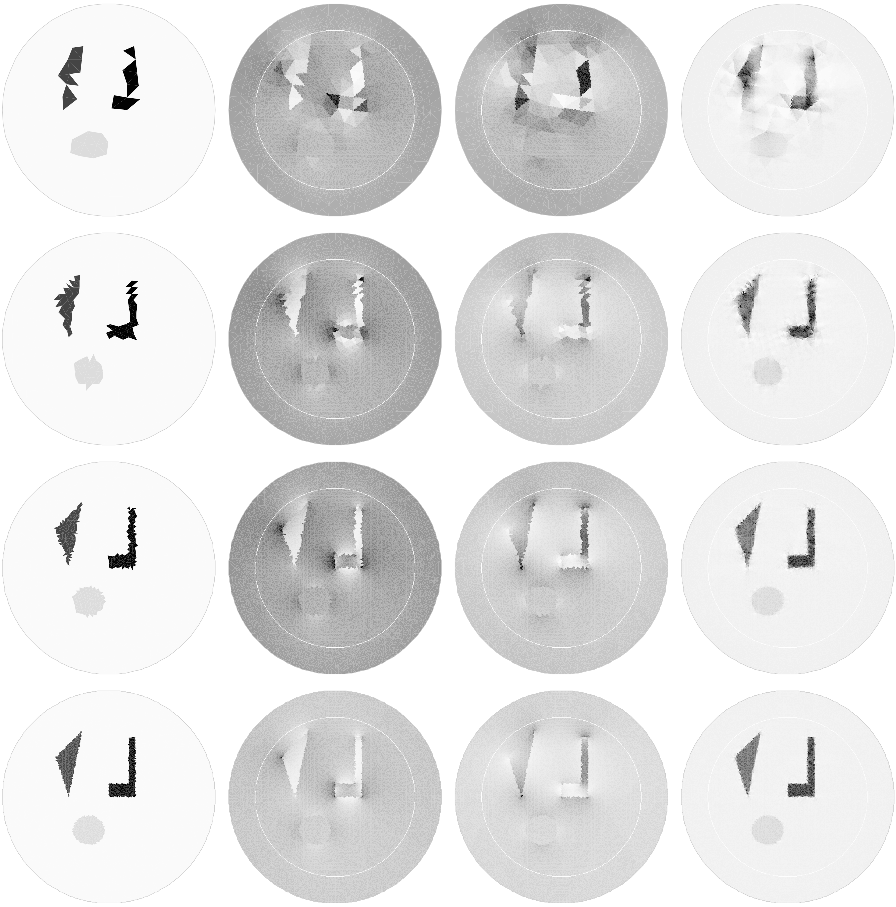


FIG. 5. *Reconstruction tests. From top to bottom, using a regular mesh with* 50*,* 100*,* 200*, and* 400 *boundary points. From left to right, the initial guess, the collected data $\mathcal{S}$ (for $x/|x|$ and $y/|y|$), and the reconstructed conductivity.*

Moreover, we also intend to address some of the mathematical questions raised by this imaging approach, among which are the uniqueness for solutions to the PDE (1.2), the uniqueness of the inverse problem of recovering the conductivity distribution with two measurements, and the convergence analysis of the reconstruction algorithm.

Appendix. Useful estimate.

LEMMA A.1. *Let $a \in L^\infty(\Omega)$ be a positive function satisfying $C_0 > a > c_0 > 0$, and let $F \in L^\infty(\Omega)^d$. Let V be a closed subset of $H^1(\Omega)$ such that*

$$H^1_0(\Omega) \subset V \subset H^1(\Omega).$$

Assume that $\phi \in V$ is such that

$$\nabla \cdot (a\nabla\phi) = \nabla \cdot (\mathbf{1}_\omega(x)F) \text{ in } \Omega. \tag{A.1}$$

Suppose that Ω contains a subset of $\Omega' \subset \Omega$ of class $\mathcal{C}^2$, such that dist$(\Omega', \partial\Omega) > d_0 > 0$, and such that $\omega \subset \Omega'$, or alternatively that Ω is a cube and that ϕ is periodic

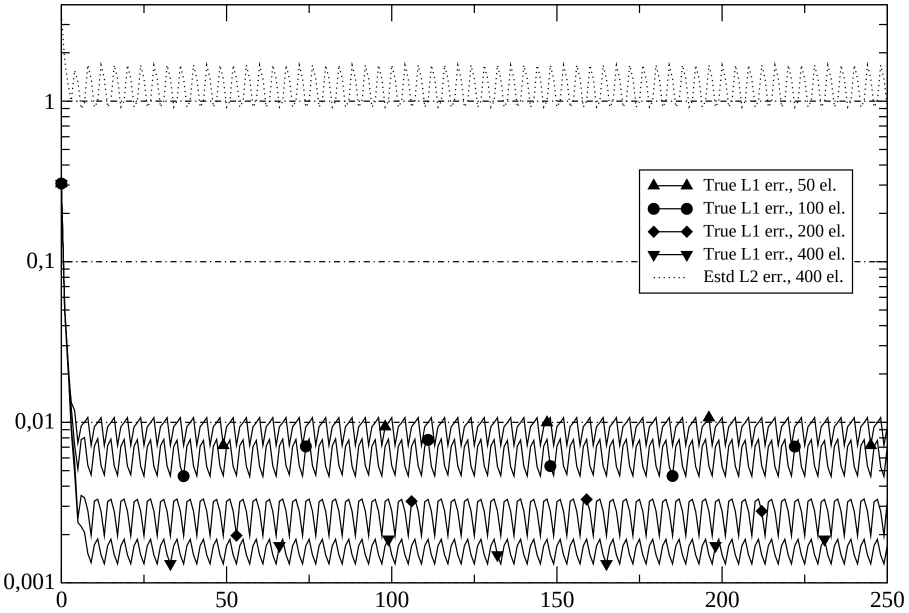


FIG. 6. *Convergence results. The curves labeled "True L1 err." correspond to the L^1 norm of the discrepancy between the real and reconstructed conductivity. The curve labeled "Estd L2 err., 400 el." represents the estimated error for the regular mesh designed with* 400 *boundary points.*

on that cube. Then,

$$\|\nabla\phi\|_{L^2(\Omega)^d} \leq \frac{1}{\sqrt{c_0}}|\omega|^{1/2}\|F\|_{L^\infty(\Omega)^d}. \tag{A.2}$$

Furthermore, there exist $\kappa > 0$ and $C > 0$, two positive constants depending only on Ω', d_0, c_0, and C_0, such that

$$\|\phi\|_{L^2(\Omega)} \leq C|\omega|^{\frac{1}{2}+\kappa}\|F\|_{L^\infty(\Omega)^d}. \tag{A.3}$$

At this point, let us emphasize the fact that the background is not required to be smooth. The proof uses Meyers's theorem [38].

Proof. We shall now prove Lemma A.1 for $V = H^1(\Omega)$ or $V = H_0^1(\Omega)$. Integrating (A.1) against ϕ, we obtain

$$\int_\Omega a\nabla\phi\cdot\nabla\phi\,dx = \int_\Omega \mathbf{1}_\omega F\cdot\nabla\phi\,dx.$$

Thus, using the Cauchy–Schwarz inequality, we get (A.2). Define $\psi \in H^1(\Omega)$ as the unique solution to

$$-\nabla\cdot(a\nabla\psi) = \phi \text{ in } \Omega,$$
$$\psi = 0 \text{ on } \partial\Omega.$$

Choose $f \in \mathcal{C}_0^\infty(\Omega)$ to be a cut-off function such that $f \equiv 1$ on Ω' and $0 \leq f \leq 1$. According to Meyers's theorem [38] (see also [15, pp. 35–45]), there exists an $\eta > 0$ depending only on Ω, c_0, C_0, and f such that $\psi \in W^{1,2+\eta}(\Omega')$, and we have

$$\|\nabla(\psi f)\|_{L^{2+\eta}(\Omega)} \leq C\,\|\phi\|_{L^{2+\eta}(\Omega)}.$$

Using the Gagliardo–Nirenberg inequality,

$$\begin{aligned}\|\phi\|_{L^{2+\eta}(\Omega)} &\leq C\,\|\phi\|^{\alpha}_{L^2(\Omega)}\,\|\nabla\phi\|^{1-\alpha}_{L^2(\Omega)^d} \text{ with } \alpha = \frac{\eta}{2+\eta}\\ &\leq C\,\|\phi\|^{\alpha}_{L^2(\Omega)}\,|\omega|^{\frac{1}{2+\eta}}\,.\end{aligned}$$

We then compute

$$\begin{aligned}\int_\Omega \phi^2\,dx &= \int_\Omega a\nabla\psi\cdot\nabla\phi\,dx\\ &= \int_\Omega \mathbf{1}_\omega F\cdot\nabla\psi\,dx\\ &\leq \|F\|_{L^\infty(\Omega)}\int_\Omega |\mathbf{1}_\omega\nabla\psi|\,dx.\end{aligned}$$

Using Hölder's inequality, we then obtain

$$\begin{aligned}\int_\Omega \phi^2\,dx &\leq \|F\|_{L^\infty(\Omega)}\,|\omega|^{\frac{1+\eta}{2+\eta}}\,\|\nabla(f\psi)\|_{L^{2+\eta}(\Omega)}\\ &\leq C\,\|F\|_{L^\infty(\Omega)}\,|\omega|\,\|\phi\|^{\alpha}_{L^2(\Omega)}\,.\end{aligned}$$

Consequently,

$$\|\phi\|_{L^2(\Omega)} \leq C\,\|F\|_{L^\infty(\Omega)}\,|\omega|^{\frac{1}{2}+\frac{\eta}{2(4+\eta)}}\,,$$

and therefore, choosing $\kappa = \frac{\eta}{2(4+\eta)}$ concludes the proof. □

REFERENCES

[1] G. Alessandrini, *Examples of instability in inverse boundary value problems*, Inverse Problems, 13 (1997), pp. 887–897.

[2] G. Alessandrini, V. Isakov, and J. Powell, *Local uniqueness in the inverse conductivity problem with one measurement*, Trans. Amer. Math. Soc., 347 (1995), pp. 3031–3041.

[3] G. Alessandrini and R. Magnanini, *The index of isolated critical points and solutions of elliptic equations in the plane*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 567–589.

[4] G. Alessandrini and V. Nesi, *Univalent σ-harmonic mappings*, Arch. Ration. Mech. Anal., 158 (2001), pp. 155–171.

[5] C. Alves and H. Ammari, *Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium*, SIAM J. Appl. Math., 62 (2001), pp. 94–106.

[6] H. Ammari and H. Kang, *High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter*, SIAM J. Math. Anal., 34 (2003), pp. 1152–1166.

[7] H. Ammari and H. Kang, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[8] H. Ammari and H. Kang, *Polarization and Moment Tensors with Applications to Inverse Problems and Effective Medium Theory*, Appl. Math. Sci. 162, Springer-Verlag, New York, 2007.

[9] H. Ammari and H. Kang, *Boundary layer techniques for solving the Helmholtz equation in the presence of small inhomogeneities*, J. Math. Anal. Appl., 296 (2004), pp. 190–208.

[10] H. Ammari, H. Kang, G. Nakamura, and K. Tanuma, *Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion*, J. Elasticity, 67 (2002), pp. 97–129.

[11] H. Ammari, O. Kwon, J. K. Seo, and E. J. Woo, *T-scan electrical impedance imaging system for anomaly detection*, SIAM J. Appl. Math., 65 (2004), pp. 252–266.

[12] H. Ammari, M. Vogelius, and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter* II. *The full Maxwell equations*, J. Math. Pure Appl., 80 (2001), pp. 769–814.

[13] M. Assenheimer, O. Laver-Moskovitz, D. Malonek, D. Manor, U. Nahliel, R. Nitzan, and A. Saad, *The T-scan technology: Electrical impedance as a diagnostic tool for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 1–8.

[14] D. C. Barber and B. H. Brown, *Applied potential tomography*, J. Phys. Sci. Instrum., 17 (1984), pp. 723–733.

[15] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.

[16] J. Bercoff, M. Tanter, and M. Fink, *Supersonic shear imaging: A new technique for soft tissue elasticity mapping*, IEEE Trans. Ultrasonics Ferro. Freq. Control, 51 (2004), pp. 396–409.

[17] E. Beretta, E. Francini, and M. Vogelius, *Asymptotic formulas for steady state voltage potentials in the presence of thin inhomogeneities. A rigorous error analysis*, J. Math. Pure Appl., 82 (2003), pp. 1277–1301.

[18] M. Bruhl, M. Hanke, and M. S. Vogelius, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[19] Y. Capdeboscq, J. Fehrenbach, F. de Gournay, and O. Kavian, *An optimal control approach to imaging by modification*, in preparation.

[20] Y. Capdeboscq and M. S. Vogelius, *A general representation formula for boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 159–173.

[21] Y. Capdeboscq and M. S. Vogelius, *Optimal asymptotic estimates for the volume of internal inhomogeneities in terms of multiple boundary measurements*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 227–240.

[22] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[23] A. Friedman and M. S. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[24] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[25] V. A. Cherepenin, A. Karpov, A. Korjenevsky, V. Kornienko, A. Mazaletskaya, D. Mazourov, and D. Meister, *A* 3*D electrical impedance tomography (EIT) system for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 9–18.

[26] E. Fabes, H. Kang, and J. K. Seo, *Inverse conductivity problem with one measurement: Error estimates and approximate identification for perturbed disks*, SIAM J. Math. Anal., 30 (1999), pp. 699–720.

[27] D. B. Geselowitz, *An application of electrocardiographic lead theory to impedance plethysmography*, IEEE Trans. Biomed. Eng., 18 (1971), pp. 38–41.

[28] F. Hecht, O. Pironneau, K. Ohtsuka, and A. Le Hyaric, *FreeFem++*, http://www.freefem.org/ (2007).

[29] D. Holder, *Clinical and Physiological Applications of Electrical Impedance Tomography*, UCL Press, London, 1993.

[30] J. Jossinet, B. Lavandier, and D. Cathignol, *Impedance modulation by pulsed ultrasound*, Ann. New York Acad. Sci., 873 (1999), pp. 396–407.

[31] J. Jossinet, C. Trillaud, and S. Chesnais, *Impedance changes in liver tissue exposed in vitro to high-energy ultrasound*, Physiol. Meas., 26 (2005), pp. 49–58.

[32] Y. J. Kim, O. Kwon, J. K. Seo, and E. J. Woo, *Uniqueness and convergence of conductivity image reconstruction in magnetic resonance electrical impedance tomography*, Inverse Problems, 19 (2003), pp. 1213–1225.

[33] S. Kim, O. Kwon, J. K. Seo, and J.-R. Yoon, *On a nonlinear partial differential equation arising in magnetic resonance electrical impedance tomography*, SIAM J. Math. Anal., 34 (2002), pp. 511–526.

[34] R. V. Kohn and M. S. Vogelius, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–298.

[35] F. Korber, *Über der Einfluss Des Druckes auf das elektrolytische Leitvermögen von Lösungen*, Z. Phys. Chem., 67 (1909), pp. 212–248.

[36] O. Kwon, J. K. Seo, and J.-R. Yoon, *A real-time algorithm for the location search of discontinuous conductivities with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[37] O. Kwon, E. J. Woo, J.-R. Yoon, and J. K. Seo, *Magnetic resonance electrical impedance tomography (MREIT): Simulation study of J-substitution algorithm*, IEEE Trans. Biomed. Eng., 49 (2002), pp. 160–167.

[38] N. G. Meyers, *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.

[39] F. Santosa and M. Vogelius, *A backprojection algorithm for electrical impedance imaging*, SIAM J. Appl. Math., 50 (1990), pp. 216–243.

[40] B. Scholz, *Towards virtual electrical breast biopsy: Space-frequency MUSIC for trans-admittance data*, IEEE Trans. Med. Imaging, 21 (2002), pp. 588–595.

[41] J. K. Seo, *A uniqueness result on inverse conductivity problem with two measurements*, J. Fourier Anal. Appl., 2 (1996), pp. 227–235.

[42] E. J. Woo, J. G. Webster, and W. J. Tompkins, *A robust image reconstruction algorithm and its parallel implementation in electrical impedance tomography*, IEEE Trans. Med. Imaging, 12 (1993), pp. 137–146.

[43] T. Yorkey, J. Webster, and W. Tompkins, *Comparing reconstruction algorithms for electrical impedance tomography*, IEEE Trans. Biomed. Eng., 34 (1987), pp. 843–852.

[44] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1574–1599

EFFECTIVE TRANSPORT EQUATIONS AND ENHANCED BACKSCATTERING IN RANDOM WAVEGUIDES[*]

JOSSELIN GARNIER[†] AND KNUT SØLNA[‡]

Abstract. In this paper we derive a general system of transport equations for the moments of reflected and transmitted mode amplitudes in a randomly perturbed waveguide, in a regime where backscattering is significant. The derivation is based on a limit theorem for the system of coupled differential equations for the mode amplitudes, in the limit where the amplitude of the random fluctuations of the medium is small, the correlation lengths in the transverse and longitudinal directions are of the same order of the wavelength, and the waveguide is long. Using this system we derive several results in specific regimes, including the enhanced backscattering phenomenon for the reflected wave: when an incoming monochromatic wave with a specific incidence angle is present, the mean reflected power has a local maximum in the backward direction twice as large as the mean reflected power in the other directions.

Key words. acoustic waveguides, random media, asymptotic analysis

AMS subject classifications. 76B15, 35Q99, 60F05

DOI. 10.1137/070694909

1. Introduction.

This paper is devoted to the analysis of wave propagation in a random waveguide. We use a separation of scales technique introduced by Papanicolaou and his co-authors. Although this technique was originally introduced for wave propagation in one-dimensional random media [1], it has been shown recently that it is possible to extend the technique to three-dimensional random media in the context of waveguides [5, 7, 6]. By writing the coupled mode equations for the complex mode amplitudes, diffusion approximation theorems can be applied, leading to differential equations driven by Brownian motions whose solutions are Itô diffusion processes. In [5, 6, 7] the analysis was restricted to the forward scattering approximation, where the conversion from forward-going to backward-going modes is neglected. This regime is characterized by the equipartition of energy: the mean transmitted mode powers become uniformly distributed when the waveguide is long enough. In this paper, we revisit this analysis in the general case and take into account backscattering. We derive a system of transport equations for the moments of the reflected and transmitted mode amplitudes in the regime where the fluctuations of the random medium have a small amplitude and a correlation length of the same order as the typical wavelength. This allows us to exhibit the enhanced backscattering phenomenon: when a monochromatic input mode is applied, the mean reflected mode powers become uniformly distributed, except for the mode corresponding to the backward direction, where the mean reflected power is twice the mean power of the other modes. This phenomenon, also known as weak localization, is well referenced in the physical literature and it has been observed in several experimental contexts, such as in optics with

[*] Received by the editors June 20, 2007; accepted for publication (in revised form) February 29, 2008; published electronically June 6, 2008. This work was supported by ONR grants N00014-02-1-0089 and N00014-05-1-0442.

http://www.siam.org/journals/siap/68-6/69490.html

[†] Laboratoire de Probabilités et Modèles Aléatoires & Laboratoire Jacques-Louis Lions, Université Paris VII, 2 Place Jussieu, 75251 Paris Cedex 5, France (garnier@math.jussieu.fr). This author was supported by ANR.

[‡] Department of Mathematics, University of California, Irvine, CA 92697 (ksolna@math.uci.edu). This author was supported by NSF grant DMS0307011 and the Sloan Foundation.

powder suspensions [17, 15], with biological tissues [18], and with ultracold atoms [11] as well as in acoustics [14]. The physical analysis of the weak localization is based on diagrammatic expansions [16], where interference effects between direct and reverse paths play a crucial role. Here we give a mathematical derivation of this phenomenon by an asymptotic analysis in the context of random waveguides. We also compute the second moments of the reflected mode powers and show that these quantities are not statistically stable in the sense that their fluctuations are of the same order as their mean values. This means that it is necessary to average the reflected power to detect the enhanced backscattering. This point was already mentioned in the physical literature, and we give here a quantitative analysis of this phenomenon.

2. Propagation in a random waveguide.

We consider wave propagation in a waveguide where the medium parameters have small random perturbations. Many modern applications involve propagation in waveguides [3, 4, 13]. We will here describe the problem in a scaling regime where the radius of the waveguide is of the order of a few wavelengths and with the medium parameters varying randomly in the longitudinal and transversal directions with a correlation length on the order of the wavelength. This scaling regime was also considered in [5, 7, 6]. The analysis could be generalized to other scaling limits whenever diffusion approximation theorems can be applied.

We consider linear acoustic waves propagating in three spatial dimensions:

$$\rho(\mathbf{x}, z)\frac{\partial \mathbf{u}}{\partial t} + \nabla p = \mathbf{0}, \quad \frac{1}{K(\mathbf{x}, z)}\frac{\partial p}{\partial t} + \nabla \cdot \mathbf{u} = 0 \quad \text{for } \mathbf{x} \in \mathcal{D} \text{ and } t, z \in \mathbb{R}, \tag{2.1}$$

where p is the pressure field, $\mathbf{u}$ is the velocity field, ρ is the density of the medium, K is the bulk modulus, and $(\mathbf{x}, z) = (x, y, z)$ stands for the space coordinates. The cross section of the waveguide is denoted by $\mathcal{D}$, and we shall use Dirichlet boundary conditions

$$p(t, \mathbf{x}, z) = 0 \qquad \text{for } \mathbf{x} \in \partial\mathcal{D} \text{ and } z \in \mathbb{R}. \tag{2.2}$$

The direction of propagation along the waveguide axis is z and the transverse coordinates are denoted by $\mathbf{x} \in \mathcal{D}$. The random part of the waveguide occupies the region $z \in [0, L/\varepsilon^2]$ and is embedded in between two homogeneous waveguide sections. Inside the perturbed waveguide the bulk modulus is randomly varying, and we assume for simplicity that the density is homogeneous:

$$\frac{1}{K(\mathbf{x}, z)} = \begin{cases} \frac{1}{\bar{K}}\left(1 + \varepsilon\nu(\mathbf{x}, z)\right) & \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in [0, L/\varepsilon^2], \\ \frac{1}{\bar{K}} & \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in (-\infty, 0) \cup (L/\varepsilon^2, \infty), \end{cases} \tag{2.3}$$

$$\rho(\mathbf{x}, z) = \bar{\rho} \quad \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in (-\infty, \infty). \tag{2.4}$$

It is possible to take into account a randomly varying density; this complicates the algebra but leads to the same general system of transport equations for the moments of reflected and transmitted mode amplitudes. Such a generalization was carried out in the case of randomly layered media in [5, section 17.3]. Here ε is a small parameter and $\nu(\mathbf{x}, z)$ is a zero-mean random process that describes the random medium fluctuations that are mixing in the z-direction. This weakly heterogeneous regime can be encountered, for instance, in underwater acoustics [4, 8].

2.1. Waveguide modes. In a homogeneous waveguide $\nu = 0$, the complex amplitude of a monochromatic wave $p(t, \mathbf{x}, z) = \hat{p}(\omega, \mathbf{x}, z)e^{-i\omega t}$ at frequency ω satisfies the time-harmonic form of the wave equation (Helmholtz equation):

$$\partial_z^2 \hat{p} + \Delta_\perp \hat{p} + k^2(\omega)\hat{p} = 0. \tag{2.5}$$

Here $\Delta_\perp$ is the transverse Laplacian, $k(\omega) = \omega/\bar{c}$ is the wavenumber, and $\bar{c} = \sqrt{\bar{K}/\bar{\rho}}$ is the homogenized wave speed. The monochromatic wave can be decomposed in terms of normal modes which are the (normalized in $L^2(\mathcal{D})$) solutions of the eigenvalue problem

$$-\Delta_\perp \phi_j(\mathbf{x}) = \lambda_j \phi_j(\mathbf{x}), \ \mathbf{x} \in \mathcal{D}, \qquad \phi_j(\mathbf{x}) = 0, \ \mathbf{x} \in \partial\mathcal{D},$$

for $j = 1, 2, \ldots$. The eigenvalues are positive and nondecreasing, and we assume for simplicity that they are simple, so we have $0 < \lambda_1 < \lambda_2 < \cdots$. The eigenmodes are real and form an orthonormal set

$$\int_{\mathcal{D}} \phi_j(\mathbf{x})\phi_l(\mathbf{x})\, d\mathbf{x} = \delta_{jl}.$$

For a given frequency ω, there exists a unique integer $N(\omega)$ such that $\lambda_{N(\omega)} \le k^2(\omega) < \lambda_{N(\omega)+1}$, with the convention that $N(\omega) = 0$ if $\lambda_1 > k^2(\omega)$. The modal wavenumbers $\beta_j(\omega)$ for $1 \le j \le N(\omega)$ are defined by

$$\beta_j(\omega) = \sqrt{k^2(\omega) - \lambda_j}\,. \tag{2.6}$$

The solutions $\hat{p}_j(\omega, \mathbf{x}, z) = \phi_j(\mathbf{x})e^{\pm i\beta_j(\omega)z}$, $j = 1, \ldots, N(\omega)$, of the wave equation (2.5) are the propagating waveguide modes. For $j > N(\omega)$ we define the modal wavenumbers by $\beta_j(\omega) = [\lambda_j - k^2(\omega)]^{1/2}$, and the corresponding solutions $\hat{q}_j(\omega, \mathbf{x}, z) = \phi_j(\mathbf{x})e^{\pm\beta_j(\omega)z}$ of the wave equation (2.5) are the evanescent modes.

From now on we consider the perturbed waveguide as described by (2.3)–(2.4). We expand the time-harmonic field inside the randomly perturbed waveguide in terms of the transverse eigenmodes of the unperturbed waveguide:

$$\hat{p}(\omega, \mathbf{x}, z) = \sum_{j=1}^{N(\omega)} \phi_j(\mathbf{x})\hat{p}_j(\omega, z) + \sum_{j=N(\omega)+1}^{\infty} \phi_j(\mathbf{x})\hat{q}_j(\omega, z), \tag{2.7}$$

where $\hat{p}_j$ is the amplitude of the jth propagating mode and $\hat{q}_j$ is the amplitude of the jth evanescent mode. For $1 \le j \le N(\omega)$, let $\hat{a}_j(\omega, z)$ and $\hat{b}_j(\omega, z)$ represent the amplitudes of the forward- and backward-propagating modes, with the forward direction referring to the z-direction. They are given by

$$\hat{p}_j(\omega, z) = \frac{1}{\sqrt{\beta_j(\omega)}}\left(\hat{a}_j(\omega, z)e^{i\beta_j(\omega)z} + \hat{b}_j(\omega, z)e^{-i\beta_j(\omega)z}\right), \tag{2.8}$$

$$\frac{d\hat{p}_j(\omega, z)}{dz} = i\sqrt{\beta_j(\omega)}\left(\hat{a}_j(\omega, z)e^{i\beta_j(\omega)z} - \hat{b}_j(\omega, z)e^{-i\beta_j(\omega)z}\right). \tag{2.9}$$

We next make a change of the z variable by introducing the rescaled processes $\hat{a}_j^\varepsilon(\omega, z)$, $\hat{b}_j^\varepsilon(\omega, z)$, $j = 1, \ldots, N(\omega)$, given by

$$\hat{a}_j^\varepsilon(\omega, z) = \hat{a}_j\left(\omega, \frac{z}{\varepsilon^2}\right), \qquad \hat{b}_j^\varepsilon(\omega, z) = \hat{b}_j\left(\omega, \frac{z}{\varepsilon^2}\right). \tag{2.10}$$

By projecting the wave equation (2.5) on the transverse eigenmodes and by expressing the amplitudes of the evanescent modes in terms of the amplitudes of the propagating modes [5, 6], we obtain the following mode coupling equations for the amplitude processes $\hat{a}^\varepsilon(\omega,z)=(\hat{a}^\varepsilon_j(\omega,z))_{j=1,\dots,N(\omega)}$ and $\hat{b}^\varepsilon(\omega,z)=(\hat{b}^\varepsilon_j(\omega,z))_{j=1,\dots,N(\omega)}$:

$$\frac{d\hat{a}^\varepsilon}{dz}=\left[\frac{1}{\varepsilon}\mathbf{H}^{(aa)}+\mathbf{G}^{(aa)}\right]\left(\omega,\frac{z}{\varepsilon^2}\right)\hat{a}^\varepsilon+\left[\frac{1}{\varepsilon}\mathbf{H}^{(ab)}+\mathbf{G}^{(ab)}\right]\left(\omega,\frac{z}{\varepsilon^2}\right)\hat{b}^\varepsilon, \tag{2.11}$$

$$\frac{d\hat{b}^\varepsilon}{dz}=\left[\frac{1}{\varepsilon}\overline{\mathbf{H}^{(ab)}}+\overline{\mathbf{G}^{(ab)}}\right]\left(\omega,\frac{z}{\varepsilon^2}\right)\hat{a}^\varepsilon+\left[\frac{1}{\varepsilon}\overline{\mathbf{H}^{(aa)}}+\overline{\mathbf{G}^{(aa)}}\right]\left(\omega,\frac{z}{\varepsilon^2}\right)\hat{b}^\varepsilon, \tag{2.12}$$

with the two-point boundary conditions

$$\hat{a}^\varepsilon_j(\omega,0)=0,\quad \hat{b}^\varepsilon_j(\omega,L)=\hat{b}^{\mathrm{inc}}_j(\omega), \tag{2.13}$$

which correspond to a left-propagating wave incoming from the right homogeneous waveguide. The matrices $\mathbf{G}$ describe coupling via evanescent modes [5]. The $N(\omega)\times N(\omega)$ coupling matrices have entries of form

$$H^{(aa)}_{jl}(\omega,z)=\frac{ik^2(\omega)}{2}\frac{C_{jl}(z)}{\sqrt{\beta_j\beta_l(\omega)}}e^{i(\beta_l(\omega)-\beta_j(\omega))z}, \tag{2.14}$$

$$G^{(aa)}_{jl}(\omega,z)=\frac{ik^4(\omega)}{4}\sum_{l'>N(\omega)}\int_{-\infty}^{\infty}\frac{C_{jl'}(z)C_{ll'}(z+s)}{\sqrt{\beta_j\beta_{l'}^2\beta_l(\omega)}}e^{i\beta_l(\omega)(z+s)-i\beta_j(\omega)z-\beta_{l'}(\omega)|s|}\,ds,$$

$$H^{(ab)}_{jl}(\omega,z)=-e^{-2i\beta_j(\omega)z}\overline{H^{(aa)}_{jl}(\omega,z)},\quad G^{(ab)}_{jl}(\omega,z)=-e^{-2i\beta_j(\omega)z}\overline{G^{(aa)}_{jl}(\omega,z)}, \tag{2.15}$$

$$C_{jl}(z)=\int_{\mathcal{D}}\phi_j(\mathbf{x})\phi_l(\mathbf{x})\nu(\mathbf{x},z)\,d\mathbf{x} \tag{2.16}$$

for $j,l=1,\dots,N(\omega)$.

2.2. Channel coupled wave approximation. We use an invariant imbedding step to convert the boundary value problem to an initial value problem with the objective being to characterize the reflected and transmitted wave fields. Accordingly we introduce the $N(\omega)\times N(\omega)$ reflection and transmission matrices $\boldsymbol{\mathcal{R}}^\varepsilon$ and $\boldsymbol{\mathcal{T}}^\varepsilon$ by

$$\hat{b}^\varepsilon(\omega,0)=\boldsymbol{\mathcal{T}}^\varepsilon(\omega,z)\hat{b}^\varepsilon(\omega,z),\qquad \hat{a}^\varepsilon(\omega,z)=\boldsymbol{\mathcal{R}}^\varepsilon(\omega,z)\hat{b}^\varepsilon(\omega,z). \tag{2.17}$$

Using (2.11)–(2.12) we find that these matrices solve the problems

$$\frac{d}{dz}\boldsymbol{\mathcal{R}}^\varepsilon=\mathbf{H}^{b,\varepsilon}+\mathbf{H}^{a,\varepsilon}\boldsymbol{\mathcal{R}}^\varepsilon-\boldsymbol{\mathcal{R}}^\varepsilon\overline{\mathbf{H}^{a,\varepsilon}}-\boldsymbol{\mathcal{R}}^\varepsilon\overline{\mathbf{H}^{b,\varepsilon}}\boldsymbol{\mathcal{R}}^\varepsilon, \tag{2.18}$$

$$\frac{d}{dz}\boldsymbol{\mathcal{T}}^\varepsilon=-\boldsymbol{\mathcal{T}}^\varepsilon\left(\overline{\mathbf{H}^{a,\varepsilon}}+\overline{\mathbf{H}^{b,\varepsilon}}\boldsymbol{\mathcal{R}}^\varepsilon\right), \tag{2.19}$$

where we defined

$$\mathbf{H}^{a,\varepsilon}(\omega,z)=\frac{1}{\varepsilon}\mathbf{H}^{(aa)}\left(\omega,\frac{z}{\varepsilon^2}\right)+\mathbf{G}^{(aa)}\left(\omega,\frac{z}{\varepsilon^2}\right),$$

$$\mathbf{H}^{b,\varepsilon}(\omega,z)=\frac{1}{\varepsilon}\mathbf{H}^{(ab)}\left(\omega,\frac{z}{\varepsilon^2}\right)+\mathbf{G}^{(ab)}\left(\omega,\frac{z}{\varepsilon^2}\right),$$

and where $\boldsymbol{\mathcal{R}}^\varepsilon(\omega,z)$ and $\boldsymbol{\mathcal{T}}^\varepsilon(\omega,z)$ take initial values at $z=0$

$$\boldsymbol{\mathcal{T}}^\varepsilon(\omega,0)=\mathbf{I},\quad \boldsymbol{\mathcal{R}}^\varepsilon(\omega,0)=\mathbf{0}. \tag{2.20}$$

We remark that energy conservation leads to the reflection-transmission conservation relation

$$\boldsymbol{\mathcal{R}}^{\varepsilon\dagger}\boldsymbol{\mathcal{R}}^{\varepsilon}+\boldsymbol{\mathcal{T}}^{\varepsilon\dagger}\boldsymbol{\mathcal{T}}^{\varepsilon}=\mathbf{I}, \tag{2.21}$$

where the sign † stands for the conjugate transpose [5].

The initial value problem (2.18) is a stochastic Riccati matrix equation, and it can be analyzed in the limit $\varepsilon \to 0$ using the theory of diffusion approximations [5, 10]. The matrix $\mathbf{H}^{(aa)}$ contains rapidly varying phase factors and is centered with respect to the randomness that fluctuates on the scale ε^2 and is mixing in the z-direction. In this white-noise scaling regime we can then identify the corresponding infinitesimal generator and the associated white-noise model that describes the joint law of the transmission and reflection matrices in the limit $\varepsilon \to 0$.

3. The reflected wave field.

We consider the problem of characterizing the modal distribution of the reflected or transmitted waves. We consider in this section the reflected waves, and we will address the transmitted waves in section 4.

3.1. The moments of the reflected time-harmonic field.

We first consider the time-harmonic reflected field for a single frequency ω. We can compute the limit of the moments of the reflection matrix by using the diffusion approximation theory.

PROPOSITION 3.1. *Let* $\mathbf{p}=\{(j_1,l_1),\dots,(j_{|\mathbf{p}|},l_{|\mathbf{p}|})\}\in\{1,\dots,N(\omega)\}^{2|\mathbf{p}|}$ *denote a multi-index (*$|\mathbf{p}|$ *is the number of index pairs in* $\mathbf{p}$*). We introduce the moments of elements* $\mathcal{R}^{\varepsilon}_{jl}$ *of the reflection matrix:*

$$\mathcal{M}^{\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,z)=\mathbb{E}\Big[\prod_{(j,l)\in\mathbf{p}}\mathcal{R}^{\varepsilon}_{jl}(\omega,z)\prod_{(m,n)\in\mathbf{q}}\overline{\mathcal{R}^{\varepsilon}_{mn}(\omega,z)}\Big].$$

These moments converge as $\varepsilon\to 0$ *to the solution* $\mathcal{M}_{\mathbf{p},\mathbf{q}}$ *of the system*

$$\frac{d\mathcal{M}_{\mathbf{p},\mathbf{q}}}{dz}=D_{\mathbf{p},\mathbf{q}}(\omega)\mathcal{M}_{\mathbf{p},\mathbf{q}}+\big[\mathcal{S}_{\omega}(\mathcal{M})\big]_{\mathbf{p},\mathbf{q}}, \tag{3.1}$$

with the initial conditions

$$\mathcal{M}_{\mathbf{p},\mathbf{q}}(\omega,z=0)=\mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|). \tag{3.2}$$

Here we have defined the linear operator $\mathcal{S}_\omega$*,*

$$\begin{aligned}
\big[\mathcal{S}_{\omega}(\mathcal{M})\big]_{\mathbf{p},\mathbf{q}} &= -\sum_{(j,l)\in\mathbf{p}} d^{(1)}_{jl}\mathcal{M}_{\mathbf{p}|\{(j,l)|(l,j)\},\mathbf{q}}-\sum_{(j,l)\in\mathbf{q}} d^{(1)}_{jl}\mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l,j)\}}\\
&-\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} d^{(2)}_{jl}\mathcal{M}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(l,\tilde{l})\},\mathbf{q}}+d^{(2)}_{\tilde{j}\tilde{l}}\mathcal{M}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(\tilde{l},l)\},\mathbf{q}}\\
&-\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} d^{(1)}_{j\tilde{l}}\mathcal{M}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(\tilde{l},l)\},\mathbf{q}}+d^{(1)}_{\tilde{j}l}\mathcal{M}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(l,\tilde{l})\},\mathbf{q}}\\
&-\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} \Big[d^{(5)}_{jl}+d^{(5)}_{\tilde{j}\tilde{l}}+d^{(1)}_{j\tilde{j}}+d^{(1)}_{l\tilde{l}}\Big]\mathcal{M}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\},\mathbf{q}}\\
&-\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} d^{(2)}_{jl}\mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(l,\tilde{l})\}}+d^{(2)}_{\tilde{j}\tilde{l}}\mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(\tilde{l},l)\}}
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} d^{(1)}_{j\tilde{l}} \mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(\tilde{l},l)\}} + d^{(1)}_{\tilde{j}l} \mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(l,\tilde{l})\}} \\
&- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} \left[d^{(5)}_{jl} + d^{(5)}_{\tilde{j}\tilde{l}} + d^{(1)}_{j\tilde{j}} + d^{(1)}_{l\tilde{l}} \right] \mathcal{M}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\}} \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} d^{(3)}_{jl\tilde{j}\tilde{l}} \mathcal{M}_{\mathbf{p}|(j,l),\mathbf{q}|(\tilde{j},\tilde{l})} \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} \sum_{k=1\neq j}^{N} \left[d^{(4)}_{jk\tilde{j}} \mathcal{M}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(k,\tilde{l})\}} + d^{(4)}_{jk\tilde{l}} \mathcal{M}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k)\}} \right] \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} \sum_{k=1\neq l}^{N} \left[d^{(4)}_{lk\tilde{l}} \mathcal{M}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k)\}} + d^{(4)}_{lk\tilde{j}} \mathcal{M}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(k,\tilde{l})\}} \right] \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} \sum_{k_1,k_2=1}^{N} d^{(2)}_{k_1k_2} \mathcal{M}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k_1),(k_2,\tilde{l})\}} \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} \sum_{k_1,k_2=1}^{N} d^{(5)}_{k_1k_2} \mathcal{M}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k_2),(k_1,\tilde{l})\}},
\end{aligned}
$$

and we have used the following notation: If $\mathbf{p}$ *is a multi-index and* $\{(j_1,l_1),\dots,(j_m,l_m)\} \subset \mathbf{p}$, *then* $\mathbf{p}|\{(j_1,l_1),\dots,(j_m,l_m)|(\tilde{j}_1,\tilde{l}_1),\dots,(\tilde{j}_n,\tilde{l}_n)\}$ *denotes the new multi-index obtained from* $\mathbf{p}$ *by removing the index pairs* $\{(j_1,l_1),\dots,(j_m,l_m)\}$ *and by adding the new index pairs* $\{(\tilde{j}_1,\tilde{l}_1),\dots,(\tilde{j}_n,\tilde{l}_n)\}$. *Finally, the coefficients* $D_{\mathbf{p},\mathbf{q}}(\omega)$ *and* $d^{(j)}(\omega)$, $j=1,\dots,5$, *are defined by*

$$
\begin{aligned}
D_{\mathbf{p},\mathbf{q}} = {} & i \sum_{(j,l)\in\mathbf{p}} \big(\kappa_j + \kappa_l\big) - i \sum_{(j,l)\in\mathbf{q}} \big(\kappa_j + \kappa_l\big) \\
& - \sum_{(j,l)\in\mathbf{p}} \sum_{k=1}^{N} \Big(\overline{\Gamma_{jk}} + \overline{\Gamma_{lk}} + \widetilde{\Gamma}_{jk} + \widetilde{\Gamma}_{lk} \Big) - \sum_{(j,l)\in\mathbf{q}} \sum_{k=1}^{N} \Big(\Gamma_{jk} + \Gamma_{lk} + \overline{\widetilde{\Gamma}_{jk}} + \overline{\widetilde{\Gamma}_{lk}} \Big) \\
& - 2 \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} \Big(\check{\Gamma}_{j\tilde{j}} + \check{\Gamma}_{l\tilde{l}} \Big) - 2 \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} \Big(\check{\Gamma}_{j\tilde{j}} + \check{\Gamma}_{l\tilde{l}} \Big) \\
& - \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{p}} 2\Re\big(\check{\Gamma}_{j\tilde{l}}\big) - \sum_{(j,l)\in\mathbf{q}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} 2\Re\big(\check{\Gamma}_{j\tilde{l}}\big) \\
& + 2 \sum_{(j,l)\in\mathbf{p}} \sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} \Big(\check{\Gamma}_{j\tilde{l}} + \check{\Gamma}_{l\tilde{j}} + \check{\Gamma}_{\tilde{j}j} + \check{\Gamma}_{\tilde{l}l} \Big),
\end{aligned}
$$

$$
\begin{aligned}
& d^{(1)}_{jl}(\omega) = 2\Re\big[\widetilde{\Gamma}_{jl}(\omega)\big]\mathbf{1}_{j\neq l}, \qquad d^{(2)}_{jl}(\omega) = 2\Re\big[\Gamma_{jl}(\omega)\big], \\
& d^{(3)}_{jl\tilde{j}\tilde{l}}(\omega) = 2\Re\big[\Gamma_{jl}(\omega)\big]\mathbf{1}_{(j,l)=(\tilde{j},\tilde{l}) \ or \ (j,l)=(\tilde{l},\tilde{j})}, \\
& d^{(4)}_{jk\tilde{j}}(\omega) = 2\Re\big[\widetilde{\Gamma}_{jk}(\omega)\big]\mathbf{1}_{j=\tilde{j}}, \qquad d^{(5)}_{k_1k_2}(\omega) = 2\Re\big[\Gamma_{k_1k_2}(\omega)\big]\mathbf{1}_{k_1\neq k_2},
\end{aligned}
$$

with $\Re(x)$ *the real part of* x, $\mathbf{1}_{j\neq l} = 1$ *if* $j \neq l$ *and* 0 *otherwise, and*

$$
\check{\Gamma}_{jl}(\omega) = \frac{k^4(\omega)}{4} \frac{\int_0^\infty \mathbb{E}[C_{jj}(0)C_{ll}(s)]\,ds}{\beta_j\beta_l(\omega)}, \tag{3.3}
$$

$$\widetilde{\Gamma}_{jl}(\omega) = \frac{k^4(\omega)}{4} \frac{\int_0^\infty e^{i(\beta_j(\omega)-\beta_l(\omega))s} \mathbb{E}[C_{jl}(0)C_{jl}(s)]\, ds}{\beta_j \beta_l(\omega)}, \tag{3.4}$$

$$\Gamma_{jl}(\omega) = \frac{k^4(\omega)}{4} \frac{\int_0^\infty e^{i(\beta_j(\omega)+\beta_l(\omega))s} \mathbb{E}[C_{jl}(0)C_{jl}(s)]\, ds}{\beta_j \beta_l(\omega)}, \tag{3.5}$$

$$\kappa_l(\omega) = \frac{k^4(\omega)}{4} \sum_{l' > N(\omega)} \frac{\int_{-\infty}^{\infty} \mathbb{E}\left[C_{ll'}(0)\, C_{ll'}(s)\right] e^{i\beta_l(\omega)s - \beta_{l'}(\omega)|s|}\, ds}{\beta_l \beta_{l'}(\omega)}. \tag{3.6}$$

We will discuss applications of this proposition in the next sections, but we first give a generalization of the result for the two-frequency case. The proof of Proposition 3.1 is a simplified version of the proof of the next proposition, so we shall present it only for this second proposition.

3.2. The transport equations for the two-frequency moments. We have the following result, which is proved in Appendix A. We use the same notation as in Proposition 3.1.

PROPOSITION 3.2. *We introduce the moments of elements $\mathcal{R}^\varepsilon_{jl}$ of the reflection matrix at two nearby frequencies:*

$$\mathcal{U}^\varepsilon_{\mathbf{p},\mathbf{q}}(\omega, h, z) = \mathbb{E}\left[\prod_{(j,l)\in\mathbf{p}} \mathcal{R}^\varepsilon_{jl}(\omega + \varepsilon^2 h/2, z) \prod_{(m,n)\in\mathbf{q}} \overline{\mathcal{R}^\varepsilon_{mn}(\omega - \varepsilon^2 h/2, z)} \right], \tag{3.7}$$

where we set $\mathcal{R}^\varepsilon_{jl}(\omega \pm \varepsilon^2 h/2, z) = 0$ if j or l is larger than $N(\omega \pm \varepsilon^2 h/2)$. The family of Fourier transforms (in h)

$$\mathcal{W}^\varepsilon_{\mathbf{p},\mathbf{q}}(\omega, \tau, z) = \frac{1}{2\pi} \int e^{-ih[\tau - \phi_{\mathbf{p},\mathbf{q}}(\omega)z]} \mathcal{U}^\varepsilon_{\mathbf{p},\mathbf{q}}(\omega, h, z)\, dh \tag{3.8}$$

converges as $\varepsilon \to 0$ to the solution $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ of the system of transport equations

$$\frac{\partial \mathcal{W}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi_{\mathbf{p},\mathbf{q}}(\omega) \frac{\partial \mathcal{W}_{\mathbf{p},\mathbf{q}}}{\partial \tau} = D_{\mathbf{p},\mathbf{q}}(\omega) \mathcal{W}_{\mathbf{p},\mathbf{q}} + \left[\mathcal{S}_\omega(\mathcal{W})\right]_{\mathbf{p},\mathbf{q}}, \tag{3.9}$$

with the initial conditions

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega, \tau, z = 0) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\delta(\tau). \tag{3.10}$$

The coefficient $\phi_{\mathbf{p},\mathbf{q}}(\omega)$ is defined by

$$\phi_{\mathbf{p},\mathbf{q}}(\omega) = \frac{1}{2} \sum_{(j,l)\in\mathbf{p}} \left(\beta'_j(\omega) + \beta'_l(\omega)\right) + \frac{1}{2} \sum_{(j,l)\in\mathbf{q}} \left(\beta'_j(\omega) + \beta'_l(\omega)\right), \tag{3.11}$$

with $\beta'_j(\omega) = d\beta_j(\omega)/d\omega$, while the coefficient $D_{\mathbf{p},\mathbf{q}}(\omega)$ and the operator $\mathcal{S}_\omega$ are given in Proposition 3.1.

The set of transport equations (3.9) describes accurately the reflected wave field, and it is the key tool in analyzing various applications with waves in random waveguides. The corresponding transport equations in the layered case with one-dimensional medium variations were first obtained in [1]. They have played a crucial role in the analysis of a wide range of applications, and they have been generalized to describe a wide range of propagation scenarios in [5]. The transport equations given in Proposition 3.2 provide a rigorous tool for studying qualitatively and quantitatively the multiple scattering effects in a nonlayered random medium.

Remark. The convergence of $\mathcal{W}^\varepsilon$ and the existence and uniqueness of the solution $\mathcal{W}$ to the system of transport equations (3.9) are established in the space $\mathcal{C}([0,L],S'_H)$, where S'_H is a generalization of the space of distributions introduced in [12] to study the analogous problem with $N=1$ (randomly layered media). The space S'_H can be identified as the dual of the space S_H of the test functions $\lambda = (\lambda_{\mathbf{p},\mathbf{q}}(\tau))_{\mathbf{p}\in\{1,\dots,N(\omega)\}^{2|\mathbf{p}|},\mathbf{q}\in\{1,\dots,N(\omega)\}^{2|\mathbf{q}|},\tau\in\mathbb{R}}$, where the $\lambda_{\mathbf{p},\mathbf{q}}(\tau)$ are infinitely differentiable in τ and are rapidly decaying as functions of τ, $|\mathbf{p}|$ and $|\mathbf{q}|$. The convergence of $\mathcal{M}^\varepsilon$ in Proposition 3.1 is established in the space $\mathcal{C}([0,L],S'_M)$, where S'_M is the dual of the space S_M of the test sequences $\lambda = (\lambda_{\mathbf{p},\mathbf{q}})_{\mathbf{p}\in\{1,\dots,N(\omega)\}^{2|\mathbf{p}|},\mathbf{q}\in\{1,\dots,N(\omega)\}^{2|\mathbf{q}|}}$ which are rapidly decaying in $|\mathbf{p}|$ and $|\mathbf{q}|$.

3.3. Interpretation of the transport equations. We make the five following observations regarding the system of transport equations.

(1) By integrating the solution of the system of transport equations in τ, it is straightforward to see that the integral quantity is the solution of the system (3.1). This shows that we have

$$\mathcal{M}_{\mathbf{p},\mathbf{q}}(\omega,z) = \int \mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z)\,d\tau. \tag{3.12}$$

Therefore, the following remarks stated in terms of the family $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ hold true for the family of moments $\mathcal{M}_{\mathbf{p},\mathbf{q}}$ as well.

(2) Consider the set of moments $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ such that $|\mathbf{p}|-|\mathbf{q}|=c$ with c a nonzero integer. These moments form a closed subfamily with each member satisfying a zero initial condition. Therefore, these moments vanish and only moments having the same number of conjugated and unconjugated terms $|\mathbf{p}|=|\mathbf{q}|$ survive in the limit $\varepsilon\to 0$.

(3) Consider the case when

$$C_{jl}(z)\equiv 0 \quad \text{for} \quad j\neq l. \tag{3.13}$$

This corresponds to the situation where modes with different modal wavenumbers are not coupled. This is the case particularly when the inhomogeneities of the waveguide do not have lateral variations $\nu(\mathbf{x},z)=\nu(z)$. It then follows that

$$\widetilde{\Gamma}_{jl}(\omega) = \widetilde{\Gamma}_j^{(0)}(\omega)\mathbf{1}_{j=l}, \qquad \widetilde{\Gamma}_j^{(0)}(\omega) = \frac{k^4(\omega)}{2\beta_j^2(\omega)}\int_{-\infty}^{\infty}\mathbb{E}[\nu(0)\nu(s)]\,ds, \tag{3.14}$$

$$\Gamma_{jl}(\omega) = \Gamma_j^{(0)}(\omega)\mathbf{1}_{j=l}, \qquad \Gamma_j^{(0)}(\omega) = \frac{k^4(\omega)}{4\beta_j^2(\omega)}\int_0^{\infty}\mathbb{E}[\nu(0)\nu(s)]e^{i2\beta_j(\omega)s}\,ds. \tag{3.15}$$

This simplification gives $d_{jl}^{(1)}=0$, $d_{jl}^{(2)}=2\Gamma_j^{(0)}\mathbf{1}_{j=l}$, $d_{jl\tilde{j}\tilde{l}}^{(3)}=2\Gamma_j^{(0)}\mathbf{1}_{j=l=\tilde{j}=\tilde{l}}$, $d_{j\tilde{j}k}^{(4)}=2\widetilde{\Gamma}_j^{(0)}\mathbf{1}_{j=\tilde{j}=k}$, and $d_{k_1k_2}^{(5)}=0$. The analysis of the system shows that the solution has the form

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z) = \begin{cases} W_{p_1}^{(1)}*\cdots*W_{p_N}^{(N)}(\omega,\tau,z) \text{ if } \mathbf{p}=\mathbf{q}=\{(1,1)^{p_1},\dots,(N,N)^{p_N}\}, \\ 0 \text{ otherwise,}\end{cases}$$

where $*$ stands for the convolution in τ and for each $j=1,\dots,N$ the family $(W_p^{(j)})_{p\in\mathbb{N}}$ is the solution of the closed system of transport equations

$$\frac{\partial W_p^{(j)}}{\partial z} + 2p\beta'_j(\omega)\frac{\partial W_p^{(j)}}{\partial \tau} = 2p^2\Re\left[\Gamma_j^{(0)}(\omega)\right]\left(W_{p+1}^{(j)}+W_{p-1}^{(j)}-2W_p^{(j)}\right), \tag{3.16}$$

with the initial conditions $W_p^{(j)}(\omega,\tau,z=0)=\mathbf{1}_0(p)\delta(\tau)$. We therefore obtain that the backward and forward jth modes are uncoupled from the other modes, but their moments are coupled together according to the system that governs the propagation of one-dimensional waves in random media [1]. This is not qualitatively surprising, but this analysis shows that a sufficient criterion for this reduction is (3.13).

(4) If the two-point statistics of the process $\nu(\mathbf{x},z)$ are such that

$$\Gamma_{jl}(\omega)\equiv 0 \quad \text{for all} \quad j,l=1,\dots,N(\omega), \tag{3.17}$$

then $d^{(2)}=d^{(3)}=d^{(5)}=0$. Consequently there is coupling in the system of transport equations only for indices $(\mathbf{p},\mathbf{q})$ and $(\mathbf{p}',\mathbf{q}')$ such that $|\mathbf{p}|=|\mathbf{p}'|$ and $|\mathbf{q}|=|\mathbf{q}'|$. Since the initial conditions are zero for all nonempty indices $(\mathbf{p},\mathbf{q})$, the moments $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ are zero as soon as $|\mathbf{p}|$ or $|\mathbf{q}|$ is positive. In other words, $\mathcal{R}_{jl}^{\varepsilon}\to 0$ for all $j,l=1,\dots,N$ in distribution as $\varepsilon\to 0$. This shows that the forward scattering approximation is valid as soon as the condition (3.17) is fulfilled. This approximation is frequently used in the literature; it consists in neglecting coupling between forward- and backward-propagating modes, while retaining the coupling between forward-going modes and the implicit coupling to the evanescent modes. Here we give the necessary and sufficient condition (3.17) for the validity of this approximation.

(5) In the full system (2.18) we do not have“reciprocity” in that in general $\mathcal{R}_{jl}^{\varepsilon}\neq\mathcal{R}_{lj}^{\varepsilon}$ because of the coupling with the evanescent modes modeled by the matrices $\mathbf{G}$. However, the following symmetry relation is satisfied:

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}=\mathcal{W}_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}}$$

for $\tilde{\mathbf{p}}_n=(l_n,j_n)$ with $\mathbf{p}_n=(j_n,l_n)$ and $\tilde{\mathbf{q}}$ correspondingly defined. This means that reciprocity is satisfied in the limit $\varepsilon\to 0$, and this follows from the following observations:

- The initial condition in (3.10) depends on the multi-index only through $|\mathbf{p}|$ and $|\mathbf{q}|$.
- The coupling matrices $\mathbf{G}^{(aa)}$ and $\mathbf{G}^{(ab)}$ in (2.18) affect only the diagonal coefficients $D_{\mathbf{p},\mathbf{q}}$ in a symmetric way in the problem for $\mathcal{W}_{\mathbf{p},\mathbf{q}}$.
- We have the symmetry relations

$$\left(\mathbf{H}^{(aa)}\right)^T=-\overline{\mathbf{H}^{(aa)}},\quad \left(\mathbf{H}^{(ab)}\right)^T=\mathbf{H}^{(ab)}$$

in the coupling matrices in (2.18).

3.4. Enhanced backscattering. In this section we consider the case where the forward coupling is strong while the coupling between the forward- and backward-going modes is weak. As seen above, the forward scattering approximation consists in neglecting completely the latter coupling, and it is valid when the matrix Γ is zero. Here we assume that Γ is not zero but Γ is small compared to $\widetilde{\Gamma}$. This allows us to simplify significantly the system of transport equations and to present very interesting results. In particular, we show in the following proposition that the enhanced backscattering phenomenon extensively discussed in the physical literature can be exhibited from the particular structure of the reflected time-harmonic wave field.

We denote by $\mathcal{P}_{jl}$ the mean reflected power of the mode j when the input wave is a mode l:

$$\mathcal{P}_{jl}^{(1)}(\omega,z)=\mathcal{M}_{(j,l),(j,l)}(\omega,z)=\lim_{\varepsilon\to 0}\mathbb{E}[|\mathcal{R}_{jl}^{\varepsilon}(\omega,z)|^2],\qquad j,l=1,\dots,N(\omega),$$

for a random waveguide with length z.

Proposition 3.3. *If the matrix norm of $\Re[\Gamma(\omega)]$ is small compared to $1/L$ and the positive spectral gap* (3.20) *of the operator $\mathcal{L}_\omega$ in* (3.18) *below, which is defined in terms of $\widetilde{\Gamma}(\omega)$, is large compared to $1/L$, then the mean reflected mode powers are*

$$\mathcal{P}_{jl}^{(1)}(\omega, L) = \begin{cases} P_0(\omega) & \text{if } j \neq l, \\ 2P_0(\omega) & \text{if } j = l, \end{cases}$$

where $P_0(\omega)$ is given by

$$P_0(\omega) = \frac{2}{N(\omega)(N(\omega)+3)}\left[\sum_{j\neq l}\Re[\Gamma_{jl}(\omega)] + 2\sum_{j}\Re[\Gamma_{jj}(\omega)]\right]L.$$

The first condition "$\Re[\Gamma(\omega)L] \ll 1$" means that the coupling between backward- and forward-going modes is weak. The second condition about the spectral gap means that the coupling between forward-going modes is strong (as well as the coupling between backward-going modes). The most striking result of this proposition is that, if the incident wave is a pure mode l, then the mean reflected power of the lth mode $\mathcal{P}_{ll}^{(1)}$ is twice the mean reflected power of any other mode $\mathcal{P}_{jl}^{(1)}$, $j \neq l$.

First, this result shows that the reflected wave has a memory of the initial conditions. This is in contrast to the transmitted wave field in the same regime, where the equipartition of energy means that the mean transmitted mode powers acquire a uniform distribution over the modes, independently of the initial conditions (see section 4.4).

Second, since a mode corresponds to a particular wavevector angle, this result means that we observe a uniform mean reflected power in all outgoing directions, except in the backscattered direction (corresponding to the input one), where we observe twice as much power. The physical reason for this enhancement of backscattered power is the constructive interference between the direct and reverse paths in the backscattering direction. Enhanced backscattering was first predicted in three-dimensional random media in [2] and was detected by several groups [9, 15, 17]. It is also referred to as the weak localization effect. The most popular techniques amongst physicists for analyzing the weak localization effect, and more generally for taking into account interference effects, are based on diagrammatic expansions [16]. Here we give a mathematical derivation of this phenomenon in the context of random waveguides.

Proof. The initial conditions for the solution $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ of the system of transport equations is zero as soon as $|\mathbf{p}| > 0$ or $|\mathbf{q}| > 0$. Since the coupling terms from $|\mathbf{p}|$ to $|\mathbf{p}| \pm 1$ and from $|\mathbf{q}|$ to $|\mathbf{q}| \pm 1$ are proportional to $\Re(\Gamma)$, this shows that the only coefficients $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ of order $\Re(\Gamma)$ are the ones with $|\mathbf{p}| = 1$ and $|\mathbf{q}| = 1$. Up to terms of higher order, we find that

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z) = \begin{cases} \delta(\tau) & \text{if } \mathbf{p} = \mathbf{q} = \emptyset, \\ W_{jl}(\omega,\tau,z) & \text{if } \{\mathbf{p} = (j,l),\ \mathbf{q} = (j,l)\} \text{ or } \{\mathbf{p} = (j,l),\ \mathbf{q} = (l,j)\}, \\ 0 & \text{otherwise,} \end{cases}$$

where W_{jl} is the solution of

$$\frac{\partial W_{jl}}{\partial z} + \left[\beta_j'(\omega) + \beta_l'(\omega)\right]\frac{\partial W_{jl}}{\partial \tau} = (\mathcal{L}_\omega W)_{jl} + 2\Re\left[\Gamma_{jl}(\omega)\right]\delta(\tau),$$

with the initial conditions $W_{jl}(\omega,\tau,z) = 0$. Here $\mathcal{L}_\omega$ is the linear operator from $\mathbb{R}^{N\times N}$ into $\mathbb{R}^{N\times N}$ defined by

$$(3.18)\qquad (\mathcal{L}_\omega \mathcal{P})_{jl} = \begin{cases} \displaystyle\sum_{k\neq j} \tilde{\gamma}_{jk}(\mathcal{P}_{kl} - \mathcal{P}_{jl}) + \sum_{k\neq l} \tilde{\gamma}_{kl}(\mathcal{P}_{jk} - \mathcal{P}_{jl}) - 2\tilde{\gamma}_{jl}\mathcal{P}_{jl} & \text{if } j \neq l, \\ \displaystyle 2\sum_{k\neq j} \tilde{\gamma}_{jk}(2\mathcal{P}_{jk} - \mathcal{P}_{jj}) & \text{if } j = l, \end{cases}$$

where $\tilde{\gamma}_{jl}(\omega) = 2\Re[\widetilde{\Gamma}_{jl}(\omega)]$. By integrating in τ and using (3.12), we obtain the system for the mean reflected powers

$$(3.19)\qquad \frac{d\mathcal{P}^{(1)}_{jl}}{dz} = (\mathcal{L}_\omega \mathcal{P}^{(1)})_{jl} + 2\Re\big[\Gamma_{jl}(\omega)\big].$$

Interpreting $\mathcal{L}_\omega$ as an $N^2(\omega) \times N^2(\omega)$ matrix acting on $N^2(\omega)$-dimensional vectors, it is straightforward to check that the vector $\mathcal{P}^*(\omega)$ defined by

$$\mathcal{P}^*_{jl}(\omega) = \begin{cases} \dfrac{1}{\sqrt{N(\omega)(N(\omega)+3)}} & \text{if } j \neq l, \\ \dfrac{2}{\sqrt{N(\omega)(N(\omega)+3)}} & \text{if } j = l \end{cases}$$

is a unit eigenvector of $\mathcal{L}_\omega$ associated with the eigenvalue zero. Additionally, using the positivity of the matrix $\Re(\widetilde{\Gamma})$ and the Perron–Frobenius theorem, one can show that zero is a simple eigenvalue and all other eigenvalues are negative. Let us denote by $\lambda^{(N^2)}(\omega) \leq \cdots \leq \lambda^{(2)}(\omega) < 0$ these eigenvalues and by $\mathcal{Q}^{(N^2)}(\omega), \dots, \mathcal{Q}^{(2)}(\omega)$ the corresponding unit eigenvectors. The spectral gap mentioned in the proposition is $|\lambda^{(2)}(\omega)|$, which is also given by

$$(3.20)\qquad |\lambda^{(2)}(\omega)| = \inf_{\mathcal{P}\in\mathbb{R}^{N^2(\omega)},\, \langle \mathcal{P},\mathcal{P}^*(\omega)\rangle = 0} \frac{-\langle \mathcal{P}, \mathcal{L}_\omega \mathcal{P}\rangle}{\langle \mathcal{P},\mathcal{P}\rangle},$$

where $\langle \cdot,\cdot\rangle$ denotes the scalar product in $\mathbb{R}^{N^2(\omega)}$. The integration of (3.19) gives

$$(3.21)\qquad \mathcal{P}^{(1)}_{jl}(\omega,z) = \mathcal{P}^*_{jl}\,\langle \mathcal{P}^*, 2\Re(\Gamma)\rangle\, z + \sum_{k=2}^{N^2} \mathcal{Q}^{(k)}_{jl} \left\langle \mathcal{Q}^{(k)}, 2\Re(\Gamma)\right\rangle \frac{\exp(\lambda^{(k)} z) - 1}{\lambda^{(k)}}.$$

If $|\lambda^{(2)}|z$ is much larger than 1, then the first term of the right-hand side is much larger than the other terms. This gives the desired result. □

3.5. Fluctuation theory for the reflected mode powers. The previous section describes the mean reflected powers. It is important to study the fluctuations of the reflected powers in order to predict under which conditions the enhanced backscattering can be observed. Propositions 3.1–3.2 allow us to study the fluctuations of the reflected mode powers by looking at their second moments:

$$\mathcal{P}^{(2)}_{jl,mn}(\omega,z) = \lim_{\varepsilon\to 0} \mathbb{E}\left[|\mathcal{R}^\varepsilon_{jl}(\omega,z)|^2 |\mathcal{R}^\varepsilon_{mn}(\omega,z)|^2\right].$$

We investigate the asymptotic correlation matrix (of size $N^2(\omega) \times N^2(\omega)$) of the reflected mode powers:

$$\begin{aligned} \mathrm{Cor}_{jl,mn}(\omega) &= \lim_{\varepsilon\to 0} \frac{\mathbb{E}\left[|\mathcal{R}^\varepsilon_{jl}(\omega,L)|^2 |\mathcal{R}^\varepsilon_{mn}(\omega,L)|^2\right] - \mathbb{E}\Big[|\mathcal{R}^\varepsilon_{jl}(\omega,L)|^2\Big]\mathbb{E}\Big[|\mathcal{R}^\varepsilon_{mn}(\omega,L)|^2\Big]}{\mathbb{E}\Big[|\mathcal{R}^\varepsilon_{jl}(\omega,L)|^2\Big]\mathbb{E}\Big[|\mathcal{R}^\varepsilon_{mn}(\omega,L)|^2\Big]} \\ &= \frac{\mathcal{P}^{(2)}_{jl,mn}(\omega,L) - \mathcal{P}^{(1)}_{jl}(\omega,L)\mathcal{P}^{(1)}_{mn}(\omega,L)}{\mathcal{P}^{(1)}_{jl}(\omega,L)\mathcal{P}^{(1)}_{mn}(\omega,L)}. \end{aligned}$$

Proposition 3.4. *If the matrix norm of $\Re[(\Gamma(\omega)]$ is small compared to $1/L$ and the positive spectral gap of the operator $\mathcal{L}_\omega^{(2)}$ given in* 1–7 *below, which is defined in terms of $\widetilde{\Gamma}(\omega)$, is large compared to $1/L$, then the second moments of the reflected mode powers satisfy*

$$\lim_{N(\omega)\to\infty} \mathrm{Cor}_{jl,mn}(\omega) = \begin{cases} 0 \textit{ if } (j,l) \neq (m,n) \textit{ and } (j,l) \neq (n,m), \\ 1 \textit{ if } (j,l) = (m,n) \textit{ or } (j,l) = (n,m). \end{cases}$$

The result for $(j,l) \neq (m,n)$ shows that the reflected mode powers are asymptotically uncorrelated as $N \to \infty$. The result for $(j,l) = (m,n)$ shows that they are not statistically stable quantities as $N \to \infty$, since their normalized variances are equal to one. This means that the fluctuations of the reflected mode powers are of the same order as their mean values. This implies that it is necessary to perform an averaging in order to observe the enhanced backscattering. This averaging can be done by a summation of the reflected mode powers over different experiments with different realizations of the random medium, or with the same realization of the random medium but with different frequencies of the input monochromatic wave.

Another interesting point is that the normalized variances of the background reflected powers (i.e., $\mathrm{Cor}_{jl,jl}$ for $j \neq l$) are asymptotically equal to one and equal to the normalized variance of the backscattered reflected power (i.e., $\mathrm{Cor}_{jj,jj}$).

Proof. We apply the same strategy as in the proof of Proposition 3.3. Once again, the fundamental argument is that the coupling terms from $|\mathbf{p}|$ to $|\mathbf{p}| \pm 1$ are proportional to $\Re(\Gamma)$. Therefore, the lowest order terms in $\Re(\Gamma)$ of the coefficients $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ with $|\mathbf{p}| = |\mathbf{q}| = 2$ are

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z) = \begin{cases} W_{jl,mn}(\omega,\tau,z) \text{ if } & \begin{array}{l} \mathbf{p} = \{(j,l),(m,n)\} \text{ and } \mathbf{q} = \{(j,l),(m,n)\} \\ \text{or } \{(l,j),(m,n)\} \text{ or } \{(j,l),(n,m)\} \\ \text{or } \{(l,j),(n,m)\}, \end{array} \\ 0 \text{ otherwise,} & \end{cases}$$

where $W_{jl,mn}$ is the solution of

$$\frac{\partial W_{jl,mn}}{\partial z} + (\beta_j' + \beta_l' + \beta_m' + \beta_n') \frac{\partial W_{jl,mn}}{\partial \tau} = (\mathcal{L}_\omega^{(2)} W)_{jl,mn} + 2\Re(\Gamma_{jl}) W_{mn} + 2\Re(\Gamma_{mn}) W_{jl},$$

with the initial conditions $W_{jl,mn}(\omega,\tau,z=0) = 0$. Here $\mathcal{L}_\omega^{(2)}$ is the linear operator from $\mathbb{R}^{N^2\times N^2}$ into $\mathbb{R}^{N^2\times N^2}$ defined by the following:

1. If $j = l = m = n$,

$$(\mathcal{L}_\omega^{(2)} \mathcal{P}^{(2)})_{jj,jj} = \sum_{k\neq j} \tilde{\gamma}_{jk} [16 \mathcal{P}^{(2)}_{jk,jj} - 4\mathcal{P}^{(2)}_{jj,jj}],$$

where $\tilde{\gamma}_{jl}(\omega) = 2\Re[\widetilde{\Gamma}_{jl}(\omega)]$.

2. If $j = l = m \neq n$,

$$(\mathcal{L}_\omega^{(2)} \mathcal{P}^{(2)})_{jj,jn} = \sum_{k\neq j} \tilde{\gamma}_{jk} [4\mathcal{P}^{(2)}_{jk,jn} - 2\mathcal{P}^{(2)}_{jj,jn}] + \sum_{k\neq j} \tilde{\gamma}_{jk} [\mathcal{P}^{(2)}_{jj,kn} - \mathcal{P}^{(2)}_{jj,jn}]$$
$$+ \sum_{k\neq m} \tilde{\gamma}_{nk} [\mathcal{P}^{(2)}_{jj,jk} - \mathcal{P}^{(2)}_{jj,jn}] - 6\tilde{\gamma}_{jn} \mathcal{P}^{(2)}_{jj,jn}.$$

A formula of the same form holds true if $j = l = n \neq m$ or $j \neq l = m = n$ or $l \neq j = m = n$.

3. If $j = l \neq m = n$,

$$(\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jj,nn} = \sum_{k\neq j}\tilde{\gamma}_{jk}[4\mathcal{P}^{(2)}_{jk,nn} - 2\mathcal{P}^{(2)}_{jj,nn}] + \sum_{k\neq n}\tilde{\gamma}_{nk}[4\mathcal{P}^{(2)}_{jj,kn} - 2\mathcal{P}^{(2)}_{jj,nn}].$$

4. If $j = m \neq l = n$,

$$(\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jl,jl} = \sum_{k\neq j}\tilde{\gamma}_{jk}[4\mathcal{P}^{(2)}_{kl,jl} - 2\mathcal{P}^{(2)}_{jl,jl}] + \sum_{k\neq l}\tilde{\gamma}_{lk}[4\mathcal{P}^{(2)}_{jk,jl} - 2\mathcal{P}^{(2)}_{jl,jl}] - 4\tilde{\gamma}_{jl}\mathcal{P}^{(2)}_{jl,jl}.$$

A formula of the same form holds true if $j = n \neq l = m$.

5. If $j = l \neq m \neq n$,

$$\begin{aligned}(\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jj,mn} = {} & \sum_{k\neq j}\tilde{\gamma}_{jk}[4\mathcal{P}^{(2)}_{jk,mn} - 2\mathcal{P}^{(2)}_{jj,mn}] + \sum_{k\neq m}\tilde{\gamma}_{mk}[\mathcal{P}^{(2)}_{jj,kn} - \mathcal{P}^{(2)}_{jj,mn}] \\ & + \sum_{k\neq n}\tilde{\gamma}_{nk}[\mathcal{P}^{(2)}_{jj,mk} - \mathcal{P}^{(2)}_{jj,mn}] - 2\tilde{\gamma}_{mn}\mathcal{P}^{(2)}_{jj,mn}.\end{aligned}$$

A formula of the same form holds true if $m = n \neq j \neq l$.

6. If $j = m \neq l \neq n$,

$$\begin{aligned}(\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jl,jn} = {} & \sum_{k\neq j}\tilde{\gamma}_{jk}[\mathcal{P}^{(2)}_{kl,jn} - \mathcal{P}^{(2)}_{jl,jn}] + \sum_{k\neq l}\tilde{\gamma}_{lk}[\mathcal{P}^{(2)}_{jk,jn} - \mathcal{P}^{(2)}_{jl,jn}] \\ & + \sum_{k\neq j}\tilde{\gamma}_{jk}[\mathcal{P}^{(2)}_{jl,kn} - \mathcal{P}^{(2)}_{jl,jn}] + \sum_{k\neq n}\tilde{\gamma}_{nk}[\mathcal{P}^{(2)}_{jl,jk} - \mathcal{P}^{(2)}_{jl,jn}] \\ & - 2\left[\tilde{\gamma}_{jl} + \tilde{\gamma}_{jn} + \tilde{\gamma}_{ln}\right]\mathcal{P}^{(2)}_{jl,jn}.\end{aligned}$$

A formula of the same form holds true if $j = n \neq l \neq m$ or $l = m \neq j \neq n$ or $l = n \neq j \neq m$.

7. In the other cases,

$$\begin{aligned}(\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jl,mn} = {} & \sum_{k\neq j}\tilde{\gamma}_{jk}[\mathcal{P}^{(2)}_{kl,mn} - \mathcal{P}^{(2)}_{jl,mn}] + \sum_{k\neq l}\tilde{\gamma}_{lk}[\mathcal{P}^{(2)}_{jk,mn} - \mathcal{P}^{(2)}_{jl,mn}] \\ & + \sum_{k\neq m}\tilde{\gamma}_{mk}[\mathcal{P}^{(2)}_{jl,kn} - \mathcal{P}^{(2)}_{jl,mn}] + \sum_{k\neq n}\tilde{\gamma}_{nk}[\mathcal{P}^{(2)}_{jl,mk} - \mathcal{P}^{(2)}_{jl,mn}] \\ & - 2\left[\tilde{\gamma}_{jl} + \tilde{\gamma}_{mn}\right]\mathcal{P}^{(2)}_{jl,mn}.\end{aligned}$$

By integrating in τ, we find the system for the second moments of the reflected powers:

$$\frac{d\mathcal{P}^{(2)}_{jl,mn}}{dz} = (\mathcal{L}_\omega^{(2)}\mathcal{P}^{(2)})_{jl,mn} + 2\Re(\Gamma_{jl})\mathcal{P}^{(1)}_{mn} + 2\Re(\Gamma_{mn})\mathcal{P}^{(1)}_{jl}. \tag{3.22}$$

Interpreting $\mathcal{L}_\omega^{(2)}$ as an $N^4(\omega) \times N^4(\omega)$ matrix acting on $N^4(\omega)$-dimensional vectors, it is possible to check that the vector $\mathcal{P}^{(2),*}(\omega)$ defined by

$$\mathcal{P}^{(2),*}_{jl,mn} = \frac{1}{\sqrt{T^*}}\begin{cases} 8 & \text{if } j = l = m = n, \\ 4 & \text{if } j = l \neq m = n, \\ 2 & \text{if } \left|\begin{array}{l} j = m \neq l = n \text{ or } j = n \neq l = m \text{ or } j = l \neq m \neq n \\ \text{ or } j \neq l \neq m = n \text{ or } j = l = m \neq n \text{ or } j = l = n \neq m \\ \text{ or } j \neq l = m = n \text{ or } l \neq j = m = n, \end{array}\right. \\ 1 & \text{otherwise} \end{cases}$$

is the unit eigenvector of $\mathcal{L}_\omega^{(2)}$ associated with the eigenvalue zero, where $T^* = N^4+6N^3+15N^2+42N$. The other eigenvalues are negative. As a result, taking into account the expression in (3.21) for $\mathcal{P}^{(1)}$, the integration of (3.22) gives

$$\frac{1}{z^2}\mathcal{P}^{(2)}_{jl,mn}(\omega,z) \overset{z\to\infty}{\longrightarrow} \mathcal{P}^{(2),*}_{jl,mn} \sum_{\tilde{j},\tilde{l},\tilde{m},\tilde{n}=1}^{N} \mathcal{P}^{(2),*}_{\tilde{j}\tilde{l},\tilde{m}\tilde{n}} \left(\Re(\Gamma_{\tilde{j}\tilde{l}})\mathcal{P}^*_{\tilde{m}\tilde{n}} + \Re(\Gamma_{\tilde{m}\tilde{n}})\mathcal{P}^*_{\tilde{j}\tilde{l}}\right) \langle \mathcal{P}^*, 2\Re(\Gamma)\rangle\,,$$

which in turn gives the result of the proposition. □

4. The transmitted wave field.

We consider the problem of characterizing the distribution of the transmitted field.

4.1. The moments of the transmitted time-harmonic field.

We consider the time-harmonic transmitted field for a frequency ω. We can compute the limit of moments of the transmission matrix as in the case of the reflection matrix. We use the same notation as in Proposition 3.1.

Proposition 4.1. *We introduce the joint moments of elements of the reflection matrix along with a pair of elements of the transmission matrix:*

(4.1)
$$\mathcal{M}^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,z;j_1,j_2) = \mathbb{E}\left[\mathcal{T}^\varepsilon_{j_1l_1}(\omega,z)\overline{\mathcal{T}^\varepsilon_{j_2l_2}(\omega,z)} \prod_{(j,l)\in\mathbf{p}} \mathcal{R}^\varepsilon_{jl}(\omega,z) \prod_{(m,n)\in\mathbf{q}} \overline{\mathcal{R}^\varepsilon_{mn}(\omega,z)}\right]$$

for $\mathbf{t}=(l_1,l_2)$. The family of moments $\mathcal{M}^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}$ converges as $\varepsilon\to 0$ to the solution $\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ of the system

$$\frac{d\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}}{dz} = D^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} + \left[\mathcal{S}_\omega(\mathcal{M})\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} + \left[\mathcal{Z}_\omega(\mathcal{M})\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}\,, \tag{4.2}$$

with the initial conditions $\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,z=0;j_1,j_2) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\mathbf{1}_{j_1}(l_1)\mathbf{1}_{j_2}(l_2)$. The linear operator $\mathcal{Z}_\omega$ is defined by

$$\begin{aligned}
\left[\mathcal{Z}_\omega(\mathcal{M})\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} &= \sum_{k=1}^{N} d^{(4)}_{l_1kl_2}\mathcal{M}^{(k,k)}_{\mathbf{p},\mathbf{q}} \\
&+ \sum_{k_1,k_2=1}^{N} d^{(2)}_{k_1k_2}\mathcal{M}^{(k_1,k_1)}_{\mathbf{p}\cup\{(k_2,l_1)\},\mathbf{q}\cup\{(k_2,l_2)\}} + d^{(5)}_{k_1k_2}\mathcal{M}^{(k_2,k_1)}_{\mathbf{p}\cup\{(k_1,l_1)\},\mathbf{q}\cup\{(k_2,l_2)\}} \\
&- \sum_{(j,l)\in\mathbf{p}} d^{(6)}_{l_1j}\mathcal{M}^{(j,l_2)}_{\mathbf{p}|\{(j,l)|(l_1,l)\},\mathbf{q}} + d^{(6)}_{l_1l}\mathcal{M}^{(l,l_2)}_{\mathbf{p}|\{n|(j,l_1)\},\mathbf{q}} \\
&+ \sum_{(j,l)\in\mathbf{q}} \sum_{k=1}^{N} d^{(4)}_{jkl_1}\mathcal{M}^{(k,l_2)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(k,l)\}} + d^{(4)}_{lkl_1}\mathcal{M}^{(k,l_2)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k)\}} \\
&- \sum_{(j,l)\in\mathbf{q}} d^{(6)}_{jl_2}\mathcal{M}^{(l_1,j)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l_2,l)\}} + d^{(6)}_{ll_2}\mathcal{M}^{(l_1,l)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,l_2)\}} \\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{k=1}^{N} d^{(4)}_{jkl_2}\mathcal{M}^{(l_1,k)}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}} + d^{(4)}_{lkl_2}\mathcal{M}^{(l_1,k)}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}} \\
&- \sum_{(j,l)\in\mathbf{p}} d^{(2)}_{jl}\mathcal{M}^{(j,l_2)}_{\mathbf{p}|\{(j,l)|(l,l_1)\},\mathbf{q}} + d^{(5)}_{jl}\mathcal{M}^{(l,l_2)}_{\mathbf{p}|\{(j,l)|(j,l_1)\},\mathbf{q}}
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{(j,l)\in\mathbf{q}} d^{(2)}_{jl}\mathcal{M}^{(l_1,j)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l,l_2)\}} + d^{(5)}_{jl}\mathcal{M}^{(l_1,l)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,l_2)\}}\\
&+ \sum_{(j,l)\in\mathbf{q}} \sum_{k_1,k_2=1}^{N} d^{(2)}_{k_1k_2}\mathcal{M}^{(k_1,l_2)}_{\mathbf{p}\cup\{(k_2,l_1)\},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}}\\
&+ \sum_{(j,l)\in\mathbf{q}} \sum_{k_1,k_2=1}^{N} d^{(5)}_{k_1k_2}\mathcal{M}^{(k_2,l_2)}_{\mathbf{p}\cup\{(k_1,l_1)\},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}}\\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{k_1,k_2=1}^{N} d^{(2)}_{k_1k_2}\mathcal{M}^{(l_1,k_1)}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}\cup\{(k_2,l_2)\}}\\
&+ \sum_{(j,l)\in\mathbf{p}} \sum_{k_1,k_2=1}^{N} d^{(5)}_{k_1k_2}\mathcal{M}^{(l_1,k_2)}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}\cup\{(k_1,l_2)\}}.
\end{aligned}
$$

The coefficient $D^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)$ *is defined by*

$$
\begin{aligned}
D^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = D_{\mathbf{p},\mathbf{q}} &+ i\left(\kappa_{l_1}-\kappa_{l_2}\right) - \Gamma_{kl_1} - \overline{\Gamma_{kl_2}} - \sum_{k=1}^{N}\Big(\widetilde{\Gamma}_{kl_1} + \widetilde{\Gamma}_{l_2k} - 2\check{\Gamma}_{l_1l_2}\mathbf{1}_{l_1\neq l_2}\Big)\\
&- 2\sum_{(j,l)\in\mathbf{p}}\big(\check{\Gamma}_{jl_1}\mathbf{1}_{j\neq l_1} + \check{\Gamma}_{ll_1}\mathbf{1}_{l\neq l_1} - \check{\Gamma}_{jl_2}\mathbf{1}_{j\neq l_2} - \check{\Gamma}_{ll_2}\mathbf{1}_{l\neq l_2}\big)\\
&+ 2\sum_{(j,l)\in\mathbf{q}}\big(\check{\Gamma}_{jl_1}\mathbf{1}_{j\neq l_1} + \check{\Gamma}_{ll_1}\mathbf{1}_{l\neq l_1} - \check{\Gamma}_{jl_2}\mathbf{1}_{j\neq l_2} - \check{\Gamma}_{ll_2}\mathbf{1}_{l\neq l_2}\big).
\end{aligned}
$$

The coefficient $d^{(6)}(\omega)$ *is given by*

$$
d^{(6)}_{jl}(\omega) = 2\Re\big[\widetilde{\Gamma}_{jl}(\omega)\big].
$$

The linear operator $\mathcal{S}_\omega$ *and the coefficients* $D_{\mathbf{p},\mathbf{q}}(\omega)$*,* $\kappa_l(\omega)$*, and* $d^{(j)}(\omega)$*,* $j=1,\ldots,5$*, are defined in Proposition* 3.1*.*

4.2. Transmission transport equations. We consider here the two-frequency statistics of the transmitted field. We have the following result that is proved in Appendix B.

PROPOSITION 4.2. *We introduce the joint moments of elements of the reflection matrix at two nearby frequencies along with a pair of elements of the transmission matrix:*

$$
\begin{aligned}
(4.3)\quad \mathcal{U}^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z;j_1,j_2) = \mathbb{E}\Bigg[&\mathcal{T}^{\varepsilon}_{j_1l_1}(\omega+\varepsilon^2h/2,z)\overline{\mathcal{T}^{\varepsilon}_{j_2l_2}(\omega-\varepsilon^2h/2,z)}\\
&\times\prod_{(j,l)\in\mathbf{p}}\mathcal{R}^{\varepsilon}_{jl}(\omega+\varepsilon^2h/2,z)\prod_{(m,n)\in\mathbf{q}}\overline{\mathcal{R}^{\varepsilon}_{mn}(\omega-\varepsilon^2h/2,z)}\Bigg]
\end{aligned}
$$

for $\mathbf{t}=(l_1,l_2)$*. The family of Fourier transforms*

$$
(4.4)\qquad \mathcal{W}^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2) = \frac{1}{2\pi}\int e^{-ih[\tau-\phi_{\mathbf{p},\mathbf{q}}(\omega)z]}\mathcal{U}^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z;j_1,j_2)\,dh
$$

converges as $\varepsilon\to0$ *to the solution* $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ *of the system of transport equations*

$$
(4.5)\qquad \frac{\partial\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)\frac{\partial\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}}{\partial\tau} = D^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} + \big[\mathcal{S}_\omega(\mathcal{W})\big]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} + \big[\mathcal{Z}_\omega(\mathcal{W})\big]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}},
$$

with the initial conditions $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z=0;j_1,j_2) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\mathbf{1}_{j_1}(l_1)\mathbf{1}_{j_2}(l_2)\delta(\tau)$. *The coefficient* $\phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)$ *is given by*

$$\phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega) = \phi_{\mathbf{p},\mathbf{q}}(\omega) + \frac{\beta'_{l_1}(\omega)+\beta'_{l_2}(\omega)}{2}. \tag{4.6}$$

This generalized set of transport equations describes accurately the transmitted wave field and is the key tool in analyzing various applications with wave propagation in random waveguides. The corresponding transport equations in the layered case are presented in [5].

4.3. Interpretation of the transmission transport equations. We make the following observations regarding the system of transport equations.

(1) By integrating the solution of the system of transport equations in τ, it is straightforward to see that the integral quantity is the solution of the system (4.2). This shows that we have

$$\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,z;j_1,j_2) = \int \mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2)\,d\tau. \tag{4.7}$$

Therefore, the following remarks stated in terms of the family $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ hold true for the family of moments $\mathcal{M}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ as well.

(2) Consider the set of moments $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ such that $|\mathbf{p}|-|\mathbf{q}|=c$ with c a nonzero integer. These moments form a closed subfamily with each member satisfying a zero initial condition. Therefore, these moments vanish, and again only moments having the same number of conjugated and unconjugated terms survive in the small ε limit.

(3) Consider the case (3.13) when $C_{jl}\equiv 0$ for $j\neq l$, as described under (3) in section 3.3. Recall that this corresponds to the situation where modes with different modal wavenumbers are not coupled, which is the case particularly when the inhomogeneities of the waveguide do not have lateral variations. The analysis of the system in (4.2) then shows that the solution has the form

$$\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2) = W^{(1)}_{p_1} * \cdots * W^{(l-1)}_{p_{l-1}} * V^{(l)}_{p_l} * W^{(l+1)}_{p_{l+1}} * \cdots * W^{(N)}_{p_N}(\omega,\tau,z)$$

if $\mathbf{t}=(j_1,j_2)=(l,l)$ and $\mathbf{p}=\mathbf{q}=\{(1,1)^{p_1},\dots,(N,N)^{p_N}\}$, and $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2)=0$ otherwise. For each j, $(W^{(j)}_p)_{p\in\mathbb{N}}$ is given by (3.16) and for each l, $(V^{(l)}_p)_{p\in\mathbb{N}}$ is the solution of the closed system of transport equations

$$\frac{\partial V^{(l)}_p}{\partial z} + (2p+1)\beta'_l(\omega)\frac{\partial V^{(l)}_p}{\partial \tau} = 2\Re\big[\Gamma^{(0)}_l(\omega)\big]\Big[(p+1)^2(V^{(l)}_{p+1}-V^{(l)}_p) + p^2(V^{(l)}_{p-1}-V^{(l)}_p)\Big],$$

with the initial conditions $V^{(l)}_p(\omega,\tau,z=0)=\mathbf{1}_0(p)\delta(\tau)$. We therefore obtain that the backward and forward jth modes are uncoupled from the other modes, but they are coupled together according to the system that governs the propagation of one-dimensional waves in random media [1].

4.4. Forward scattering approximation. To contrast with the fully coupled case discussed above, we address in this section the forward scattering approximation analyzed in detail in [5, 7]. As shown above, this approximation is valid when Γ is zero or very small (in the sense that $\Re(\Gamma)L \ll 1$). The system of transport equations of Proposition 4.2 can be dramatically simplified, since only the terms $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ with $\mathbf{p}=\mathbf{q}=\emptyset$ contribute at the leading order, and these terms satisfy a closed system of transport equations as described in the following proposition.

Proposition 4.3. *If $\Gamma(\omega) = \mathbf{0}$, then the (transformed) autocorrelation function of the transmission coefficients at two nearby frequencies,*

$$\mathcal{V}_{jl}^{\varepsilon}(\omega, \tau, z) = \frac{1}{2\pi} \int e^{-ih[\tau - \beta_l'(\omega) z]} \mathbb{E}\left[\mathcal{T}_{jl}^{\varepsilon}(\omega + \varepsilon^2 h/2, z) \overline{\mathcal{T}_{jl}^{\varepsilon}(\omega - \varepsilon^2 h/2, z)} \right] dh,$$

has a limit as $\varepsilon \to 0$:

$$\lim_{\varepsilon \to 0} \mathcal{V}_{jl}^{\varepsilon}(\omega, \tau, z) = \mathcal{V}_{jl}(\omega, \tau, z).$$

For any fixed $l \in \{1, \ldots, N(\omega)\}$, the subfamily $(\mathcal{V}_{jl}(\omega, \tau, z))_{j=1,\ldots,N(\omega)}$ is the solution of the system of transport equations

$$\frac{\partial \mathcal{V}_{jl}}{\partial z} + \beta_j'(\omega) \frac{\partial \mathcal{V}_{jl}}{\partial \tau} = \sum_{n \neq j} 2\Re\big[\widetilde{\Gamma}_{jn}(\omega)\big] \left(\mathcal{V}_{nl} - \mathcal{V}_{jl}\right), \tag{4.8}$$

with the initial conditions $\mathcal{V}_{jl}(\omega, \tau, z = 0) = \delta(\tau) \mathbf{1}_l(j)$. Here $\widetilde{\Gamma}(\omega)$ is given by (3.4).

Let us introduce the mean transmitted power of the mode j when the input wave is a mode l:

$$\mathcal{P}_{jl}^{(t)}(\omega, z) = \mathcal{M}_{\emptyset,\emptyset}^{(l,l)}(\omega, \tau, z; j, j) = \lim_{\varepsilon \to 0} \mathbb{E}\left[|\mathcal{T}_{jl}^{\varepsilon}(\omega, z)|^2 \right].$$

The next proposition shows the equipartition of energy of the transmitted wave.

Proposition 4.4. *If $\Gamma(\omega) = \mathbf{0}$, then the mean transmitted powers converge to the uniform distribution, that is,*

$$\mathcal{P}_{jl}^{(t)}(\omega, z) \overset{z \to \infty}{\longrightarrow} \frac{1}{N(\omega)},$$

uniformly in j, l and exponentially in z.

Proof. By integrating (4.8) in τ, we get that, for any fixed l, the subfamily $(\mathcal{P}_{jl}^{(t)}(\omega, z))_{j=1,\ldots,N(\omega)}$ is the solution of the linear system

$$\frac{\partial \mathcal{P}_{jl}^{(t)}}{\partial z} = \sum_{n=1}^{N(\omega)} \mathcal{L}_{jn}^{(t)}(\omega) \mathcal{P}_{nl}^{(t)}, \tag{4.9}$$

starting from $\mathcal{P}_{jl}^{(t)}(\omega, z = 0) = \mathbf{1}_l(j)$. Here $\mathcal{L}^{(t)}(\omega)$ is the $N(\omega) \times N(\omega)$ matrix

$$\mathcal{L}_{jn}^{(t)}(\omega) = \begin{cases} 2\Re\big[\widetilde{\Gamma}_{jn}(\omega)\big] & \text{if } j \neq n, \\ -2 \sum_{m \neq j} 2\Re\big[\widetilde{\Gamma}_{jm}(\omega)\big] & \text{if } j = n. \end{cases}$$

Using the positivity of the coefficients $\Re[\widetilde{\Gamma}_{jn}]$ and the Perron–Frobenius theorem, we find that the matrix $\mathcal{L}^{(t)}$ has zero as an isolated eigenvalue, and all other eigenvalues are negative. It is straightforward to check that the eigenvector corresponding to the zero eigenvalue is the uniform vector, which establishes the proposition. □

Appendix A. Derivation of channel reflection-transport equations.

A.1. Propagator equations. We prove here Proposition 3.2. Note first that we can write the first equation in (2.18) in the form

$$\frac{d}{dz} \boldsymbol{\mathcal{R}}^{\varepsilon} = -\boldsymbol{\Phi}^{\varepsilon} \overline{\mathbf{H}^{a,\varepsilon}} + \boldsymbol{\mathcal{R}}^{\varepsilon} \overline{\boldsymbol{\Phi}^{\varepsilon}} \mathbf{H}^{a,\varepsilon} \boldsymbol{\mathcal{R}}^{\varepsilon} + \mathbf{H}^{a,\varepsilon} \boldsymbol{\mathcal{R}}^{\varepsilon} - \boldsymbol{\mathcal{R}}^{\varepsilon} \overline{\mathbf{H}^{a,\varepsilon}}, \tag{A.1}$$

where $\mathbf{\Phi}^\varepsilon(\omega, z)$ is the $N(\omega) \times N(\omega)$ diagonal matrix with diagonal entries:

$$\Phi^\varepsilon_{jj}(\omega, z) = e^{-2i\beta_j(\omega)z/\varepsilon^2}.$$

Our objective is now to compute cross moments of reflection matrix entries using diffusion approximation, and we remark that the phase factors in $\mathbf{\Phi}^\varepsilon$ then act as decoupling terms, decoupling the entries in (A.1). We introduce the quantities $U^\varepsilon_{\mathbf{p},\mathbf{q}}$ that give high-order products of elements $\mathcal{R}^\varepsilon_{jl}$ of the reflection matrix at two nearby frequencies:

$$U^\varepsilon_{\mathbf{p},\mathbf{q}}(\omega, h, z) = \prod_{(j,l)\in\mathbf{p}} \mathcal{R}^\varepsilon_{jl}(\omega + \varepsilon^2 h/2, z) \prod_{(m,n)\in\mathbf{q}} \overline{\mathcal{R}^\varepsilon_{mn}(\omega - \varepsilon^2 h/2, z)}. \tag{A.2}$$

It now follows from (2.18) that the $U^\varepsilon_{\mathbf{p},\mathbf{q}}$'s solve evolution equations of the form

$$\frac{\partial U^\varepsilon_{\mathbf{p},\mathbf{q}}}{\partial z} = \left[\mathcal{H}^\varepsilon_U(U^\varepsilon)\right]_{\mathbf{p},\mathbf{q}}. \tag{A.3}$$

Here $\left[\mathcal{H}^\varepsilon_U(U^\varepsilon)\right]_{\mathbf{p},\mathbf{q}}$ is a finite sum of $U^\varepsilon_{\mathbf{p}^{(1)},\mathbf{q}^{(1)}}, \dots, U^\varepsilon_{\mathbf{p}^{(m)},\mathbf{q}^{(m)}}$, where the multi-indices $\mathbf{p}^{(1)}, \mathbf{q}^{(1)}, \dots, \mathbf{p}^{(N)}, \mathbf{q}^{(N)}$ are obtained from $\mathbf{p}$ and $\mathbf{q}$ by one or two replacements. We have explicitly

$$\begin{aligned}\left[\mathcal{H}^\varepsilon_U(U^\varepsilon)\right]_{\mathbf{p},\mathbf{q}} = &\sum_{(j,l)\in\mathbf{p}} U^\varepsilon_{\mathbf{p}|(j,l),\mathbf{q}} \\ &\times \left\{ H^{b,\varepsilon}_{jl} - \sum_{k_1,k_2=1}^{N} \mathcal{R}^\varepsilon_{jk_1} \overline{H^{b,\varepsilon}_{k_1k_2}} \mathcal{R}^\varepsilon_{k_2l} + \sum_{k=1}^{N} \left[H^{a,\varepsilon}_{jk} \mathcal{R}^\varepsilon_{kl} - \mathcal{R}^\varepsilon_{jk} \overline{H^{a,\varepsilon}_{kl}} \right] \right\}_{\omega+h\varepsilon^2/2} \\ &\quad + \sum_{(j,l)\in\mathbf{q}} U^\varepsilon_{\mathbf{p},\mathbf{q}|(j,l)} \\ &\overline{\times \left\{ H^{b,\varepsilon}_{jl} - \sum_{k_1,k_2=1}^{N} \mathcal{R}^\varepsilon_{jk_1} \overline{H^{b,\varepsilon}_{k_1k_2}} \mathcal{R}^\varepsilon_{k_2l} + \sum_{k=1}^{N} \left[H^{a,\varepsilon}_{jk} \mathcal{R}^\varepsilon_{kl} - \mathcal{R}^\varepsilon_{jk} \overline{H^{a,\varepsilon}_{kl}} \right] \right\}}_{\omega-h\varepsilon^2/2},\end{aligned}$$

which can also be written as

$$\begin{aligned}\left[\mathcal{H}^\varepsilon_U(U^\varepsilon)\right]_{\mathbf{p},\mathbf{q}} = &\sum_{(j,l)\in\mathbf{p}} \Bigg\{ H^{b,\varepsilon}_{jl} U^\varepsilon_{\mathbf{p}|(j,l),\mathbf{q}} - \sum_{k_1,k_2=1}^{N} \overline{H^{b,\varepsilon}_{k_1k_2}} U^\varepsilon_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}} \\ &+ \sum_{k=1}^{N} \left[H^{a,\varepsilon}_{jk} U^\varepsilon_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}} - \overline{H^{a,\varepsilon}_{kl}} U^\varepsilon_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}} \right] \Bigg\}_{\omega+h\varepsilon^2/2} \\ &+ \sum_{(j,l)\in\mathbf{q}} \Bigg\{ \overline{H^{b,\varepsilon}_{jl}} U^\varepsilon_{\mathbf{p},\mathbf{q}|(j,l)} - \sum_{k_1,k_2=1}^{N} H^{b,\varepsilon}_{k_1k_2} U^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}} \\ &+ \sum_{k=1}^{N} \left[\overline{H^{a,\varepsilon}_{jk}} U^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(k,l)\}} - H^{a,\varepsilon}_{kl} U^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k)\}} \right] \Bigg\}_{\omega-h\varepsilon^2/2}.\end{aligned} \tag{A.4}$$

Next we observe that

$$
\begin{aligned}
H_{jl}^{a,\varepsilon}\,|_{\omega\pm\varepsilon^2 h/2} &\sim \alpha_{jl}^{\varepsilon}\,(\omega,h,z)\, e^{i(\beta_l(\omega)-\beta_j(\omega))z/\varepsilon^2} e^{\pm i(\beta_l'(\omega)-\beta_j'(\omega))zh/2}\\
&\equiv \alpha_{jl}^{\pm,\varepsilon}\,(\omega,h,z)\, e^{i(\beta_l(\omega)-\beta_j(\omega))z/\varepsilon^2},\\
H_{jl}^{b,\varepsilon}\,|_{\omega\pm\varepsilon^2 h/2} &\sim -\overline{\alpha_{jl}^{\varepsilon}\,(\omega,h,z)} e^{-i(\beta_l(\omega)+\beta_j(\omega))z/\varepsilon^2} e^{\mp i(\beta_l'(\omega)+\beta_j'(\omega))zh/2}\\
&\equiv \widetilde{\alpha}_{jl}^{\pm,\varepsilon}\,(\omega,h,z)\, e^{-i(\beta_l(\omega)+\beta_j(\omega))z/\varepsilon^2}
\end{aligned}
$$

as $\varepsilon \to 0$ for

$$
\begin{aligned}
\alpha_{jl}^{\varepsilon}(\omega,h,z) &= \frac{ik^2(\omega)}{2\varepsilon}\frac{C_{jl}\left(\frac{z}{\varepsilon^2}\right)}{\sqrt{\beta_j\beta_l(\omega)}}\\
&\quad + \frac{ik^4(\omega)}{4}\sum_{l'>N(\omega)}\int_{-\infty}^{\infty}\frac{C_{jl'}\left(\frac{z}{\varepsilon^2}\right)C_{ll'}\left(\frac{z}{\varepsilon^2}+s\right)}{\sqrt{\beta_j\beta_{l'}^2\beta_l(\omega)}} e^{i\beta_l(\omega)s-\beta_{l'}(\omega)|s|}\,ds.
\end{aligned}
$$

Using this notation, we get from (A.4)

$$
\begin{aligned}
\left[\mathcal{H}_U^{\varepsilon}(U^{\varepsilon})\right]_{\mathbf{p},\mathbf{q}} = \sum_{(j,l)\in\mathbf{p}} &\left\{\widetilde{\alpha}_{jl}^{+,\varepsilon} U^{\varepsilon}_{\mathbf{p}|(j,l),\mathbf{q}} e^{-i(\beta_l+\beta_j)z/\varepsilon^2}\right.\\
&+\sum_{k=1}^{N}\left[\alpha_{jk}^{+,\varepsilon}U^{\varepsilon}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}}e^{i(\beta_k-\beta_j)z/\varepsilon^2} - \overline{\alpha_{kl}^{+,\varepsilon}}U^{\varepsilon}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}}e^{i(\beta_k-\beta_l)z/\varepsilon^2}\right]\\
&\left.-\sum_{k_1,k_2=1}^{N}\overline{\widetilde{\alpha}_{k_1k_2}^{+,\varepsilon}}U^{\varepsilon}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}}e^{i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2}\right\}\\
&\qquad + \sum_{(j,l)\in\mathbf{q}}\left\{\overline{\widetilde{\alpha}_{jl}^{-,\varepsilon}}U^{\varepsilon}_{\mathbf{p},\mathbf{q}|(j,l)}e^{i(\beta_j+\beta_l)z/\varepsilon^2}\right.\\
&+\sum_{k=1}^{N}\left[\overline{\alpha_{jk}^{-,\varepsilon}}U^{\varepsilon}_{\mathbf{p},\mathbf{q}|\{(j,l)|(k,l)\}}e^{i(\beta_j-\beta_k)z/\varepsilon^2} - \alpha_{kl}^{-,\varepsilon}U^{\varepsilon}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k)\}}e^{i(\beta_l-\beta_k)z/\varepsilon^2}\right]\\
&\left.-\sum_{k_1,k_2=1}^{N}\widetilde{\alpha}_{k_1k_2}^{-,\varepsilon}U^{\varepsilon}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}}e^{-i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2}\right\},
\end{aligned}
\tag{A.5}
$$

where the β_j's are evaluated at ω.

A.2. The homogeneous propagator equations. In order the eliminate the h-dependence in the coefficients of (A.5), we now introduce the transformation

$$
V^{\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z) = \frac{1}{2\pi}\int e^{-ih[\tau-\phi_{\mathbf{p},\mathbf{q}}(\omega)z]}U^{\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z)\,dh,
\tag{A.6}
$$

where $\phi_{\mathbf{p},\mathbf{q}}(\omega)$ is given by (3.11). We then obtain from (A.5) that V^{ε} solves the infinite-dimensional system of partial differential equations

$$
\frac{\partial V^{\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi_{\mathbf{p},\mathbf{q}}(\omega)\frac{\partial V^{\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial \tau} = \left[\mathcal{H}_V^{\varepsilon}(V^{\varepsilon})\right]_{\mathbf{p},\mathbf{q}},
$$

with the initial conditions $V^{\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z=0) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\delta(\tau)$. The source term now has the form

$$\begin{aligned}\big[\mathcal{H}^\varepsilon_V(V^\varepsilon)\big]_{\mathbf{p},\mathbf{q}} = \sum_{(j,l)\in\mathbf{p}} \Bigg\{ & -\overline{\alpha^\varepsilon_{jl}} V^\varepsilon_{\mathbf{p}|(j,l),\mathbf{q}} e^{-i(\beta_j+\beta_l)z/\varepsilon^2} \\ & + \sum_{k_1,k_2=1}^{N} \alpha^\varepsilon_{k_1k_2} V^\varepsilon_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}} e^{i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2} \\ & + \sum_{k=1}^{N} \alpha^\varepsilon_{jk} V^\varepsilon_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}} e^{i(\beta_k-\beta_j)z/\varepsilon^2} \\ & - \sum_{k=1}^{N} \overline{\alpha^\varepsilon_{kl}} V^\varepsilon_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}} e^{i(\beta_k-\beta_l)z/\varepsilon^2} \Bigg\} \\ + \sum_{(j,l)\in\mathbf{q}} \Bigg\{ & -\alpha^\varepsilon_{jl} V^\varepsilon_{\mathbf{p},\mathbf{q}|(j,l)} e^{i(\beta_j+\beta_l)z/\varepsilon^2} \\ & + \sum_{k_1,k_2=1}^{N} \overline{\alpha^\varepsilon_{k_1k_2}} V^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}} e^{-i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2} \\ & + \sum_{k=1}^{N} \overline{\alpha^\varepsilon_{jk}} V^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(k,l)\}} e^{i(\beta_j-\beta_k)z/\varepsilon^2} \\ & - \sum_{k=1}^{N} \alpha^\varepsilon_{kl} V^\varepsilon_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k)\}} e^{i(\beta_l-\beta_k)z/\varepsilon^2} \Bigg\}, \end{aligned} \tag{A.7}$$

where the β_j's are evaluated at ω.

A.3. Transport equations. We next apply the diffusion approximation to get transport equations for the moments; see [5] for background material on and related to applications of the diffusion approximation theory. Observe that the function $\mathcal{H}^\varepsilon_V$ is linear and the random coefficients are rapidly fluctuating. Those coefficients whose amplitudes are of order ε^{-1} are centered and fluctuate on the scale ε^2; moreover, they are assumed to be rapidly mixing, giving a white-noise scaling situation. We can thus apply diffusion approximation results to obtain transport equations for the moments $\mathbb{E}[V^\varepsilon_{\mathbf{p},\mathbf{q}}]$ in the limit $\varepsilon \to 0$:

$$\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z) = \lim_{\varepsilon\to 0} \mathbb{E}[V^\varepsilon_{\mathbf{p},\mathbf{q}}(\omega,\tau,z)].$$

We then obtain from (A.7) that $\mathcal{W}_{\mathbf{p},\mathbf{q}}$ solves the infinite-dimensional system of partial differential equations

$$\frac{\partial \mathcal{W}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi_{\mathbf{p},\mathbf{q}}(\omega) \frac{\partial \mathcal{W}_{\mathbf{p},\mathbf{q}}}{\partial \tau} = i\Big[\sum_{(j,l)\in\mathbf{p}} \big(\kappa_j+\kappa_l\big) - \sum_{(j,l)\in\mathbf{q}} \big(\kappa_j+\kappa_l\big)\Big]\mathcal{W}_{\mathbf{p},\mathbf{q}} + \big[\mathcal{H}(\mathcal{W})\big]_{\mathbf{p},\mathbf{q}},$$

with the initial conditions $\mathcal{W}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z=0) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\delta(\tau)$, and where we defined $\kappa_l(\omega)$ (which is real) by (3.6). The source term now takes the form

$$\big[\mathcal{H}(\mathcal{W})\big]_{\mathbf{p},\mathbf{q}} = \sum_{k=1}^{6} \mathcal{I}_k, \tag{A.8}$$

and we next identify the coupling terms $\mathcal{I}_k$. We remark that in applying the diffusion approximation there is no coupling between terms that contain phase modulation of the type $\exp[i(\beta_j-\beta_l)z/\varepsilon^2]$ with terms that contain phase modulation of the type $\exp[i(\beta_m+\beta_n)z/\varepsilon^2]$ since the rapid phases then cannot cancel. There are eight terms in the expression for $\mathcal{H}_V^\varepsilon$ in (A.7); we label the first four associated with the multi-index $\mathbf{p}$ by $1_p, \dots, 4_p$ and the last by four by $1_q, \dots, 4_q$. First we consider the cross interaction of the terms 1_p and 2_p and also the corresponding combination 1_q and 2_q that is associated with complex conjugate coefficients. We label their contribution by the term $\mathcal{I}_1$, which is given by

$$\begin{aligned}\mathcal{I}_1 = &- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} 2\Re\big(\Gamma_{jl}\big)\Big(\mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(l,\tilde{l})\},\mathbf{q}} + \mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\},\mathbf{q}}\mathbf{1}_{j\neq l}\Big)\\ &- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} 2\Re\big(\Gamma_{\tilde{j}\tilde{l}}\big)\Big(\mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(\tilde{l},l)\},\mathbf{q}} + \mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(l,\tilde{j}),(\tilde{l},j)\},\mathbf{q}}\mathbf{1}_{\tilde{j}\neq\tilde{l}}\Big)\\ &- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} 2\Re\big(\Gamma_{jl}\big)\Big(\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},j),(l,\tilde{l})\}} + \mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\}}\mathbf{1}_{j\neq l}\Big)\\ &- \sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} 2\Re\big(\Gamma_{\tilde{j}\tilde{l}}\big)\Big(\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(j,\tilde{j}),(\tilde{l},l)\}} + \mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(l,\tilde{j}),(\tilde{l},j)\}}\mathbf{1}_{\tilde{j}\neq\tilde{l}}\Big)\\ &- \sum_{(j,l)\in\mathbf{p}}\sum_{k=1}^{N}\left(\Gamma_{jk}+\Gamma_{lk}\right)\mathcal{W}_{\mathbf{p},\mathbf{q}} - \sum_{(j,l)\in\mathbf{q}}\sum_{k=1}^{N}\left(\overline{\Gamma_{jk}}+\overline{\Gamma_{lk}}\right)\mathcal{W}_{\mathbf{p},\mathbf{q}},\end{aligned}$$

where Γ is defined by (3.5).

Next we consider the cross interaction of the terms 1_p and 2_p with the terms 1_q and 2_q. We label their contribution by the term $\mathcal{I}_2$, which is given by

$$\begin{aligned}\mathcal{I}_2 = &\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} 2\Re\big(\Gamma_{jl}\big)\mathcal{W}_{\mathbf{p}|(j,l),\mathbf{q}|(\tilde{j},\tilde{l})}\mathbf{1}_{(j,l)\cong(\tilde{j},\tilde{l})}\\ &+ \sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\sum_{k_1,k_2=1}^{N} 2\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k_1),(k_2,\tilde{l})\}}\\ &+ \sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\sum_{k_1,k_2=1}^{N} 2\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k_2),(k_1,\tilde{l})\}}\mathbf{1}_{k_1\neq k_2},\end{aligned}$$

where $(j,l) \cong (\tilde{j},\tilde{l})$ if $(j,l) = (\tilde{j},\tilde{l})$ or $(j,l) = (\tilde{l},\tilde{j})$.

We have completed the analysis of the terms associated with phase modulation of the form $\exp[i(\beta_j+\beta_l)z/\varepsilon^2]$ and consider now terms associated with phases of the form $\exp[i(\beta_j-\beta_l)z/\varepsilon^2]$. Consider first the interaction of the terms 3_p, 4_p, 3_q, and 4_q with themselves. We label this contribution by $\mathcal{I}_3$, which is given by

$$\begin{aligned}\mathcal{I}_3 = &-\Big\{2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}}\Big[\check{\Gamma}_{j\tilde{j}}+\check{\Gamma}_{l\tilde{l}}\Big] + \sum_{(j,l)\in\mathbf{p}}\sum_{k=1}^{N}\Big[\overline{\widetilde{\Gamma}_{jk}}+\overline{\widetilde{\Gamma}_{lk}}\Big]\Big\}\mathcal{W}_{\mathbf{p},\mathbf{q}}\\ &- 2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}}\Big[\mathbf{1}_{j\neq\tilde{j}}\Re\big(\widetilde{\Gamma}_{j\tilde{j}}\big)+\mathbf{1}_{l\neq\tilde{l}}\Re\big(\widetilde{\Gamma}_{l\tilde{l}}\big)\Big]\,\mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\},\mathbf{q}}\\ &-\Big\{2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}}\Big[\check{\Gamma}_{j\tilde{j}}+\check{\Gamma}_{l\tilde{l}}\Big] + \sum_{(j,l)\in\mathbf{q}}\sum_{k=1}^{N}\Big[\widetilde{\Gamma}_{jk}+\widetilde{\Gamma}_{lk}\Big]\Big\}\mathcal{W}_{\mathbf{p},\mathbf{q}}\\ &- 2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}}\Big[\mathbf{1}_{j\neq\tilde{j}}\Re\big(\widetilde{\Gamma}_{j\tilde{j}}\big)+\mathbf{1}_{l\neq\tilde{l}}\Re\big(\widetilde{\Gamma}_{l\tilde{l}}\big)\Big]\,\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{j},l),(j,\tilde{l})\}},\end{aligned}$$

where $\check{\Gamma}$ and $\widetilde{\Gamma}$ are defined by (3.3)–(3.4).

Next, we deal with the cross interaction between the terms 3_p and 4_p and correspondingly between 3_q and 4_q. We label this contribution by $\mathcal{I}_4$ and obtain

$$\begin{aligned}\mathcal{I}_4 = &-\Big\{\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{p}} 2\check{\Gamma}_{j\tilde{l}}\Big\}\mathcal{W}_{\mathbf{p},\mathbf{q}} - \sum_{(j,l)\in\mathbf{p}} 2\Re\big(\widetilde{\Gamma}_{jl}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(l,j)\},\mathbf{q}}\mathbf{1}_{j\neq l}\\
&-2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} \Re\big(\widetilde{\Gamma}_{j\tilde{l}}\big)\mathbf{1}_{j\neq\tilde{l}}\mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{l},l),(\tilde{j},j)\},\mathbf{q}}\\
&-2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{p}} \Re\big(\widetilde{\Gamma}_{\tilde{j}l}\big)\mathbf{1}_{\tilde{j}\neq l}\mathcal{W}_{\mathbf{p}|\{(j,l),(\tilde{j},\tilde{l})|(l,\tilde{l}),(j,\tilde{j})\},\mathbf{q}}\\
&-\Big\{\sum_{(j,l)\in\mathbf{q}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}} 2\check{\Gamma}_{j\tilde{l}}\Big\}\mathcal{W}_{\mathbf{p},\mathbf{q}} - \sum_{(j,l)\in\mathbf{q}} 2\Re\big(\widetilde{\Gamma}_{jl}\big)\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l,j)\}}\mathbf{1}_{j\neq l}\\
&-2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} \Re\big(\widetilde{\Gamma}_{j\tilde{l}}\big)\mathbf{1}_{j\neq\tilde{l}}\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(\tilde{l},l),(\tilde{j},j)\}}\\
&-2\sum_{\{(j,l),(\tilde{j},\tilde{l})\}\in\mathbf{q}} \Re\big(\widetilde{\Gamma}_{\tilde{j}l}\big)\mathbf{1}_{\tilde{j}\neq l}\mathcal{W}_{\mathbf{p},\mathbf{q}|\{(j,l),(\tilde{j},\tilde{l})|(l,\tilde{l}),(j,\tilde{j})\}}.\end{aligned}$$

Now we consider the cross interaction between the terms 3_p and 3_q and correspondingly between 4_p and 4_q. We label this contribution by $\mathcal{I}_5$ and obtain

$$\begin{aligned}\mathcal{I}_5 = &\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\left[2\check{\Gamma}_{\tilde{j}j}\mathcal{W}_{\mathbf{p},\mathbf{q}} + \sum_{k=1\neq j}^{N} 2\Re\left[\widetilde{\Gamma}_{jk}\right]\mathcal{W}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(k,\tilde{l})\}}\mathbf{1}_{j=\tilde{j}}\right]\\
&+\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\left[2\check{\Gamma}_{\tilde{l}l}\mathcal{W}_{\mathbf{p},\mathbf{q}} + \sum_{k=1\neq l}^{N} 2\Re\big(\widetilde{\Gamma}_{lk}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k)\}}\mathbf{1}_{l=\tilde{l}}\right].\end{aligned}$$

Finally, we analyze the cross interaction between the terms 3_p and 4_q and correspondingly between 4_p and 3_q. We label this contribution by $\mathcal{I}_6$ and obtain

$$\begin{aligned}\mathcal{I}_6 = &\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\left[2\check{\Gamma}_{j\tilde{l}}\mathcal{W}_{\mathbf{p},\mathbf{q}} + \sum_{k=1\neq j}^{N} 2\Re\big(\widetilde{\Gamma}_{jk}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(\tilde{j},k)\}}\mathbf{1}_{j=\tilde{l}}\right]\\
&+\sum_{(j,l)\in\mathbf{p}}\sum_{(\tilde{j},\tilde{l})\in\mathbf{q}}\left[2\check{\Gamma}_{l\tilde{j}}\mathcal{W}_{\mathbf{p},\mathbf{q}} + \sum_{k=1\neq l}^{N} 2\Re\big(\widetilde{\Gamma}_{lk}\big)\mathcal{W}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}|\{(\tilde{j},\tilde{l})|(k,\tilde{l})\}}\mathbf{1}_{l=\tilde{j}}\right].\end{aligned}$$

We can now assemble the terms in the source term $\mathcal{H}$ for the transport equation, and this completes the proof of Proposition 3.2.

Appendix B. Derivation of channel transmission-transport equations. We consider next the wave field that has been transmitted through the waveguide and develop a family of transport equations that generalize those we derived above for the characterization of the reflected field. The transmitted field can be characterized by the transmission operator in (2.17). Recall that the transmission and reflection matrices solve (2.18). In order to obtain a closed system of transport equations, we introduce the quantities

$$U^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z;j_1,j_2) = \mathcal{T}^{\varepsilon}_{j_1l_1}(\omega+\varepsilon^2h/2,z)\overline{\mathcal{T}^{\varepsilon}_{j_2l_2}(\omega-\varepsilon^2h/2,z)}U^{\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z)$$

for $\mathbf{t} = (l_1, l_2)$. Then we find, using (A.3),

$$\begin{aligned}\frac{\partial U^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial z} &= \left[\mathcal{H}^\varepsilon_U(U^{\mathbf{t},\varepsilon})\right]_{\mathbf{p},\mathbf{q}} \\ &\quad - U^\varepsilon_{\mathbf{p},\mathbf{q}}\left\{\overline{\mathcal{T}^\varepsilon_{j_2 l_2}}\right\}_{\omega-h\varepsilon^2/2}\left\{\sum_{k_1=1}^N \mathcal{T}^\varepsilon_{j_1k_1}\left(\overline{H^{a,\varepsilon}_{k_1l_1}}+\sum_{k_2=1}^N \overline{H^{b,\varepsilon}_{k_1k_2}}\mathcal{R}^\varepsilon_{k_2l_1}\right)\right\}_{\omega+h\varepsilon^2/2} \\ &\quad - U^\varepsilon_{\mathbf{p},\mathbf{q}}\left\{\mathcal{T}^\varepsilon_{j_1 l_1}\right\}_{\omega+h\varepsilon^2/2}\left\{\sum_{k_1=1}^N \overline{\mathcal{T}^\varepsilon_{j_2k_1}}\left(H^{a,\varepsilon}_{k_1l_2}+\sum_{k_2=1}^N H^{b,\varepsilon}_{k_1k_2}\overline{\mathcal{R}^\varepsilon_{k_2l_2}}\right)\right\}_{\omega-h\varepsilon^2/2},\end{aligned}$$

with $\mathcal{H}^\varepsilon_U$ defined in (A.5). We remark that the family of coefficients $U^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega, h, z; j_1, j_2)$ for fixed j_1 and j_2 form a closed subfamily, which allows us to rewrite the previous system as

$$\text{(B.1)}\qquad \frac{\partial U^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial z} = \left[\mathcal{H}^\varepsilon_U(U^{\mathbf{t},\varepsilon})\right]_{\mathbf{p},\mathbf{q}} + \left[\mathcal{H}^{\varepsilon,1}_U(U^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} + \left[\mathcal{H}^{\varepsilon,2}_U(U^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}},$$

$$\left[\mathcal{H}^{\varepsilon,1}_U(U^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = -\sum_{k=1}^N\left(\left\{\overline{H^{a,\varepsilon}_{kl_1}}\right\}_{\omega+h\varepsilon^2/2}U^{(k,l_2),\varepsilon}_{\mathbf{p},\mathbf{q}} + \left\{H^{a,\varepsilon}_{kl_2}\right\}_{\omega-h\varepsilon^2/2}U^{(l_1,k),\varepsilon}_{\mathbf{p},\mathbf{q}}\right),$$

$$\begin{aligned}\left[\mathcal{H}^{\varepsilon,2}_U(U^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = -\sum_{k_1,k_2=1}^N &\left(\left\{\mathcal{R}^\varepsilon_{k_2l_1}\overline{H^{b,\varepsilon}_{k_1k_2}}\right\}_{\omega+h\varepsilon^2/2}U^{(k_1,l_2),\varepsilon}_{\mathbf{p},\mathbf{q}}\right. \\ &\qquad\left. + \left\{\overline{\mathcal{R}^\varepsilon_{k_2l_2}}H^{b,\varepsilon}_{k_1k_2}\right\}_{\omega-h\varepsilon^2/2}U^{(l_1,k_1),\varepsilon}_{\mathbf{p},\mathbf{q}}\right).\end{aligned}$$

B.1. Homogeneous propagator equations in the transmission case. In order the eliminate the h-dependence in the coefficients of (B.1), we introduce the transformation

$$\text{(B.2)}\qquad V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2) = \frac{1}{2\pi}\int e^{-ih[\tau-\phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)z]}U^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,h,z;j_1,j_2)\,dh,$$

with $\phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)$ defined in (4.6). We then obtain from (B.1) that $V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}$ solves the infinite-dimensional system of partial differential equations

$$\text{(B.3)}\qquad \frac{\partial V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)\frac{\partial V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}}{\partial \tau} = \left[\widetilde{\mathcal{H}}^\varepsilon_V(V^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}},$$

with the initial conditions $V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z=0;j_1,j_2) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\mathbf{1}_{j_1}(l_1)\mathbf{1}_{j_2}(l_2)\delta(\tau)$. We decompose the source term as

$$\text{(B.4)}\qquad \widetilde{\mathcal{H}}^\varepsilon_V = \mathcal{H}^\varepsilon_V + \mathcal{H}^{\varepsilon,1}_V + \mathcal{H}^{\varepsilon,2}_V,$$

with $\mathcal{H}^\varepsilon_V$ defined in (A.7) and the specific transmission source terms given by

(B.5)

$$\left[\mathcal{H}^{\varepsilon,1}_V(V^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = -\sum_{k=1}^N\left[\overline{\alpha^\varepsilon_{kl_1}}V^{(k,l_2),\varepsilon}_{\mathbf{p},\mathbf{q}}e^{i(\beta_k-\beta_{l_1})z/\varepsilon^2} + \alpha^\varepsilon_{kl_2}V^{(l_1,k),\varepsilon}_{\mathbf{p},\mathbf{q}}e^{i(\beta_{l_2}-\beta_k)z/\varepsilon^2}\right],$$

$$\begin{aligned}\left[\mathcal{H}^{\varepsilon,2}_V(V^\varepsilon)\right]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = \sum_{k_1,k_2=1}^N &\left[\alpha^\varepsilon_{k_1k_2}V^{(k_1,l_2),\varepsilon}_{\mathbf{p}\cup\{(k_2,l_1)\},\mathbf{q}}e^{i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2}\right. \\ \text{(B.6)}\qquad &\left. + \overline{\alpha^\varepsilon_{k_1k_2}}V^{(l_1,k_1),\varepsilon}_{\mathbf{p},\mathbf{q}\cup\{(k_2,l_2)\}}e^{-i(\beta_{k_1}+\beta_{k_2})z/\varepsilon^2}\right],\end{aligned}$$

where the β_j's are evaluated at ω.

B.2. Transport equations. We now apply the diffusion approximation to get transport equations for the above modified moments that are relevant in the transmission case. That is, we deduce transport equations for the moments $\mathbb{E}[V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}]$ in the limit $\varepsilon \to 0$:

$$\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2) = \lim_{\varepsilon\to 0} \mathbb{E}[V^{\mathbf{t},\varepsilon}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z;j_1,j_2)].$$

We then obtain from (B.3) that $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$ solves the infinite-dimensional system of partial differential equations

$$\frac{\partial \mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}}{\partial z} + \phi^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega)\frac{\partial \mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}}{\partial \tau} = i\Big[\kappa_{l_1} - \kappa_{l_2} + \sum_{(j,l)\in\mathbf{p}}\big(\kappa_j+\kappa_l\big) - \sum_{(j,l)\in\mathbf{q}}\big(\kappa_j+\kappa_l\big)\Big]\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}$$
$$+ \big[\mathcal{H}(\mathcal{W}^{\mathbf{t}})\big]_{\mathbf{p},\mathbf{q}} + \big[\mathcal{H}^1(\mathcal{W})\big]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}},$$

with the initial conditions $\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}(\omega,\tau,z=0;j_1,j_2) = \mathbf{1}_0(|\mathbf{p}|)\mathbf{1}_0(|\mathbf{q}|)\mathbf{1}_{j_1}(l_1)\mathbf{1}_{j_2}(l_2)\delta(\tau)$. The source term $\mathcal{H}$ is defined in (A.8), and the specific transmission source term has the form

$$\big[\mathcal{H}^1(\mathcal{W})\big]^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} = \sum_{k=1}^{4}\widetilde{\mathcal{I}}_k, \tag{B.7}$$

and we next identify the coupling terms $\widetilde{\mathcal{I}}_k$.

First, we consider the terms that correspond to the interaction of the terms $\mathcal{H}^{\varepsilon,1}_V$ in (B.5) with themselves. This contribution is

$$\widetilde{\mathcal{I}}_1 = 2\sum_{k=1}^{N}\Re\big(\widetilde{\Gamma}_{kl_1}\big)\mathcal{W}^{(k,k)}_{\mathbf{p},\mathbf{q}}\mathbf{1}_{l_1=l_2} - \sum_{k=1}^{N}\Big[\widetilde{\Gamma}_{kl_1} + \widetilde{\Gamma}_{l_2k} - 2\check{\Gamma}_{l_1l_2}\Big]\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}}\mathbf{1}_{l_1\neq l_2}.$$

Then, we consider the cross interaction of the terms in $\mathcal{H}^{\varepsilon,2}_V$ in (B.5). This gives the contribution

$$\widetilde{\mathcal{I}}_2 = 2\sum_{k_1,k_2=1}^{N}\Re\big(\Gamma_{k_1k_2}\big)\Big[\mathcal{W}^{(k_1,k_1)}_{\mathbf{p}\cup\{(k_2,l_1)\},\mathbf{q}\cup\{(k_2,l_2)\}} + \mathcal{W}^{(k_2,k_1)}_{\mathbf{p}\cup\{(k_1,l_1)\},\mathbf{q}\cup\{(k_2,l_2)\}}\mathbf{1}_{k_2\neq k_1}\Big].$$

The terms in $\mathcal{H}^{\varepsilon,1}_V$ interact with those in $\mathcal{H}^{\varepsilon}_V$ having phase modulations of the form $\exp[i(\beta_j-\beta_l)z/\varepsilon^2]$, giving the following contribution to the diffusion approximation:

$$\begin{aligned}
\widetilde{\mathcal{I}}_3 = {} & -2\sum_{(j,l)\in\mathbf{p}}\big[\check{\Gamma}_{jl_1}\mathbf{1}_{j\neq l_1} + \check{\Gamma}_{ll_1}\mathbf{1}_{l\neq l_1} - \check{\Gamma}_{jl_2}\mathbf{1}_{j\neq l_2} - \check{\Gamma}_{ll_2}\mathbf{1}_{l\neq l_2}\big]\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} \\
& + 2\sum_{(j,l)\in\mathbf{q}}\big[\check{\Gamma}_{jl_1}\mathbf{1}_{j\neq l_1} + \check{\Gamma}_{ll_1}\mathbf{1}_{l\neq l_1} - \check{\Gamma}_{jl_2}\mathbf{1}_{j\neq l_2} - \check{\Gamma}_{ll_2}\mathbf{1}_{l\neq l_2}\big]\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} \\
& - 2\sum_{(j,l)\in\mathbf{p}}\Big[\Re\big(\widetilde{\Gamma}_{l_1j}\big)\mathcal{W}^{(j,l_2)}_{\mathbf{p}|\{(j,l)|(l_1,l)\},\mathbf{q}} + \Re\big(\widetilde{\Gamma}_{l_1l}\big)\mathcal{W}^{(l,l_2)}_{\mathbf{p}|\{(j,l)|(j,l_1)\},\mathbf{q}}\Big] \\
& + 2\sum_{(j,l)\in\mathbf{q}}\sum_{k=1}^{N}\Big[\Re\big(\widetilde{\Gamma}_{jk}\big)\mathcal{W}^{(k,l_2)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(k,l)\}}\mathbf{1}_{j=l_1} + \Re\big(\widetilde{\Gamma}_{lk}\big)\mathcal{W}^{(k,l_2)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,k)\}}\mathbf{1}_{l=l_1}\Big] \\
& - 2\sum_{(j,l)\in\mathbf{q}}\Big[\Re\big(\widetilde{\Gamma}_{jl_2}\big)\mathcal{W}^{(l_1,j)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l_2,l)\}} + \Re\big(\widetilde{\Gamma}_{ll_2}\big)\mathcal{W}^{(l_1,l)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,l_2)\}}\Big] \\
& + 2\sum_{(j,l)\in\mathbf{p}}\sum_{k=1}^{N}\Big[\Re\big(\widetilde{\Gamma}_{kj}\big)\mathcal{W}^{(l_1,k)}_{\mathbf{p}|\{(j,l)|(k,l)\},\mathbf{q}}\mathbf{1}_{j=l_2} + \Re\big(\widetilde{\Gamma}_{kl}\big)\mathcal{W}^{(l_1,k)}_{\mathbf{p}|\{(j,l)|(j,k)\},\mathbf{q}}\mathbf{1}_{l=l_2}\Big].
\end{aligned}$$

Finally, we consider the cross interaction of the terms in $\mathcal{H}_V^{\varepsilon,2}$ with those in $\mathcal{H}_V^{\varepsilon}$. This gives the contribution

$$\begin{aligned}
\widetilde{\mathcal{I}}_4 = & -\sum_{k=1}^{N}\Gamma_{kl_1}\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} - 2\sum_{(j,l)\in\mathbf{p}}\Re\big(\Gamma_{jl}\big)\Big[\mathcal{W}^{(j,l_2)}_{\mathbf{p}|\{(j,l)|(l,l_1)\},\mathbf{q}} + \mathcal{W}^{(l,l_2)}_{\mathbf{p}|\{(j,l)|(j,l_1)\},\mathbf{q}}\mathbf{1}_{j\neq l}\Big] \\
& -\sum_{k=1}^{N}\overline{\Gamma_{kl_2}}\mathcal{W}^{\mathbf{t}}_{\mathbf{p},\mathbf{q}} - 2\sum_{(j,l)\in\mathbf{q}}\Re\big(\Gamma_{jl}\big)\Big[\mathcal{W}^{(l_1,j)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(l,l_2)\}} + \mathcal{W}^{(l_1,l)}_{\mathbf{p},\mathbf{q}|\{(j,l)|(j,l_2)\}}\mathbf{1}_{j\neq l}\Big] \\
& + 2\sum_{(j,l)\in\mathbf{q}}\sum_{k_1,k_2=1}^{N}\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}^{(k_1,l_2)}_{\mathbf{p}\cup\{(k_2,l_1)\},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}} \\
& + 2\sum_{(j,l)\in\mathbf{q}}\sum_{k_1,k_2=1}^{N}\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}^{(k_2,l_2)}_{\mathbf{p}\cup\{(k_1,l_1)\},\mathbf{q}|\{(j,l)|(j,k_1),(k_2,l)\}}\mathbf{1}_{k_1\neq k_2} \\
& + 2\sum_{(j,l)\in\mathbf{p}}\sum_{k_1,k_2=1}^{N}\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}^{(l_1,k_1)}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}\cup\{(k_2,l_2)\}} \\
& + 2\sum_{(j,l)\in\mathbf{p}}\sum_{k_1,k_2=1}^{N}\Re\big(\Gamma_{k_1k_2}\big)\mathcal{W}^{(l_1,k_2)}_{\mathbf{p}|\{(j,l)|(j,k_1),(k_2,l)\},\mathbf{q}\cup\{(k_1,l_2)\}}\mathbf{1}_{k_1\neq k_2}.
\end{aligned}$$

We can now assemble the terms in the source term $\mathcal{H}^1$ for the transport equation, and this completes the proof of Proposition 4.2.

REFERENCES

[1] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[2] Y. N. Barabanenkov, *Wave corrections for the transfer equation for backward scattering*, Izv. Vyssh. Uchebn. Zaved. Radiofiz., 16 (1973), pp. 88–96.

[3] R. Burridge and G. Papanicolaou, *The geometry of coupled mode propagation in one-dimensional random media*, Comm. Pure Appl. Math., 25 (1972), pp. 715–757.

[4] L. B. Dozier and F. D. Tappert, *Statistics of normal mode amplitudes in a random ocean* I & II, J. Acoust. Soc. Am., 63 (1978), pp. 353–365, 533–547.

[5] J.-P. Fouque, J. Garnier, G. Papanicolaou, and K. Sølna, *Wave Propagation and Time Reversal in Randomly Layered Media*, Springer-Verlag, New York, 2007.

[6] J. Garnier, *The role of evanescent modes in randomly perturbed single-mode waveguides*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 455–472.

[7] J. Garnier and G. Papanicolaou, *Pulse propagation and time reversal in random waveguides*, SIAM J. Appl. Math., 67 (2007), pp. 1718–1739.

[8] W. Kohler and G. Papanicolaou, *Wave propagation in randomly inhomogeneous ocean*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Phys. 70, J. B. Keller and J. S. Papadakis, eds., Springer-Verlag, Berlin, 1977, pp. 153–223.

[9] Y. Kuga and A. Ishimaru, *Retroreflectance from a dense distribution of spherical particles*, J. Opt. Soc. Amer., 1 (1984), pp. 831–835.

[10] H. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press Ser. Signal Process. Optim. Control 6, MIT Press, Cambridge, MA, 1984.

[11] G. Labeyrie, F. de Tomasi, J.-C. Bernard, C. A. Müller, C. Miniatura, and R. Kaiser, *Coherent backscattering of light by atoms*, Phys. Rev. Lett., 83 (1999), pp. 5266–5269.

[12] G. Papanicolaou and S. Weinryb, *A functional limit theorem for waves reflected by a random medium*, Appl. Math. Optim., 30 (1994), pp. 307–334.

[13] H. E. Rowe, *Electromagnetic Propagation in Multi-Mode Random Media*, Wiley, New York, 1999.

[14] A. Tourin, A. Derode, P. Roux, B. A. van Tiggelen, and M. Fink, *Time-dependent coherent backscattering of acoustic waves*, Phys. Rev. Lett., 79 (1997), pp. 3637–3639.

[15] M. P. van Albada and A. Lagendijk, *Observation of weak localization of light in a random medium*, Phys. Rev. Lett., 55 (1985), pp. 2692–2695.

[16] M. C. W. van Rossum and Th. M. Nieuwenhuizen, *Multiple scattering of classical waves: Microscopy, mesoscopy, and diffusion*, Rev. Modern Phys., 71 (1999), pp. 313–371.

[17] P. E. Wolf and G. Maret, *Weak localization and coherent backscattering of photons in disordered media*, Phys. Rev. Lett., 55 (1985), pp. 2696–2699.

[18] K. M. Yoo, G. C. Tang, and R. R. Alfano, *Coherent backscattering of light from biological tissues*, Appl. Optics, 29 (1990), pp. 3237–3239.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1600–1617

COMPETITIVE EXCLUSION OF MICROBIAL SPECIES FOR A SINGLE NUTRIENT WITH INTERNAL STORAGE*

SZE-BI HSU† AND TING-HAO HSU†

Dedicated to Professor Hal Smith on the occasion of his 60*th birthday*

Abstract. We study a chemostat model that describes competition between n microbial species for a single-limited resource based on storage. The model incorporates internal resource storage variables that serve the direct connection between species growth and external resource availability. Mathematical analysis for the global dynamics of the model is carried out by using the fluctuating method. It is shown that the competitive exclusion principle holds for the limiting system of the model. The species with the smallest ambient nutrient concentration wins the competition. We extend the result of competitive exclusion in the paper [H. L. Smith and P. Waltmam, *SIAM J. Appl. Math.*, 54 (1994), pp. 1113–1131] from two species to n species.

Key words. chemostat, single-limited resource, competition, competitive exclusion, fluctuating lemma

AMS subject classification. 92A15

DOI. 10.1137/070700784

1. Introduction.

One of the basic hypotheses in the mathematical modeling of competition of microorganisms for a single-limited nutrient in a continuous culture [JM], [HHW], [T], [FS], [AM], [SW2] is that the rate of consumption of the nutrient and the rate of growth of the organism are directly proportional [M]: (rate of growth of organism)= y (rate of consumption of nutrient); y is called the yield constant and is determined over a finite period of time by

$$y = \frac{\text{weight of organism formed}}{\text{weight of the nutrient used}}.$$

In phytoplankton ecology, it has long been known that the yield can vary depending on the growth rate [D], [G1], [G2], [CM], [CN1], [CN2]. Droop [D] was the first to give a variable yield model or so-called *internal storage* model. He proposed the idea that the organism consumes the nutrient and converts the nutrient into internal storage (cell quota). When the internal storage is below the minimum cell quota, the organism ceases to grow. If the cell quota is above the minimum cell quota, then the growth rate increases with the cell quota. Furthermore, the nutrient uptake rate increases with nutrient concentration and decreases with cell quota. The model of growth with one limiting nutrient incorporating these relations has been tested in both constant and fluctuating environments [G3], [NG], [SC]. Thus the variable yield models are well supported experimentally.

In [SW1], the authors studied the competition between two species competing for a single-limited resource with internal storage. They applied the method of a monotone dynamical system [S] to show that the competitive exclusion principle holds.

*Received by the editors August 22, 2007; accepted for publication (in revised form) March 4, 2008; published electronically June 6, 2008. This work was supported by the National Council of Science, Republic of China, NSC 95-2115-M-007-008.

http://www.siam.org/journals/siap/68-6/70078.html

†Department of Mathematics, National Tsing-Hua University, Hsinchu 300, Taiwan (sbhsu@math.nthu.edu.tw, g933245@alumni.nthu.edu.tw).

When the number of species is greater than two, the method of the monotione dynamical system no longer works. In this paper we shall rigorously prove that the competitive exclusion principle also holds for the competition between n microbial species; $n \geq 2$ for a single-limited resource with internal storage. The result is similar to that of the classical simple chemostat model [HHW]: the species with the smallest ambient nutrient concentration wins the competition.

In section 2, we present the mathematical model and state the main results. In section 3 we give the proof of the main theorem. The main tools in the proof are the conservation principle, which allows the reduction of the $(2n+1)$-dimensional system of ordinary differential equations to a $(2n)$-dimensional system; the fluctuating method [HHG, WX], which provides tools to determine the global behavior of the $(2n)$-dimensional reduced system; and finally, results on the asymptotically autonomous system due to Thieme [Th], which show that the $(2n+1)$-dimensional system and the reduced $(2n)$-dimensional system have the same global asymptotic behavior. In section 4, we discuss the updated mathematical models of microorganisms competing for multiple nutrients in phytoplankton ecology. Several open problems are presented for future research.

2. The model and main result. The model of n species, $n \geq 2$, competing for a single-limited resource with internal storage in a chemostat, takes the form

$$\begin{aligned}
&S'(t) = (S^{(0)} - S(t))D - \sum_{i=1}^{n} x_i(t) f_i(S(t), Q_i(t)),\\
&x_i'(t) = [\mu_i(Q_i(t)) - D]x_i(t),\\
&Q_i'(t) = f_i(S(t), Q_i(t)) - \mu_i(Q_i(t))Q_i(t),\\
&S(0) \geq 0,\ x_i(0) > 0,\ Q_i(0) \geq Q_{\min,i}, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{2.1}$$

Here $S(t)$ denotes the concentration of the external limiting resource in the chemostat at time t, $x_i(t)$ denotes the concentration of species i at time t, $Q_i(t)$ represents the average amount of stored nutrient per cell of species i at time t, $\mu_i(Q_i)$ is the growth rate of species i as a function of cell quota Q_i, $f_i(S, Q_i)$ is the per capita uptake rate of species i as a function of resource concentration S and cell quota Q_i, $S^{(0)}$ is the input concentration, D is the dilution rate of the chemostat, and $Q_{\min,i}$ denotes the threshold cell quota below which no growth of species i occurs. The growth $\mu_i(Q_i)$ takes the forms [D, G1, G2, CN1, CN2]

$$\begin{aligned}
\mu_i(Q_i) &= \mu_{i\infty}\left(1 - \frac{Q_{\min,i}}{Q_i}\right),\\
\mu_i(Q_i) &= \mu_{i\infty}\frac{(Q_i - Q_{\min,i})_+}{K_i + (Q_i - Q_{\min,i})_+},
\end{aligned}$$

where $Q_{\min,i}$ is the minimum cell quota necessary to allow cell division, $(Q_i - Q_{\min,i})_+$ is the positive part of $(Q_i - Q_{\min,i})$, and $\mu_{i\infty}$ is the maximal growth rate of the species. According to Grover [G2],

$$\begin{aligned}
f_i(S, Q_i) &= \rho_i(Q_i)\frac{S}{a_i + S},\\
\rho_i(Q_i) &= \rho_{\max}^{\text{high}} - (\rho_{\max}^{\text{high}} - \rho_{\max}^{\text{low}})\frac{Q_i - Q_{\min,i}}{Q_{\max,i} - Q_{\min,i}},
\end{aligned}$$

where $Q_{\min,i} \le Q_i \le Q_{\max,i}$. Cunningham and Nisbet [CN1, CN2], Klausmeier and Litchman [KL], and Klausmeier, Litchman, and Levin [KLL] took $\rho_i(Q_i)$ to be a constant.

Motivated by these examples, we assumed that $\mu_i(Q_i)$ is defined and continuously differentiable for $Q_i \ge P_i > 0$ and satisfies

$$\mu_i(Q_i) \ge 0,\ \mu_i'(Q_i) > 0 \text{ and continuous for } Q_i \ge P_i,\ \mu_i(P_i) = 0. \tag{2.2}$$

In both examples above, $P_i = Q_{\min,i}$. We assume that $f_i(S, Q_i)$ is continuous differentiable for $S > 0$ and $Q_i \ge P_i$ and satisfies

$$f_i(0, Q_i) = 0,\ \frac{\partial f_i}{\partial S} > 0,\ \frac{\partial f_i}{\partial Q_i} \le 0. \tag{2.3}$$

In particular, $f_i(S, Q_i) > 0$ when $S > 0$.

From (2.2) and (2.3), it follows that $Q_i' \ge 0$ if $Q_i = P_i$ and the interval of Q_i values $[P_i, \infty)$ is positively invariant under the dynamics of (2.1). Therefore, we assume that the initial values satisfy

$$x_i(0) > 0,\ Q_i(0) \ge P_i,\ S(0) \ge 0,\ i = 1, 2, \ldots, n. \tag{2.4}$$

Assume that the equilibrium E takes the form

$$E = (S, x_1, Q_1, \ldots, x_n, Q_n).$$

Then we have the following steady states.

(i) The washout steady state

$$E_0 = (S^{(0)}, 0, Q_1^0, 0, Q_2^0, \ldots, 0, Q_n^0)$$

always exists. Here Q_i^0 is the unique solution of

$$f_i(S^{(0)}, Q_i) - Q_i \mu_i(Q_i) = 0. \tag{2.5}$$

(ii)

$$\begin{aligned}
E_1 &= (\lambda_1, x_1^*, Q_1^*, 0, \hat{Q}_2^1, 0, \hat{Q}_3^1, \ldots, 0, \hat{Q}_n^1),\\
E_2 &= (\lambda_2, 0, \hat{Q}_1^2, x_2^*, Q_2^*, 0, \hat{Q}_3^2, \ldots, 0, \hat{Q}_n^2),\\
&\ \ \vdots\\
E_n &= (\lambda_n, 0, \hat{Q}_1^n, 0, \hat{Q}_2^n, \ldots, 0, \hat{Q}_{n-1}^n, x_n^*, Q_n^*).
\end{aligned}$$

The equilibrium E_i corresponds to the presence of the ith population and the absence of the others. The parameters $\lambda_i, Q_i^*, x_i^*, \hat{Q}_j^i$, $j \ne i$, satisfy

$$\mu_i(Q_i^*) = D, \tag{2.6}$$

$$f_i(\lambda_i, Q_i^*) = \mu_i(Q_i^*) Q_i^* = D Q_i^*, \tag{2.7}$$

$$x_i^* = \frac{(S^{(0)} - \lambda_i) D}{f_i(\lambda_i, Q_i^*)} = \frac{S^{(0)} - \lambda_i}{Q_i^*}, \tag{2.8}$$

$$f_j(\lambda_j, \hat{Q}_j^i) = \mu_j(\hat{Q}_j^i) \hat{Q}_j^i, \quad j \ne i. \tag{2.9}$$

The steady state E_i exists if and only if the equation $\mu_i(Q_i) = D$ has a unique solution Q_i^* and

$$f_i(S^{(0)}, Q_i^*) > DQ_i^*.$$

Lemma 2.1. *The solutions $S(t), x_1(t), Q_1(t), \ldots, x_n(t), Q_n(t)$ of system* (2.1) *are positive and bounded for all $t \geq 0$. Furthermore,*

$$S(t) + \sum_{i=1}^{n} Q_i(t)x_i(t) = S^{(0)} + O(e^{-Dt}),\ t \to \infty, \tag{2.10}$$

and there exist $\gamma_i > P_i$, $t_0 > 0$ such that $Q_i(t) \geq \gamma_i$ for all $t \geq t_0$, for $i = 1, 2, \ldots, n$.

The above lemma is a statement that system (2.1) is as "well behaved" as one that is intuited from the biological problem. Equation (2.10) is the conservation principle. Therefore, all solutions of (2.1) asymptotically approach

$$S(t) + \sum_{i=1}^{n} Q_i(t)x_i(t) = S^{(0)} \tag{2.11}$$

as $t \to \infty$. Consequently, as a first step in the analysis of (2.1), we consider the restriction of (2.1) to the exponentially attracting invariant subset given by (2.11). Dropping S from (2.1) and letting $U_i = Q_i x_i$, $1 \leq i \leq n$, we obtain the following system:

$$\begin{aligned}
&U_i'(t) = f_i\left(S^{(0)} - \sum_{i=1}^{n} U_i(t),\ Q_i(t)\right)\frac{U_i(t)}{Q_i(t)} - DU_i(t),\\
&Q_i'(t) = f_i\left(S^{(0)} - \sum_{i=1}^{n} U_i(t),\ Q_i(t)\right) - \mu_i(Q_i(t))Q_i(t),\\
&U_i(0) > 0,\ Q_i(0) \geq P_i, \quad 1 \leq i \leq n,\ \sum_{i=1}^{n} U_i(0) \leq S^{(0)}.
\end{aligned} \tag{2.12}$$

We note that $U_i(t)$ is the total amount of stored nutrient of the ith species at time t. In the next section, we shall study the reduced limiting system (2.12). The relevant domain for (2.12) is

$$\Omega = \left\{(U_1, Q_1, \ldots, U_n, Q_n) \in \mathbb{R}^{2n} : \begin{array}{l} \sum_{i=1}^{n} U_i \leq S^{(0)},\ U_k \geq 0, \\ Q_k \geq P_k,\ k = 1, 2, \ldots, n \end{array}\right\}, \tag{2.13}$$

which is positively invariant under (2.12).

Lemma 2.2. *Let $(S(t), x_1(t), Q_1(t), \ldots, x_n(t), Q_n(t))$ be the system of* (2.1). *For $1 \leq i \leq n$, if one of the following cases holds:*

(i) *$\mu_i(Q_i) < D$ for all $Q_i \in [P_i, \infty)$,*

(ii) (2.6) *holds with $f_i(S, Q_i^*) < \mu_i(Q_i^*)Q_i^*$ for all $S \in [0, S^{(0)}]$,*

(iii) (2.6) *and* (2.7) *hold with $S^{(0)} < \lambda_i$,*

then

$$\lim_{t\to\infty} x_i(t) = 0.$$

In the first two cases, we denote $\lambda_i = +\infty$.

This lemma states that if the maximal growth rate of the ith organism is less than the dilution rate D or if the input concentration $S^{(0)}$ is too small, then the ith organism will die out as time becomes large. Note that the resulting behavior is competition independent.

Our basic hypothesis is

$$(\mathrm{H_n}) \qquad \begin{aligned} &0 < \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n, \\ &\lambda_1 < S^{(0)}. \end{aligned}$$

For an equilibrium $E = (S, x_1, Q_1, \ldots, x_n, Q_n)$ of system (2.1), we denote

$$\hat{E} = (U_1, Q_1, \ldots, U_n, Q_n)$$

as the corresponding equilibrium of system (2.12).

Lemma 2.3. *Let* $(\mathrm{H_n})$*hold; then the equilibrium* $\hat{E}_1$ *is locally asymptotically stable and the rest of the equilibria* $\hat{E}_0, \hat{E}_2, \ldots, \hat{E}_n$ *are saddles if they exist. Furthermore, if* $S^{(0)} > \lambda_i,\ i = 1, 2, \ldots, n,$ *then the stable manifolds of* $\hat{E}_0$ *and* $\hat{E}_k,\ k = 2, 3, \ldots, n,$ *are*

$$M^+(\hat{E}_0) = \{(0, Q_1, 0, Q_2, \ldots, 0, Q_n) : P_i < Q_i,\ i = 1, 2, \ldots, n\}$$

and

$$M^+(\hat{E}_k) = \left\{(0, Q_1, \ldots, 0, Q_{k-1}, U_k, Q_k, , \ldots, U_n, Q_n) : \begin{array}{c} P_i < Q_i,\ i = 1, 2, \ldots, n, \\ U_i > 0,\ i = k, k+1, \ldots, n \end{array}\right\}.$$

The following is our main theorem.

Theorem 2.4. *Let* $(\mathrm{H_n})$*hold. The solution of* (2.1) *satisfies*

$$\begin{aligned} &\lim_{t \to \infty} (S(t), x_1(t), Q_1(t), x_2(t), Q_2(t), \ldots, x_n(t), Q_n(t)) = E_1 \\ &= (\lambda_1, x_1^*, Q_1^*, 0, \hat{Q}_2^1, 0, \hat{Q}_3^1, \ldots, 0, \hat{Q}_n^1), \end{aligned}$$

where $Q_1^*,\ \lambda_1,\ x_1^*,\ \hat{Q}_j^1,\ j = 2, 3, \ldots, n,$ *satisfy*

$$\begin{aligned} &\mu_1(Q_1^*) = D, \\ &f_1(\lambda_1, Q_1^*) = DQ_1^*, \\ &x_1^* = \frac{S^{(0)} - \lambda_1}{Q_1^*}, \\ &f_j(\lambda_1, \hat{Q}_j^1) = \mu_j(\hat{Q}_j^1)\hat{Q}_j^1,\ j = 2, \ldots, n. \end{aligned}$$

This theorem states that under the hypothesis $(\mathrm{H_n})$only one species survives, the one with the lowest value of λ_i, and gives the limiting nutrient concentrations.

3. Proofs.

From the differential inequality [H2], the proof of Lemma 3.1 is easy, and so we omit it.

Lemma 3.1. *Let* $x : \mathbb{R}_+ \to [a, \infty),\ y : \mathbb{R}_+ \to [b, \infty),$ *and* $g : [a, \infty) \times [b, \infty) \to \mathbb{R}$ *be continuously differentiable and satisfy*

$$x'(t) \leq g(x(t), y(t)),\ t \geq 0.$$

Suppose

$$\frac{\partial g}{\partial x}(x,y) < 0,\ \frac{\partial g}{\partial y}(x,y) > 0,$$

and suppose that for each $y \in [b,\infty)$ there exists a unique solution $x^* = x^*(y) \in [a,\infty)$ of $g(x,y) = 0$. If $\limsup_{t\to\infty} y(t) \le \alpha$, then

$$\limsup_{t\to\infty} x(t) \le x^*(\alpha).$$

Proof of Lemma 2.1. From (2.2), (2.3), and (2.4), it is easy to verify that the solutions $S(t), Q_i(t), x_i(t)$, $1 \le i \le n$, are positive for all $t \ge 0$. The first equation of (2.1) gives

$$S' \le (S^{(0)} - S)D;$$

then obviously we have

$$\limsup_{t\to\infty} S(t) \le S^{(0)}. \tag{3.1}$$

For $i = 1, 2, \ldots, n$, consider the differential equation of Q_i in (2.1):

$$Q_i' = f_i(S, Q_i) - \mu_i(Q_i)Q_i.$$

From (2.2), (2.3), (3.1), and Lemma 3.1 it follows that

$$\limsup_{t\to\infty} Q_i(t) \le Q_i^0, \tag{3.2}$$

where $Q_i^0 > P_i$ is defined in (2.5).

Let $T = S + \sum_{i=1}^n Q_i x_i$. Then T satisfies

$$T' = (S^{(0)} - T)D.$$

Therefore,

$$T = S^{(0)} + O(e^{-Dt}) \text{ as } t \to \infty. \tag{3.3}$$

Thus the conservation principle (2.10) holds.

Next we show that there exist $\gamma_i > P_i$ and $t_0 > 0$ such that $Q_i(t) \ge \gamma_i$ for $t \ge t_0$. We show $S(t)$ is bounded below by a constant $\gamma > 0$. Let $U_i = x_i Q_i$. Rewrite the first equation in (2.1) as

$$S' + \left(D + \sum_{i=1}^n \frac{U_i}{Q_i}\frac{f_i(S,Q_i)}{S}\right) S = S^{(0)}D.$$

Then from (3.3), (2.3) it follows that

$$S' + \left[D + S^{(0)}\left(\max_{1\le i\le n}\frac{1}{P_i}\right)\cdot \max_{\substack{1\le i\le n\\ 0\le S\le S^{(0)}}} \frac{\partial f_i}{\partial S}(S,P_i)\right] S \ge S^{(0)}D.$$

Then there exists $\gamma > 0$ such that $S(t) \ge \gamma$, $t \ge t_0$.

From (2.1), we have

$$Q_i' = f_i(S, Q_i) - \mu_i(Q_i)Q_i \geq f_i(\gamma, Q_i) - \mu(Q_i)Q_i.$$

Then it follows that $Q_i(t) \geq \gamma_i$ for $t \geq t_0$, where γ_i satisfies

$$f_i(\gamma, \gamma_i) = \mu(\gamma_i)\gamma_i,\ \gamma_i > P_i.$$

For each $1 \leq i \leq n$, we have

$$x_i(t) = U_i(t)/Q_i(t) \leq T(t)/P_i \leq (S^{(0)} + \varepsilon)/P_i \text{ for } t \text{ large.}$$

Consequently, the solution

$$(S(t), x_1(t), Q_1(t), \ldots, x_n(t), Q_n(t))$$

is bounded for $t \geq 0$. □

Proof of Lemma 2.2. Suppose case (i) holds. Then

$$\mu_i(Q_i^0) < D, \tag{3.4}$$

where Q_i^0 is defined in (2.5). In case (ii) or (iii), we have

$$f_i(S^{(0)}, Q_i^*) < \mu_i(Q_i^*)Q_i^*.$$

Since $g_i(Q) = f_i(S^{(0)}, Q) - \mu_i(Q)Q$ is strictly decreasing in Q, from (2.5) it follows that $Q_i^* > Q_i^0$. Thus from (2.2) we obtain (3.4) again.

To complete the proof, it remains to show that the inequality (3.4) implies that $\lim_{t\to\infty} x_i(t) = 0$. Let $\eta = (D - \mu_i(Q_i^0))/2$. Since $\mu_i(Q_i)$ is increasing in Q_i, there exists $\delta > 0$ such that

$$\mu_i(Q_i) \leq \mu_i(Q_i^0) + \eta = D - \eta \quad \text{whenever } Q_i \leq Q_i^0 + \delta.$$

By (3.2) there exists $t_\delta > 0$ such that

$$Q_i(t) < Q_i^0 + \delta \text{ for all } t \geq T_\delta > 0.$$

It follows that

$$\begin{aligned} x_i(t) &= x_i(T_\delta) \exp\left(\int_{T_\delta}^t (\mu_i(Q_i(\tau)) - D)\, d\tau\right) \\ &\leq x_i(T_\delta)e^{-\eta(t-T_\delta)} \to 0 \text{ as } t \to \infty. \end{aligned}$$ □

Proof of Lemma 2.3. Assume that the equilibrium $\hat{E}$ takes the form

$$\hat{E} = (U_1, Q_1, \ldots, U_n, Q_n).$$

Let the variational matrix evaluated at $\hat{E}$ be $J(\hat{E}) = (a_{ij})_{i,j=1}^{2n}$.

Let $\hat{E} = \hat{E}_0$. Then it is easy to verify that the eigenvalues of $J(\hat{E}_0)$ are $a_{11}, a_{22}, \ldots, a_{2n,2n}$, where

$$a_{2i-1,2i-1} = \mu_i(Q_i^0) - D,$$

$$a_{2i,2i} = \frac{\partial f_i}{\partial Q_i}(S^{(0)}, Q_i^0) - \mu_i'(Q_i^0)Q_i^0 - \mu_i(Q_i^0) < 0,\ i = 1, 2, \ldots, n. \tag{3.5}$$

From (2.3), (2.5), and (2.7) we have $S^{(0)} > \lambda_i$ if and only if $Q_i^0 > Q_i^*$. Therefore,

$$a_{1,1} > \mu_1(Q_1^*) - D = 0,$$

and consequently $\hat{E}_0$ is unstable. Furthermore, it is a saddle since (3.5) holds. It is easy to verify that if $S^{(0)} > \lambda_i$, $i = 1, 2, \ldots, n$, then $a_{2i-1,2i-1} > 0$, $i = 1, 2, \ldots, n$, and $\hat{E}_0$ is a saddle point with n-dimensional stable manifold

$$M^+(\hat{E}_0) = \left\{(0, Q_1, 0, Q_2, \ldots, 0, Q_n) : P_i < Q_i,\ i = 1, 2, \ldots, n\right\}.$$

Let $\hat{E} = \hat{E}_k$, $1 \le k \le n$. Then for $i \ne k$,

$$\begin{aligned} a_{2i-1,2i-1} &= \mu_i(\hat{Q}_i^k) - D, \\ a_{2i,2i} &= \frac{\partial f_i}{\partial Q_i}(\lambda_k, \hat{Q}_i^k) - \hat{Q}_i^k \mu_i'(\hat{Q}_i^k) - \mu_i(\hat{Q}_i^k) < 0. \end{aligned}$$

It is easy to verify that the set of eigenvalues of $J(\hat{E}_k)$ is the union of

$$\{a_{2i-1,2i-1}, a_{2i,2i} : 1 \le i \le n, i \ne k\}$$

and the set of eigenvalues of M_k, where

$$M_k = \begin{pmatrix} -\frac{\partial f_k}{\partial S} x_k^* & -f_k(\lambda_k, Q_k^*)\frac{x_k^*}{Q_k^*} + \frac{\partial f_k}{\partial Q_k} x_k^* \\ -\frac{\partial f_k}{\partial S} & \frac{\partial f_k}{\partial Q_k} - \mu_k' Q_k^* - \mu_k \end{pmatrix}.$$

Since

$$\begin{aligned} \operatorname{trace}(M_k) &= -\frac{\partial f_k}{\partial S} x_k^* + \frac{\partial f_k}{\partial Q_k} - \mu_k' Q_k^* - \mu_k < 0, \\ \det(M_k) &= \frac{\partial f_k}{\partial S} x_k^* \mu_k' Q_k^* > 0, \end{aligned}$$

the eigenvalues of M_k have negative real part.

Consider $\hat{E} = \hat{E}_1$. The assumption $(\mathrm{H_n})$implies that

$$\hat{Q}_i^1 < Q_i^*, \quad i = 2, \ldots, n. \tag{3.6}$$

Therefore, from (3.6) it follows that

$$a_{2i-1,2i-1} = \mu_i(\hat{Q}_i^1) - D < \mu_i(Q_i^*) - D = 0, \quad i = 2, \ldots, n,$$

and consequently $\hat{E}_1$ is locally asymptotically stable.

Consider $\hat{E} = \hat{E}_k$, $k \in \{2, \ldots, n\}$. The assumption $(\mathrm{H_n})$implies that $\lambda_1 < \lambda_k$. Then from (2.3) we have

$$\begin{aligned} &f_1(\lambda_1, \hat{Q}_1^k) < f_1(\lambda_k, \hat{Q}_1^k) = \mu_1(\hat{Q}_1^k)\hat{Q}_1^k, \\ &f_1(\lambda_1, \hat{Q}_1^k) - \mu_1(\hat{Q}_1^k)\hat{Q}_1^k < 0 = f_1(\lambda_1, Q_1^*) - \mu_1(\hat{Q}_1^*)Q_1^*. \end{aligned}$$

Thus

$$Q_1^* < \hat{Q}_1^k.$$

Therefore,

$$a_{1,1} = \mu_1(\hat{Q_1^k}) - D > \mu_1(Q_1^*) - D = 0,$$

and consequently $\hat{E}_k$ is unstable. Furthermore, from (3.5) it is a saddle . Similarly, it is easy to verify that if $S^{(0)} > \lambda_i$, $i = 1, 2, \ldots, n$, then $a_{2i-1,2i-1} > 0$, $i = 1, 2, \ldots, k-1$, and $\hat{E}_k$ is a saddle point with a $(2n+1-k)$-dimensional stable manifold. From the results of [SW1] and induction on n, it follows that

$$M^+(\hat{E}_k) = \left\{ (0, Q_1, \ldots, 0, Q_{k-1}, U_k, Q_k, \ldots, U_n, Q_n) : \begin{array}{l} P_i < Q_i, \\ i = 1, 2, \ldots, n \end{array} \right\}. \qquad \square$$

We note now the following lemma.

Lemma 3.2 (see [C]). *Let $f(t) \in C^2[t_0, \infty)$. If $f(t) \to$ constant and $|f''(t)|$ is bounded for $t \geq t_0$, then*

$$\lim_{t\to\infty} f'(t) = 0.$$

The following is the so-called fluctuating lemma, which will be used to prove our main result.

Lemma 3.3 (see [HHG]). *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a differentiable function. If*

$$\liminf_{t\to\infty} f(t) < \limsup_{t\to\infty} f(t),$$

then there are sequences $\{t_m\} \nearrow \infty$ and $\{\tau_m\} \nearrow \infty$ such that for all m

$$f'(t_m) = 0, \quad f(t_m) \to \limsup_{t\to\infty} f(t) \text{ as } m \to \infty,$$
$$f'(\tau_m) = 0, \quad f(\tau_m) \to \liminf_{t\to\infty} f(t) \text{ as } m \to \infty.$$

Now we prove our main result.

Lemma 3.4. *Let $S(t) = S^{(0)} - \sum_{i=1}^n U_i(t)$. Consider the solution*

$$(U_1(t), Q_1(t), \ldots, U_n(t), Q_n(t))$$

of the reduced system (2.12) with initial conditions $U_i(0) > 0, Q_i(0) \geq P_i$, $1 \leq i \leq n$, $S(0) \geq 0$. Suppose $\lim_{t\to\infty} S(t)$ does not exist; then $\limsup_{t\to\infty} S(t) \leq \lambda_j$ for some $j \in \{1, 2, \ldots, n\}$.

Proof. Since $\lim_{t\to\infty} S(t)$ does not exist, it follows that

$$\liminf_{t\to\infty} S(t) < \limsup_{t\to\infty} S(t).$$

From Lemma 3.3, there exists $\{t_m\} \nearrow \infty$ such that

$$S'(t_m) = 0 \quad \text{and} \quad S(t_m) \to \limsup_{t\to\infty} S(t) \text{ as } m \to \infty. \tag{3.7}$$

Since

$$S'(t) = -(U_1'(t) + \cdots + U_n'(t)),$$

for each t_m there exists $j_m \in \{1, 2, \ldots, n\}$ such that

$$U_{j_m}'(t_m) \leq 0, \quad m = 1, 2, \ldots.$$

We may choose a subsequence $\{\bar{t}_m\}$ of $\{t_m\}$ such that

$$U_j'(\bar{t}_m) \le 0$$

for some $j \in \{1, 2, \ldots, n\}$ and for all m. Thus without loss of generality we may assume that

$$U_j'(t_m) \le 0$$

for some $j \in \{1, 2, \ldots, n\}$ and for all m. Thus

$$f_j(S(t_m), Q_j(t_m)) \le DQ_j(t_m).$$

Let $\gamma_S = \limsup_{t\to\infty} S(t)$ and $\gamma_Q = \limsup_{t\to\infty} Q_j(t)$. Let $\{\tilde{t}_m\}$ be a subsequence of $\{t_m\}$ such that $\lim_{m\to\infty} Q_j(\tilde{t}_m) = \bar{Q}_j$. Then $\bar{Q}_j \le \limsup_{t\to\infty} Q_j(t) = \gamma_Q$, and from the above inequality we have $f_j(\gamma_S, \bar{Q}_j) \le D\bar{Q}_j$. Since $f_j(\gamma_S, Q_j) - DQ_j$ is strictly decreasing in Q_j, then $f_j(\gamma_S, \gamma_Q) - D\gamma_Q < f_j(\gamma_S, \bar{Q}_j) - D\bar{Q}_j \le 0$. Thus we have

$$f_j(\gamma_S, \gamma_Q) < D\gamma_Q. \tag{3.8}$$

Consider the differential equation of Q_j in (2.1):

$$Q_j' = f_j(S, Q_j) - \mu_j(Q_j)Q_j. \tag{3.9}$$

From (3.1), (2.3), and Lemma 3.1 it follows that

$$\gamma_Q = \limsup_{t\to\infty} Q_j(t) \le K^{(0)}, \tag{3.10}$$

where

$$f_j(S^{(0)}, K^{(0)}) - \mu_j(K^{(0)})K^{(0)} = 0. \tag{3.11}$$

If $\lambda_j > S^0$, from (3.1) the assertion of the lemma holds. Thus we assume that $\lambda_j \le S^0$. From (2.3) and (3.11) it follows that

$$f_j(\lambda_j, K^{(0)}) - \mu_j(K^{(0)})K^{(0)} \le 0.$$

Compare the above inequality with (2.7):

$$f_j(\lambda_j, Q_j^*) - \mu_j(Q_j^*)Q_j^* = 0. \tag{3.12}$$

From (2.2), (2.3), (3.11), and (3.12) it follows that

$$K^{(0)} \ge Q_j^*. \tag{3.13}$$

Let $L^{(1)}$ satisfy

$$f_j(L^{(1)}, K^{(0)}) - DK^{(0)} = 0. \tag{3.14}$$

Then from (2.3), (3.10) we have

$$0 = f_j(L^{(1)}, K^{(0)}) - DK^{(0)} \le f_j(L^{(1)}, \gamma_Q) - D\gamma_Q.$$

From (2.3), (3.8) it follows that

$$f_j(L^{(1)}, \gamma_Q) \ge D\gamma_Q \ge f_j(\gamma_S, \gamma_Q),$$

$$\gamma_S \le L^{(1)}. \tag{3.15}$$

Since $K^{(0)} \ge Q_j^*$, from (3.14) and (2.3) it follows that

$$f_j(L^{(1)}, Q_j^*) - DQ_j^* \ge 0.$$

From (3.12) we have

$$L^{(1)} \ge \lambda_j.$$

On the other hand, the inequality $K^{(0)} \ge Q_j^*$ implies that

$$f_j(L^{(1)}, K^{(0)}) = DK^{(0)} = \mu_j(Q_j^*)K^{(0)} \le \mu_j(K^{(0)})K^{(0)} = f_j(S^{(0)}, K^{(0)}).$$

Thus we have

$$S^{(0)} \ge L^{(1)} \ge \lambda_j. \tag{3.16}$$

By (3.9), (3.15), and Lemma 3.1, we have

$$\limsup_{t\to\infty} Q_j(t) \le K^{(1)}, \tag{3.17}$$

where

$$f_j(L^{(1)}, K^{(1)}) = \mu_j(K^{(1)})K^{(1)}. \tag{3.18}$$

Since $\lambda_j \le L^{(1)}$, it follows that

$$f_j(\lambda_j, K^{(1)}) - \mu_j(K^{(1)})K^{(1)} \le 0.$$

By (3.12), we have

$$K^{(1)} \ge Q_j^*.$$

Since $S^{(0)} \ge L^{(1)}$, from (3.11), (3.16), and (3.18) it follows that

$$K^{(0)} \ge K^{(1)} \ge Q_j^*. \tag{3.19}$$

Inductively we construct two sequences $\{L^{(m)}\}_{m=1}^{\infty}$ and $\{K^{(m)}\}_{m=1}^{\infty}$ satisfying

$$\begin{aligned} &S^{(0)} \ge L^{(1)} \ge L^{(2)} \ge \cdots \ge \lambda_j,\\ &K^{(0)} \ge K^{(1)} \ge K^{(2)} \ge \cdots \ge Q_j^*, \end{aligned}$$

and for any $m = 1, 2, \dots,$

$$\begin{aligned} &\limsup_{t\to\infty} S(t) \le L^{(m)},\\ &\limsup_{t\to\infty} Q_j(t) \le K^{(m)}, \end{aligned} \tag{3.20}$$

$$\begin{aligned} &f_j(L^{(m+1)}, K^{(m)}) = DK^{(m)},\\ &f_j(L^{(m)}, K^{(m)}) = \mu_j(K^{(m)})K^{(m)}. \end{aligned} \tag{3.21}$$

Let $L = \lim_{m\to\infty} L^{(m)}$ and $K = \lim_{m\to\infty} K^{(m)}$. Then from (3.21) it follows that

$$f_j(L, K) = DK,$$
$$f_j(L, K) = \mu_j(K)K.$$

Thus $K = Q_j^*$ and $L = \lambda_j$. By (3.20) it follows that

$$\limsup_{t\to\infty} S(t) \le \lambda_j,$$
$$\limsup_{t\to\infty} Q_j(t) \le Q_j^*.$$

Hence we complete the proof of Lemma 3.4. □

THEOREM 3.5. *Let* ($\mathrm{H_n}$)*hold. Then the solution*

$$(U_1(t), Q_1(t), \ldots, U_n(t), Q_n(t))$$

of the reduced system (2.12) *in the relevant domain* Ω *(see* (2.13)*) satisfies*

$$\lim_{t\to\infty}(U_1(t), Q_1(t), \ldots, U_n(t), Q_n(t)) = \hat{E}_1 = (U_1^*, Q_1^*, 0, \hat{Q}_2^1, \ldots, 0, \hat{Q}_n^1). \tag{3.22}$$

Proof. Let $S(t) = S^{(0)} - \sum_{i=1}^n U_i(t)$. If $\lim_{t\to\infty} S(t)$ exists, we claim that $\lim_{t\to\infty} S(t) = \lambda_1$. Let $\lim_{t\to\infty} S(t) = c$.

If $c > \lambda_1$, then for $\varepsilon > 0$ small there exists $T_\varepsilon > 0$ such that

$$Q_1' > f_1(\lambda_1 + \varepsilon, Q_1) - \mu_1(Q_1)Q_1 \text{ for t} \ge \mathrm{T}_\varepsilon.$$

Thus $Q_1(t) \ge Q_1^* + \eta$, $\eta > 0$ small, $t \ge T_\varepsilon$. Hence

$$\frac{x_1'}{x_1} = \mu_1(Q_1) - D \ge \mu_1(Q_1^* + \eta) - D > 0.$$

Then $x_1(t)$ is unbounded for $t \ge T_\varepsilon$. This is in contradiction to Lemma 2.1.

If $c < \lambda_1$, then for $2 \le i \le n$, by the differential equation of Q_i in (2.1) and Lemma 3.1, we have $\limsup_{t\to\infty} Q_1(t) < Q_1^*$ and $\limsup_{t\to\infty} Q_i(t) < \hat{Q}_i^1$ for $2 \le i \le n$. Hence from (3.6), $\lim_{t\to\infty} x_i(t) = 0$, $1 \le i \le n$, and $\lim_{t\to\infty} S(t) = S^{(0)} < \lambda_1$. This is in contradiction to ($\mathrm{H_n}$).

Obviously from Lemma 3.2, $\lim_{t\to\infty} S(t) = \lambda_1$ implies

$$\lim_{t\to\infty} Q_i(t) = \hat{Q}_i^1,\ \lim_{t\to\infty} x_i(t) = 0,\ 2 \le i \le n,$$
$$\lim_{t\to\infty} Q_1(t) = Q_1^*,\ \lim_{t\to\infty} x_1(t) = x_1^*.$$

Thus the trajectory $(U_1(t), Q_1(t), \ldots, U_n(t), Q_n(t))$ tends to $\hat{E}_1$ as $t \to \infty$.

If $\lim_{t\to\infty} S(t)$ does not exist, then $\limsup_{t\to\infty} S(t) > \liminf_{t\to\infty} S(t)$. From Lemma 3.4, we have $\limsup_{t\to\infty} S(t) \le \lambda_j$ for some $j \in \{1, 2, \ldots, n\}$. From ($\mathrm{H_n}$), we have

$$\limsup_{t\to\infty} S(t) \le \lambda_n.$$

Assume that (2.6) and (2.7) hold. Consider the differential equation of Q_n in (2.1):

$$Q_n' = f_n(S, Q_n) - \mu_n(Q_n)Q_n.$$

From Lemma 3.1 it follows that

$$\limsup_{t\to\infty} Q_n(t) \le \tilde{Q}_n,$$

where $\tilde{Q}_n$ satisfies

$$f_n(\lambda_n, \tilde{Q}_n) = \mu_n(\tilde{Q}_n)\tilde{Q}_n.$$

From (2.7) it follows that $\tilde{Q}_n = Q_n^*$. Thus

$$\limsup_{t\to\infty} Q_n(t) \le Q_n^*. \tag{3.23}$$

Let

$$\kappa_n = \liminf_{t\to\infty} Q_n(t).$$

If $\kappa_n = Q_n^*$, then $\lim_{t\to\infty} Q_n(t) = Q_n^*$. From (3.23) and Lemma 3.2, we have $\lim_{t\to\infty} S(t) = \lambda_n$, which contradicts the assumption that $\lim_{t\to\infty} S(t)$ does not exist. Hence we have $\kappa_n < Q_n^*$. Let

$$y_0 = (U_1(0), Q_1(0), \dots, U_n(0), Q_n(0)),\ U_i(0) > 0,\ Q_i(0) \ge P_i(0) \text{ for } 1 \le i \le n.$$

Next we claim that the ω-limit set $\omega(y_0)$ satisfies

$$\omega(y_0) \cap (\{(U_1, Q_1, \dots, U_n, Q_n) : U_n = 0\} \setminus M) \ne \emptyset, \tag{3.24}$$

where

$$M := \left(M^+(\hat{E}_0) \bigcup M^+(\hat{E}_2) \bigcup \cdots \bigcup M^+(\hat{E}_n)\right);$$

$M^+(\hat{E})$ denotes the stable manifold of the equilibrium $\hat{E}$. First we prove that

$$\omega(y_0) \setminus M \ne \emptyset.$$

If not, then $\omega(y_0) \subseteq M$. It is easy to show that $\omega(y_0) \ne \{\hat{E}_0\}$. If $\hat{E}_0 \in \omega(y_0)$, then from the Butler–McGhee lemma [BFW], there exists a point

$$q \in \left(M^+(\hat{E}_0) \setminus \{\hat{E}_0\}\right) \bigcap \omega(y_0).$$

Then the negative orbit $O^-(q) \subseteq \omega(y_0)$. But from Lemma 2.3, either $O^-(q)$ is unbounded or $(0, P_1, 0, P_2, \dots, 0, P_n) \in O^-(q)$. This contradicts Lemma 2.1. Assume that $\hat{E}_k \in \omega(y_0)$ for some $k \in \{2, \dots, n\}$. Obviously $\omega(y_0) \ne \{\hat{E}_k\}$. If $\hat{E}_k \in \omega(y_0)$, then from the Butler–McGhee lemma, there exists a point $q \in (M^+(\hat{E}_k) \setminus \{\hat{E}_k\}) \bigcap \omega(y_0)$. Then from the Lemma 2.3 the negative orbit $O^-(q)$ is unbounded, or $\hat{E}_0 \in O^-(q)$, or $(0, P_1, \dots, 0, P_{k-1}, U_k, P_k, \dots, U_n, P_n) \in O^-(q)$ for some $U_k, \dots, U_n$. For any one of the three cases, we obtain a contradiction.

Since $y_0 \notin M$, we may choose

$$\bar{y}_0 = (\bar{U}_1(0), \bar{Q}_1(0), \dots, \bar{U}_n(0), \bar{Q}_n(0)) \in (\omega(y_0) \setminus M). \tag{3.25}$$

Consider the solution of (2.12):

$$y(t, \bar{y}_0) = (U_1(t; \bar{y}_0), Q_1(t; \bar{y}_0), \dots, U_n(t; \bar{y}_0), Q_n(t; \bar{y}_0)).$$

From (3.23) and the positive invariance of $\omega(y_0)$, we have

$$Q_n(t, \bar{y}_0) \le Q_n^*, \quad t \ge 0.$$

Thus

$$\mu_n(Q_n(t; \bar{y}_0)) - D \le 0, \quad t \ge 0. \tag{3.26}$$

Let

$$\eta = D - \mu_n\left(\frac{Q_n^* + \kappa_n}{2}\right) > 0$$

and

$$\Lambda(t) = \left\{\tau \ : \ 0 \le \tau \le t, \ Q_n(\tau; \bar{y}_0) \le \frac{Q_n^* + \kappa_n}{2}\right\}, \quad t \ge 0.$$

Then

$$\mu_n\left(Q_n(\tau, \bar{y}_0)\right) - D < -\eta, \quad \tau \in \Lambda(t).$$

Since $Q_n'(t; \bar{y}_0)$ is uniformly bounded for $t \in [0, \infty)$, $Q_n(t; \bar{y}_0)$ is uniformly continuous on $[0, \infty)$. Let $\{\tau_m\} \nearrow \infty$ satisfies $Q_n(\tau_m; \bar{y}_0) \to \kappa_n$ as $m \to \infty$. Then given

$$\varepsilon = \frac{Q_n^* + \kappa_n}{2} - \kappa_n > 0,$$

there exists $\delta = \delta(\varepsilon) > 0$ such that

$$|Q_n(\tau; \bar{y}_0) - \kappa_n| < \varepsilon \text{ whenever } |\tau - \tau_m| < \delta.$$

Hence

$$Q_n(\tau; \bar{y}_0) < \kappa_n + \varepsilon = \frac{Q_n^* + \kappa_n}{2} \qquad \text{for } -\delta < \tau - \tau_m < \delta,$$

and therefore

$$|\Lambda(t)| \to +\infty \text{ as } t \to \infty.$$

Since

$$x_n'(t; \bar{y}_0) = (\mu_n(Q_n(t; \bar{y}_0)) - D)x_n(t; \bar{y}_0),$$

it follows that

$$\begin{aligned} x_n(t; \bar{y}_0) &= x_n(0; \bar{y}_0) \exp\left(\int_0^t (\mu_n(Q_n(\tau; \bar{y}_0)) - D)\, d\tau\right) \\ &\le x_n(0; \bar{y}_0) \exp\left(\int_{\Lambda(t)} (\mu_n(Q_n(\tau; \bar{y}_0)) - D)\, d\tau\right) \\ &\le x_n(0; \bar{y}_0) e^{-\eta|\Lambda(t)|} \to 0 \text{ as } t \to \infty. \end{aligned}$$

Therefore,

$$\begin{aligned}\limsup_{t\to\infty} U_n(t;\bar{y}_0) &\leq \left(\limsup_{t\to\infty} x_n(t;\bar{y}_0)\right)\left(\limsup_{t\to\infty} Q_n(t;\bar{y}_0)\right)\\ &\leq \left(\limsup_{t\to\infty} x_n(t;\bar{y}_0)\right) Q_n^* = 0.\end{aligned}$$

Hence

$$\omega(\bar{y}_0) \subseteq \{(U_1, Q_1, \ldots, U_n, Q_n) \in \Omega : U_n = 0\}.$$

Since $\bar{y}_0 \notin M$ by (3.25), it follows that

$$\omega(\bar{y}_0) \cap (\{(U_1, Q_1, \ldots, U_n, Q_n) \in \Omega : U_n = 0\} \setminus M) \neq \emptyset.$$

By the invariance of ω-limit sets, we have

$$\omega(\bar{y}_0) \subseteq \omega(y_0).$$

It follows that

$$\omega(y_0) \cap (\{(U_1, Q_1, \ldots, U_n, Q_n) \in \Omega : U_n = 0\} \setminus M) \neq \emptyset.$$

Continuing the above arguments, we consider the systems (2.12) with $1 \leq i \leq n-1$. Then from the positive invariance of the ω-limit set,

$$\omega(y_0) \cap (\{(U_1, Q_1, \ldots, U_n, Q_n) \in \Omega : U_{n-1} = U_n = 0\} \setminus M) \neq \emptyset.$$

Inductively we have

$$\omega(y_0) \cap (\Gamma \setminus M) \neq \emptyset,$$

where

$$\Gamma = \{(U_1, Q_1, \ldots, U_n, Q_n) \in \Omega : U_2 = U_3 = \cdots = U_n = 0\}.$$

In particular,

$$\omega(y_0) \cap (\Gamma \setminus \{\hat{E}_0\}) \neq \emptyset.$$

It is easy to verify that

$$\omega(\Gamma \setminus \{\hat{E}_0\}) = \{\hat{E}_1\}.$$

Consequently, we have

$$\hat{E}_1 \in \omega(y_0).$$

By Lemma 2.3, the assumption $(\mathrm{H_n})$implies that $\hat{E}_1$ is asymptotically stable. Thus

$$\omega(y_0) = \{\hat{E}_1\}.$$

That is,

$$\lim_{t\to\infty} (U_1(t), Q_1(t), \ldots, U_n(t), Q_n(t)) = \hat{E}_1.$$

The above equality contradicts the assumption that $\lim_{t\to\infty} S(t)$ does not exist. Thus $\lim_{t\to\infty} S(t)$ exists, and we complete the proof of Theorem 3.5. $\square$

Proof of Theorem 2.4. From Lemma 2.1 all solutions of the system (2.1) with initial conditions $S(0) > 0, x_i(0) > 0, Q_i(0) \geq P_i$ asymptotically approach

$$S + \sum_{i=1}^{n} U_i = S^{(0)}$$

as $t \to \infty$. Hence the system (2.12) is the reduced limiting system of (2.1). To apply Theorem 4.2 of [Th], we note that the equilibria of (2.12) are isolated invariant sets of (2.12) and by Theorem 3.5, every solution of (2.12) converges to the equilibrium $\hat{E}_1 = (U_1^*, Q_1^*, 0, \hat{Q}_2^1, \dots, 0, \hat{Q}_n^1)$. Furthermore, we conclude from [Th, Theorem 4.2], that every solution of (2.1) converges to the equilibrium

$$E_1 = (\lambda_1, x_1^*, Q_1^*, 0, \hat{Q}_2^1, 0, \hat{Q}_3^1, \dots, 0, \hat{Q}_n^1). \qquad \square$$

4. Discussion. It is well known that the competitive exclusion principle holds for microorganisms competing for a single-limited nutrient in a chemostat when the yields of organisms are assumed to be fixed constants [HHW], [H1]. In phytoplankton ecology, it has long been known that yield is not constant and it can vary depending on the growth rate [D]. This led to the formulation of the variable-yield model, or the internal storage model. In this paper we proved that the competitive exclusion principle also holds for the variable-yield model in the case of a single-limited nutrient. Mathematically we extend the result of competitive exclusion in [SW1] from two species to arbitrary n species. Biologically the internal storage model with one limiting nutrient has been tested successfully in both constant and fluctuating environments [G3], [SC]. It is more realistic than the constant-yield model.

However, organisms require multiple nutrients to live and reproduce. In phytoplankton ecology, there are many studies in the competition of species for multiple nutrients. Narang and Pilyugin [NP] studied the dynamics of microbial growth by constructing some new physiological models. In [LC] Legović and Cruzado proposed an internal storage model of one species consuming multiple complementary nutrients in a continuous culture. Then in [LLSK] Leenheer et al. proved the global stability for the above model by the method of monotone dynamical systems. Li and Smith [LS1] studied the internal storage model for two species competing for two complementary nutrients. By using the method of monotone dynamical systems, they established the global dynamics of the model. It is shown that basically the model exhibits the familiar Lotka–Volterra alternatives: competitive exclusion, stable coexistence, and bistability. In phytoplankton ecology, many people studied the competition of organisms for multiple complementary nutrients by using the internal storage model. In [KL] Klausmeier and Litchman studied phytoplankton growth and stoichiometry under multiple nutrient limitation. In [KLL] Klausmeier, Litchman, and Levin studied the case of two species and two essential nutrients and suggested experimental tests for the model. In [LKMSF] the authors studied the multiple-nutrient, multiple-group model for phytoplankton communities and listed many biological parameters in the internal storage model.

We conjecture that for the internal storage model there are at most two species that survive for the case of n organisms competing for two complementary nutrients. We note that even in the classical model of fixed yields, the conjecture is still unsolved [LS2]. It is also interesting to compare the mathematical analysis results of the internal storage model to those of the classical constant-yield model in the case of three or more complementary nutrients [PH]. These will be the subject of our work in the future.

Acknowledgments. We are grateful to two anonymous referees for their careful reading and helpful suggestions which led to an improvment of our original manuscript.

REFERENCES

[AM] R. A. Armstorng and R. McGehee, *Competitive exclusion*, American Naturalist, 115 (1980), pp. 151–170.

[BFW] G. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.

[C] W. A. Coppell, *Stability and Asymptotic Behavior of Solutions of Differential Equations*, Heath, Boston, 1965.

[CM] A. Cunningham and P. Maas, *Time lag and nutrient storage effects in the transient growth response of Chelamydomonas reinhardii in nitrogen limited batch and continuous culture*, J. Gen. Microbiol., 104 (2978), pp. 227–231.

[CN1] A. Cunningham and R. M. Nisbet, *Time lag and co-operativity in the transient growth dynamics of microalgae*, J. Theoret. Biol., 84 (1980), pp. 189–203.

[CN2] A. Cunningham and R. M. Nisbet, *Transient and oscillation in continuous culture*, in Mathematics in Microbiology, M. J. Bazin, ed., Academic Press, New York, 1983, pp. 77–103.

[D] M. Droop, *Some thoughts on nutrient limitation in algae*, J. Phycol., 9 (1973), pp. 264–272.

[FS] A. G. Frederickson and G. Stephanopoulus, *Microbial competition*, Science, 243 (1981), pp. 972–979.

[G1] J. P. Grover, *Resource competition in variable environment: Phytoplankton growing according to variable-internal-stores model*, American Naturalist, 138 (1991), pp. 811–835.

[G2] J. P. Grover, *Non-steady state dynamics of algal population growth: Experiment with two Chlorophytes*, J. Phycol. 27 (1991), pp. 70–79.

[G3] J. P. Grover, *Constant- and variable-yield models of population growth: Responses to environmental variability and implications for competition*, J. Theoret. Biol., 158 (1992), pp. 409–428.

[H1] S. B. Hsu, *Limiting behavior for competing species*, SIAM J. Appl. Math., 34 (1978), pp. 760–763.

[H2] S. B. Hsu, *Ordinary Differential Equations with Applications*, World Scientific Press, Hackensack, NJ, 2006.

[HHG] W. M. Hirsch, H. Hanisch, and J. P. Gabriel, *Differential equation models of some parasitic infections: Methods for the study of asymptotic behavior*, Comm. Pure Appl. Math., 38 (1985), pp. 733–753.

[HHW] S. B. Hsu, S. Hubbell, and P. Waltman, *Mathematical theory for single-nutrient competition in continuous cultures of micro-organisms*, SIAM J. Appl. Math., 32 (1977), pp. 366–383.

[JM] H. W. Jannash and R. T. Matiles, *Experimental bacterial ecology studied in continuous culture*, Adv. Microbial Phys., 11 (1974), pp. 423–439.

[KL] C. A. Klausmeier and E. Litchman, *Phytoplankton growth and stoichioetry under multiple nutrient limitation*, Limnol. Oceanogr., 49 (2004), pp. 1463–1470.

[KLL] C. A. Klausmeier, E. Litchman, and S. A. Levin, *A model of flexible uptake of two essential resources*, J. Theoret. Biol., 246 (2007), pp. 278–289.

[LC] T. Legović and A. Cruzado, *A model of phytoplankton growth on mutiple nutrients based on the Michaelis–Menten–Monod uptake, Droop's growth, and Liebig's law*, Ecol. Model., 99 (1997), pp. 19–31.

[LKMSF] E. Litchmam, C. A. Klausmeier, J. R. Miller, O. M. Schofield, and P. G. Falkowski, *Mutiple-nutrient, multi-group of present and future oceanic phytoplankton communities*, Biogeoscience, 3 (2006), pp. 585–606.

[LLSK] P. De Leenheer, S. A. Levin, E. D. Sontag, and C. A. Klausmeier, *Global stability in a chemostat with multiple nutrients*, J. Math. Biol., 52 (2006), pp. 419–438.

[LS1] B. Li and H. L. Smith, *Global dynamics of microbial competition for two resources with internal storage*, J. Math. Biol., 55 (2007), pp. 481–515.

[LS2] B. Li and H. L. Smith, *Competition for essential resources: A brief review*, in Dynamical Systems and Their Applications in Biology, S. Ruan, G. S. K. Wolkowicz, and J. Wu, eds., Fields Inst. Commun. 36, AMS, Providence, RI, 2003, pp. 213–227.

[M] J. Monod, *Recherches sur la Croissance des Cultures Bacteriennes*, Hermann, Paris, 1942.

[NG] R. M. Nisbet and W. S. C. Gurney, *Modelling Fluctuating Populations*, John Wiley & Sons, New York, 1982.

[NP] A. Narang and S. S. Pilyugin, *Towards an integrated physiological theory of microbial growth: From subcellular variables to population dynamics*, Math. Biosc. Eng., 2 (2005), pp. 173–210.

[PH] J. Passarge and J. Huisman, *Competition in well-mixed habitats: From competitive exclusion to competitive chaos*, in Competition and Coexistence, U. Sommer and B. Worm, eds., Springer-Verlag, New York, 2003, pp. 7–33.

[T] D. Tilman, *Resource Competition and Community Structure*, Princeton University Press, Princeton, NJ, 1982.

[Th] H. R. Thieme, *Convergence results and a Poincari-Bendixson trichotomy for asymptotically autonomous differential equations*, J. Math. Biol., 30 (1992), pp. 755–763.

[WX] G. S. K. Wolkowicz and H. Xia, *Global asymptotic behavior of a chemostat model with discrete delays*, SIAM J. Appl. Math., 57 (1997), pp. 1019–1043.

[S] H. L. Smith, *Monotone Dynamical Systems*, AMS, Providence, RI, 1995.

[SC] E. Spijkerman and P. F. M. Coesel, *Competition for phosphorus between planktonic desmid species in continuous-flow culture*, J. Phycol., 32 (1996), pp. 939–948.

[SW1] H. L. Smith and P. E. Waltman, *Competition for a single limiting resource in continuous culture: The variable-yield model*, SIAM J. Appl. Math., 54 (1994), pp. 1113–1131.

[SW2] H. L. Smith and P. E. Waltman, *Theory of Chemostat*, Cambridge University Press, Cambridge, UK, 1995.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1618–1640

EXISTENCE, UNIQUENESS, AND A CONSTRUCTIVE SOLUTION ALGORITHM FOR A CLASS OF FINITE MARKOV MOMENT PROBLEMS*

LAURENT GOSSE† AND OLOF RUNBORG‡

Abstract. We consider a class of finite Markov moment problems with an arbitrary number of positive and negative branches. We show criteria for the existence and uniqueness of solutions, and we characterize in detail the nonunique solution families. Moreover, we present a constructive algorithm to solve the moment problems numerically and prove that the algorithm computes the right solution.

Key words. finite Markov moment problem, inverse problems, exponential transform

AMS subject classifications. 30E05, 15A29, 65H10, 65M99

DOI. 10.1137/070692510

1. Introduction.

We aim at inverting a moment system often associated with the prestigious name of Markov. The original form of the problem is the following. Given a *finite set of moments* m_k for $k = 1, \dots, K$, find a bounded measurable *density* function f satisfying

$$m_k = \int_{\mathbb{R}} x^{k-1} f(x)dx, \qquad 0 \le f \le 1, \qquad k = 1, \dots, K. \tag{1.1}$$

The condition for the existence of solutions $f(x)$ to this problem is classical [1, 2]. In general, solutions are not unique, unless more conditions are given, e.g., based on entropy minimization [3, 4] or L^∞-minimization [19, 18]. A typical result is that the unique solution for even K is piecewise constant, taking values in $\{0, 1\}$. More precisely, if $K = 2n$, then f is of the form

$$f(x) = \sum_{j=1}^{n} \chi_{[y_i, x_i]}(x), \tag{1.2}$$

where $\chi_I(x)$ is the characteristic function for the interval I and

$$y_1 < x_1 < y_2 < x_2 < \cdots < y_n < x_n. \tag{1.3}$$

See Theorem 6.1 and consult, e.g., [5, 8, 17, 23, 25] for general background on moment problems.

A reduced form of the finite moment problem is to search for solutions to (1.1) which are precisely of the form (1.2), (1.3). One then obtains an algebraic problem for the branch values,

$$m_k = \frac{1}{k} \sum_{j=1}^{n} x_j^k - y_j^k, \qquad k = 1, \dots, K = 2n. \tag{1.4}$$

*Received by the editors May 20, 2007; accepted for publication (in revised form) March 10, 2008; published electronically June 6, 2008.
http://www.siam.org/journals/siap/68-6/69251.html

†IAC–CNR "Mauro Picone" (sezione di Bari), Via Amendola 122/D, 70126 Bari, Italy (l.gosse@ba.iac.cnr.it).

‡Department of Numerical Analysis, CSC, KTH, 10044 Stockholm, Sweden (olofr@nada.kth.se).

Finding $\{x_j\}$ and $\{y_j\}$ from $\{m_k\}$ is an ill-conditioned problem when the branch values of the solution come close to each other; the Jacobian of the problem is a Vandermonde matrix, and iterative numerical resolution routines require extremely good starting guesses when the matrix degenerates. For less than four moments, a direct method based on solving polynomial equations was presented in [21]. Routines based on the simplex algorithm were proposed in [19]. Another algorithm was presented by Koborov, Sklyar, and Fardigola in [16, 24] in the slightly modified setting where f takes values in $\{-1,1\}$ instead of $\{0,1\}$. It consists of solving a sequence of high degree polynomial equations, constructed through a rather intricate process with unclear stability properties. In [14] we showed that this algorithm can be drastically simplified and adapted to (1.4). Later, in [15], we also gave a direct proof that the simplified algorithm indeed computes the correct solution, relying on the classical Newton identities and Toeplitz matrix theory.

The moment problem has many applications in, for instance, probability and statistics [10, 7] but also in areas like wave modulation [6, 22] and "shape from moments" inverse problems [11]. Our own motivation comes from a quite different field, namely, multiphase geometrical optics [3, 4, 12, 13, 14, 21]. In this application one needs to solve a system of nonlinear hyperbolic conservation laws. To evaluate the flux function in the PDEs a system like (1.4) must be solved. In a finite difference method this means that the system must be inverted once for every point in the computational grid repeatedly in every timestep. It is thus important that the inversion can be done quickly and accurately; this difficulty has been a bottleneck in computations. In [14] we used the simplified algorithm mentioned above for numerical implementation inside a shock-capturing finite difference solver. It is our aim here to develop better algorithms and understanding to open the way for the processing of intricate wave-fields with large K and thus complement the seminal paper [4], where the multiphase geometrical optics PDEs were first proposed.

In this paper we are concerned with a generalization of (1.4). In the geometrical optics application, the number of moments K is typically not even, and one can have a variable number of positive (x_k) and negative (y_k) branches. We thus consider the problem

$$m_k = \sum_{j=1}^{n_x} x_j^k - \sum_{j=1}^{n_y} y_j^k, \qquad k = 1, \dots, K, \tag{1.5}$$

where $n_x + n_y = K$ but where n_x and n_y are not necessarily equal. We study existence and uniqueness of solutions to this problem (Theorem 4.1). In particular, we are interested in how and when uniqueness is lost. For these cases we characterize the family of solutions that exists. The reason is to understand what happens numerically close to degenerate solutions, which is an important feature in the application we have in mind: In the exact solution to the multiphase geometrical optics PDEs, the moment problem is typically degenerate for large domains; the numerical approximation is almost degenerate.

We also give constructive algorithms to solve (1.5) and prove that they generate the right solution (Theorem 2.1). In a future paper we will study the numerical stability of these algorithms. Experimentally we note, for instance, that to compute the next moment, Algorithm 3 is much more stable than Algorithm 1. The difficulty lies in understanding perturbations around degenerate solutions, which is where the algorithms are most unstable. For this the insights of this paper will be of importance.

Remark 1. *The problem* (1.5) *can be cast in the form of* (1.1) *if one demands that the density function $f(x)$ be of the form*

$$f(x) = \sum_{j=1}^{n_x} \operatorname{sgn}(x_j)\,[H(x) - H(x - |x_j|)] - \sum_{j=1}^{n_y} \operatorname{sgn}(y_j)\,[H(x) - H(x - |y_j|)]\,, \tag{1.6}$$

and we rescale the moments $m_k \to k m_k$. For the case $n_x = n_y = n$ and $K = 2n$ with interlaced branch values (1.3)*, this reduces to* (1.2)*.*

This paper is organized as follows. In section 2 we present the algorithms for solving (1.5). Notation and various ways of describing a solution are subsequently introduced in section 3. Next we derive conditions for existence and uniqueness of solutions in section 4 and also discuss various properties of the solution, particularly when it is not unique. A theorem proving the correctness of the algorithms is proved in section 5. Finally, in section 6, we give additional properties of the elements of our algorithms and use these to relate our results back to the classical Markov theory.

2. Algorithms.

In this section we detail the algorithms that we propose for solving (1.5). The solution that we obtain is what we call the *minimal degree solution*, meaning that when the solution is not unique as many branch values as possible are zero. See section 4 for a precise definition. The algorithms are as follows; they may fail in case there is no solution to (1.5).

Algorithm 1 (computing $\{x_j\}$ and $\{y_j\}$).

1. *Construct the sequence $\{a_k\}$ as follows. Set $a_0 = 1$ and $a_k = 0$ for $k < 0$. For $1 \le k \le K$, let the elements be given as the solution to*

$$\begin{pmatrix} 1 & & & \\ -m_1 & 2 & & \\ \vdots & \ddots & \ddots & \\ -m_{K-1} & \dots & -m_1 & K \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_K \end{pmatrix}. \tag{2.1}$$

2. *Construct the matrix $A_1 \in \mathbb{R}^{n_x \times n_x}$ as*

$$A_1 = \begin{pmatrix} a_{n_y} & a_{n_y-1} & \cdots & a_{n_y-n_x+1} \\ a_{n_y+1} & a_{n_y} & \cdots & a_{n_y-n_x+2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_y+n_x-1} & a_{n_y+n_x-2} & \cdots & a_{n_y} \end{pmatrix}.$$

   Compute the rank of A_1. Let $\tilde{n}_x = \operatorname{rank} A_1$ and $\tilde{n}_y = n_y - n_x + \tilde{n}_x$.

3. *Construct the matrices $\tilde{A}_0, \tilde{A}_1 \in \mathbb{R}^{\tilde{n}_x \times \tilde{n}_x}$ as*

$$\tilde{A}_0 = \begin{pmatrix} a_{\tilde{n}_y+1} & a_{\tilde{n}_y} & \cdots & a_{\tilde{n}_y-\tilde{n}_x+2} \\ a_{\tilde{n}_y+2} & a_{\tilde{n}_y+1} & \cdots & a_{\tilde{n}_y-\tilde{n}_x+3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\tilde{n}_y+\tilde{n}_x} & a_{\tilde{n}_y+\tilde{n}_x-1} & \cdots & a_{\tilde{n}_y+1} \end{pmatrix},$$

$$\tilde{A}_1 = \begin{pmatrix} a_{\tilde{n}_y} & a_{\tilde{n}_y-1} & \cdots & a_{\tilde{n}_y-\tilde{n}_x+1} \\ a_{\tilde{n}_y+1} & a_{\tilde{n}_y} & \cdots & a_{\tilde{n}_y-\tilde{n}_x+1+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\tilde{n}_y+\tilde{n}_x-1} & a_{\tilde{n}_y+\tilde{n}_x-2} & \cdots & a_{\tilde{n}_y} \end{pmatrix}.$$

4. *Solve the generalized eigenvalue problem*

$$\tilde{A}_0\boldsymbol{v} = x\tilde{A}_1\boldsymbol{v} \tag{2.2}$$

to get the $\{x_j\}$ *values of the minimal degree solution to* (1.5)*.*

5. *To compute the* $\{y_j\}$ *values, the same process is used with* m_k *replaced by* $-m_k$ *and the roles of* n_x *and* n_y *interchanged.*

An alternative to Algorithm 1 is as follows.

ALGORITHM 2 (computing $\{x_j\}$ and $\{y_j\}$).

1. *Construct the matrices* $\tilde{A}_0$ *and* $\tilde{A}_1$ *as in steps* 1–3 *in Algorithm* 1*.*
2. *Denote the first column vector in* $\tilde{A}_0$ *by* $\tilde{\boldsymbol{a}}_0$ *by and solve*

$$\tilde{A}_1\boldsymbol{c}' = -\tilde{\boldsymbol{a}}_0, \qquad \boldsymbol{c}' = (c_1, c_2, \dots, c_{\tilde{n}_x})^T. \tag{2.3}$$

3. *Construct the polynomial*

$$P(z) = c_{\tilde{n}_x} + c_{\tilde{n}_x-1}z + \cdots + c_1 z^{\tilde{n}_x-1} + z^{\tilde{n}_x}.$$

The roots of $P(z)$ *are the* $\{x_j\}$ *values of the minimal degree solution to* (1.5) *(possibly together with some zeros).*

4. *To compute the* $\{y_j\}$ *values, the same process is used with* m_k *replaced by* $-m_k$ *and the roles of* n_x *and* n_y *interchanged.*

REMARK 2. *We note that the values of* a_k *in the definition* (2.1) *are independent of* K*, since the system matrix is triangular. We therefore consider the sequence without reference to* K *in any other respect than the fact that we are only able to compute elements with* $k \leq K$ *when we are given* K *moments. The largest index of the* a_k*-sequence appearing in the matrix* A_1 *is* $n_y + n_x - 1 < K$*. In the matrices* $\tilde{A}_0, \tilde{A}_1$ *it is* $\tilde{n}_y + \tilde{n}_x = n_y - n_x + 2\tilde{n}_x \leq n_y + n_x = K$*. Hence all three matrices can be constructed from the first* K *moments. Some properties of the* A_1 *matrix are detailed in section* 6*.*

Sometimes one is not interested in finding the individual $\{x_j\}$ and $\{y_j\}$ branch values but just wants the higher moments, defined as

$$m_k = \sum_{j=1}^{n_x} x_j^k - \sum_{j=1}^{n_y} y_j^k, \tag{2.4}$$

but now for $k > K$, *given* a solution $\{x_j\} \cup \{y_j\}$ to (1.5). (That this is well defined is shown later in Theorem 4.1.) For this case there is another algorithm which has empirically proven to be more stable than first computing $\{x_j\}$ and $\{y_j\}$ from Algorithm 1 or 2 and then entering the values into (2.4). We stress that this is precisely what is needed in order to compute K-multivalued solutions of the inviscid Burgers equation in geometrical optics, following the ideas of [4].

ALGORITHM 3 (computing m_{K+1}).

1. *Construct the* A_1 *matrix as in steps* 1–2 *of Algorithm* 1*.*
2. *Let*

$$\boldsymbol{a}_0 = (a_{n_y+1}, a_{n_y+2}, \dots, a_{n_y+n_x})^T \in \mathbb{R}^{n_x},$$

and let $\bar{\boldsymbol{c}} = (c_1, c_2, \dots, c_{n_x})^T$ *be one solution to*

$$A_1\bar{\boldsymbol{c}} = -\boldsymbol{a}_0. \tag{2.5}$$

3. *The next moment is given by*

$$m_{K+1} = -(K+1)\sum_{j=1}^{n_x} c_j a_{K+1-j} - \sum_{j=1}^{K} m_j a_{K+1-j}.$$

We recall that Algorithm 1 has been shown to be numerically efficient in the paper [14]. The justification of these algorithms is given in section 5, where we show the following theorem.

THEOREM 2.1. *If a solution to* (1.5) *exists, then the following hold.*

(i) *In Algorithm* 1, *the matrix* $\tilde{A}_1$ *is nonsingular. The generalized eigenvalue problem in* (2.2) *is well defined and the generalized eigenvalues (counting algebraic multiplicity) are the* $\{x_j\}$*-values of the minimal degree solution to* (1.5) *plus* $\tilde{n}_x - D_{\min}$ *zeros. (See* (3.7) *for the definition of* $D_{\min}$*.)*

(ii) *In Algorithm* 2, $\boldsymbol{c}'$ *is well defined,*

$$P(z) = \det(zI - \tilde{A}_1^{-1}\tilde{A}_0), \tag{2.6}$$

and the roots of $P(z)$ *are the* $\{x_j\}$*-values of the minimal degree solution to* (1.5) *plus* $\tilde{n}_x - D_{\min}$ *zeros.*

(iii) *In Algorithm* 3, *the computed moment satisfies*

$$m_{K+1} = \sum_{j=1}^{n_x} x_j^{K+1} - \sum_{j=1}^{n_y} y_j^{K+1}$$

for all solutions $\{x_j\} \cup \{y_j\}$ *to* (1.5).

We postpone the proof of Theorem 2.1 to section 5. We just note here that the last point in Algorithms 1 and 2 can easily be explained by the symmetry of the problem. Indeed, the negative of (1.5),

$$-m_k = \sum_{j=1}^{n_y} y_j^k - \sum_{j=1}^{n_x} x_j^k, \qquad k = 1, \dots, K,$$

is of the same form as (1.5) itself, with the roles of n_x, $\{x_j\}$ and n_y, $\{y_j\}$ interchanged.

3. Preliminaries.

We will use three different ways of describing the solution to (1.5). First, we have a set of numbers $\{x_j\}_{j=1}^{n_x}$ and $\{y_j\}_{j=1}^{n_y}$, solving (1.5). We call those numbers branch values. Second, we have a pair of polynomials (p, q) of degrees at most n_x and n_y, respectively, in the z variable. Third, we have a pair of coefficient vectors $\boldsymbol{c} = (c_0, \dots, c_{n_x})^T \in \mathbb{R}^{n_x+1}$ and $\boldsymbol{d} = (d_0, \dots, d_{n_x})^T \in \mathbb{R}^{n_y+1}$. These three representations are related as

$$p(z) = (1 - x_1 z)\cdots(1 - x_{n_x} z) = c_0 + c_1 z + \cdots + c_{n_x-1} z^{n_x-1} + c_{n_x} z^{n_x} \tag{3.1}$$

and

$$q(z) = (1 - y_1 z)\cdots(1 - y_{n_y} z) = d_0 + d_1 z + \cdots + d_{n_y-1} z^{n_y-1} + d_{n_y} z^{n_y}. \tag{3.2}$$

It is clear that there is a one-to-one correspondence between these ways of describing the solution if we disregard the ambiguity in the ordering of the numbers $\{x_j\}$ and $\{y_j\}$. Generally, we will use the notation $\mathrm{Deg}(p)$ to denote the degree of a polynomial p, and, for a given coefficient vector $\boldsymbol{c}$, we systematically write P_c to denote the corresponding polynomial (3.1).

DEFINITION 3.1. *We call the pair of polynomials* (p, q) a (polynomial) solution *to* (1.5) *if the following hold.*

1. *The degrees of p and q are at most n_x and n_y:*

$$\text{(3.3)} \qquad \mathrm{Deg}(p) \le n_x, \qquad \mathrm{Deg}(q) \le n_y.$$

2. *They are normalized to one at the origin:*

$$\text{(3.4)} \qquad p(0) = q(0) = 1.$$

3. *Their roots $\{\tilde{x}_j\}$ and $\{\tilde{y}_j\}$ satisfy*

$$\text{(3.5)} \qquad m_k = \sum_{j=1}^{\mathrm{Deg}(p)} \tilde{x}_j^{-k} - \sum_{j=1}^{\mathrm{Deg}(q)} \tilde{y}_j^{-k}, \qquad k = 1, \dots, K.$$

We note that the roots cannot be zero because of (3.4).

Next, we have the following.

Definition 3.2. *A pair of vectors*

$$\boldsymbol{c} = (c_0, \dots, c_{n_x})^T \in \mathbb{R}^{n_x+1} \textit{ and } \boldsymbol{d} = (d_0, \dots, d_{n_y})^T \in \mathbb{R}^{n_y+1}$$

is said to be a (coefficient) solution *to* (1.5) *if the corresponding pair* (P_c, P_d) (3.1)–(3.2) *realizes a polynomial solution to* (1.5).

The number of branch values is always n_x and n_y, respectively. Some of them may be zero, and they do not need to be distinct. The number of nonzero branch values is $\mathrm{Deg}(p)$ and $\mathrm{Deg}(q)$, respectively. The degree of a solution can then also be defined.

Definition 3.3. *The* degree *of a solution to* (1.5) *is the number of nonzero x_j-values. This number is equivalent to* $\mathrm{Deg}(p)$.

Given any polynomial pair satisfying (3.4), we say that it generates the moment sequence $\{m_k\}$ if m_k is given by (3.5) for all k. In turn, each sequence of moments $\{m_k\}$ generates the corresponding $\{a_k\}$ sequence through (2.1). We define the big matrix

$$A = \begin{pmatrix} a_{n_y+1} & a_{n_y} & \cdots & a_{n_y-n_x+1} \\ a_{n_y+2} & a_{n_y+1} & \cdots & a_{n_y-n_x+2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_y+n_x} & a_{n_y+n_x-1} & \cdots & a_{n_y} \end{pmatrix} \in \mathbb{R}^{n_x \times (n_x+1)}.$$

We let the columns of A be denoted $\boldsymbol{a}_0, \dots, \boldsymbol{a}_{n_x}$, and we note that

$$\text{(3.6)} \qquad A = \begin{pmatrix} | & & | \\ \boldsymbol{a}_0 & \cdots & \boldsymbol{a}_{n_x} \\ | & & | \end{pmatrix} = \begin{pmatrix} & & | \\ & A_0 & \boldsymbol{a}_{n_x} \\ & & | \end{pmatrix} = \begin{pmatrix} | & & \\ \boldsymbol{a}_0 & A_1 & \\ | & & \end{pmatrix}.$$

Hence A_0 and A_1 constitute the first and last n_x columns of A, respectively. When $\boldsymbol{a}_0 \in \operatorname{range} A_1$ and $\boldsymbol{a}_0 \neq 0$, let

$$\text{(3.7)} \qquad D_{\min} = \operatorname{argmin}_{j>0} \boldsymbol{a}_0 \in \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_j\},$$

and set $D_{\min} = 0$ if $\boldsymbol{a}_0 = 0$. Moreover, define

$$\text{(3.8)} \qquad D_{\max} = D_{\min} + n_x - \operatorname{rank} A_1.$$

4. Existence and uniqueness of solutions.

In this section we prove results on the existence and uniqueness of solutions to (1.5). We aim at establishing the following theorem.

Theorem 4.1.

(i) *There exists a solution to* (1.5) *if and only if*

$$\boldsymbol{a}_0 \in \operatorname{range}(A_1). \tag{4.1}$$

(ii) *If d is the degree of a solution to* (1.5)*, then $D_{\min} \le d \le D_{\max}$.*

(iii) *When* (4.1) *holds, there is a unique solution (p^*, q^*) of minimal degree $D_{\min}$. For this solution, $x_j \neq y_i$ for all indices i, j representing nonzero branch values. Moreover,* $\operatorname{Deg}(q^*) \le n_y - n_x + \operatorname{rank} A_1$ *with equality if* $D_{\min} < \operatorname{rank} A_1$.

(iv) *When* (4.1) *holds, a polynomial pair (p, q) is a solution if and only if $p = p^* r$ and $q = q^* r$, where $r(z)$ is a polynomial satisfying $r(0) = 1$ and* $\operatorname{Deg}(r) \le D_{\max} - D_{\min}$.

(v) *The minimal degree solution is the only solution to* (1.5) *if and only if the matrix A_1 is nonsingular.*

(vi) *Let $\{x_j\}$ and $\{y_j\}$ be a solution to* (1.5)*. Then the higher moments defined in* (2.4) *are well defined.*

Let us proceed with several remarks.

Remark 3. *In particular, it follows from* (i) *that there exists a solution as soon as the matrix A_1 is nonsingular.*

Remark 4. *Since* (1.5) *is a system of polynomial equations of degree K, one could expect there to be a finite number of solutions, typically K solutions. However, because of the special structure of the equations there is either one unique solution (when A_1 is nonsingular) or infinitely many solutions (when A_1 is singular).*

Remark 5. *The form $(p^* r, q^* r)$ of solutions can also be stated as follows: All solutions have a core set of values $\{x_j\}$, $j = 1, \ldots, \operatorname{Deg}(p^*) = D_{\min}$, and $\{y_i\}$, $i = 1, \ldots, \operatorname{Deg}(q^*)$, corresponding to nonzero branch values of the minimal degree solution, where $x_j \neq y_i$ for all those i, j. One can then add an optional set of nonzero branch values $\{x_{D_{\min}+j}\}$ and $\{y_{\operatorname{Deg}(q^*)+j}\}$ for $j = 1, \ldots, D_{\max} - D_{\min}$ such that $x_{D_{\min}+j} = y_{\operatorname{Deg}(q^*)+j}$.*

To prove this theorem we first establish some utility results in the next subsection. We then derive different ways of characterizing the solution in section 4.2 which are subsequently used to prove Theorem 4.1 in section 4.3.

4.1. Utility results.

We start with a useful lemma on Taylor coefficients for a product of functions.

Lemma 4.2. *Suppose f, g, and h are analytic functions in a neighborhood of zero satisfying $f(z) = g(z)h(z)$. Let f have the Taylor expansion*

$$f(z) = \sum_{k=0}^{\infty} f_k z^k,$$

and let $\{g_k\}$ and $\{h_k\}$ be the corresponding coefficients for $g(x)$ and $h(x)$, respectively. Then

$$f_k = \sum_{j=0}^{k} g_j h_{k-j}. \tag{4.2}$$

Proof. Since the functions are analytic, the coefficients are given as

$$f_k = \frac{1}{k!}\frac{d^k}{dz^k}f(z)\Big|_{z=0} = \frac{1}{k!}\frac{d^k}{dz^k}g(z)h(z)\Big|_{z=0} = \frac{1}{k!}\sum_{j=0}^{k} c_{jk}g^{(j)}(0)h^{(k-j)}(0),$$

where $c_{jk} = k!/j!(k-j)!$ are the binomial coefficients. But $g^{(j)}(0) = j!g_j$ and $h^{(k-j)}(0) = (k-j)!h_{k-j}$, and therefore (4.2) follows. □

REMARK 6. *The discrete convolution* (4.2) *is, in fact, precisely an elementwise description of multiplication of a lower triangular $k \times k$ Toeplitz matrix by a vector. In the notation of* [15], *it would read $\boldsymbol{f} = \mathcal{T}(\boldsymbol{g})\boldsymbol{h}$.*

As was already known by Markov, the exponential transform of the moment sequence plays an important role in the analysis of these problems; see, e.g., [1, 2]. We show here that $\{a_k\}$ is a version of the exponential transform of $\{m_k\}$.

LEMMA 4.3. *Suppose $\{m_k\}$ is generated by the polynomials $p(z)$ and $q(z)$ and $\{a_k\}$ is generated by $\{m_k\}$. Let $m(z)$ be defined as*

$$m(z) = m_1 z + \frac{1}{2}m_2 z^2 + \frac{1}{3}m_3 z^3 + \cdots . \tag{4.3}$$

Then, if (3.4) *holds,*

$$e^{m(z)} = \frac{q(z)}{p(z)} = a_0 + a_1 z + a_2 z^2 + \cdots , \tag{4.4}$$

written as its Taylor expansion around $z = 0$.

Proof. Let us first show that $m(z)$ is a well-defined analytic function at zero. We have

$$\begin{aligned} m(z) &= \textstyle\sum_{k=0}^{\infty} \frac{m_k z^k}{k} \\ &= \textstyle\sum_{k=0}^{\infty}\sum_{j=1}^{n_x} \frac{x_j^k z^k}{k} - \sum_{k=0}^{\infty}\sum_{j=1}^{n_y} \frac{y_j^k z^k}{k} \\ &= -\textstyle\sum_{j=1}^{n_x} \log(1 - x_j z) + \sum_{j=1}^{n_y} \log(1 - y_j z). \end{aligned}$$

The last step is allowed when $|z| < 1/\max_{ij}(|x_j|, |y_i|)$, which is true for small enough z since $p(0) \neq 0$. This also shows that the function is analytic at zero. Moreover,

$$e^{m(z)} = \frac{\prod_{j=1}^{n_y}(1 - y_j z)}{\prod_{j=1}^{n_x}(1 - x_j z)} = \frac{q(z)}{p(z)}.$$

Finally, setting $a(z) := \exp(m(z))$ and differentiating gives

$$za'(z) = zm'(z)a(z),$$

where all three functions are analytic at zero. Let $a(z)$ have the Taylor coefficients $\{\tilde{a}_k\}$. Then $za'(z) = \tilde{a}_1 z + 2\tilde{a}_2 z^2 + 3\tilde{a}_3 z^3 \cdots$ and clearly $zm'(z) = m_1 z + m_2 z^2 + \cdots$. By Lemma 4.2, for $k \geq 1$,

$$k\tilde{a}_k = \sum_{j=1}^{k} m_j \tilde{a}_{k-j}.$$

Since $\tilde{a}_0 = q(0)/p(0) = 1$, we see that a_k and $\tilde{a}_k$ satisfy the same nonsingular linear system of equations (2.1), and therefore $a_k = \tilde{a}_k$, showing (4.4). □

We now have the following basic characterization of a solution.

LEMMA 4.4. *Suppose $p(z)$ and $q(z)$ are two polynomials satisfying* (3.3), (3.4)*. They form a polynomial solution to* (1.5) *if and only if their quotient has the Taylor expansion around $z = 0$*

$$\frac{q(z)}{p(z)} = a_0 + a_1 z + \cdots + a_K z^K + O\left(z^{K+1}\right), \tag{4.5}$$

where $\{a_k\}$ is generated by $\{m_k\}$. Moreover, if (p,q) is a solution, then $(\bar{p},\bar{q})$ is also a solution if and only if the pair satisfies (3.3), (3.4)*, and $\bar{p}/\bar{q} = p/q$, where these fractions are defined.*

Proof. Let $\{\tilde{m}_k\}$ be generated by p and q, and suppose (4.5) holds. Then, as in the proof of Lemma 4.3 for $1 \le k \le K$,

$$ka_k = \sum_{j=1}^{k} \tilde{m}_j a_{k-j}.$$

Since $\{m_k\}$ satisfy the linear system (2.1), we have after subtraction

$$m_n - \tilde{m}_n = -\sum_{k=1}^{n-1} (m_k - \tilde{m}_k) a_{n-k}, \qquad m_1 = \tilde{m}_1,$$

for $n = 2, \ldots, K$. By induction $\tilde{m}_k = m_k$ for $1 \le k \le K$, showing that (p,q) solves (1.5). On the other hand, if (p,q) is a solution, then (4.5) must hold by (4.4) in Lemma 4.3.

For the last statement, the "if" part is obvious since both pairs then satisfy (4.5). To show the "only if" part, suppose both (p,q) and $(\bar{p},\bar{q})$ are solutions. By definition they satisfy (3.3), (3.4), and by (4.5),

$$\frac{\bar{q}(z)}{\bar{p}(z)} - \frac{q(z)}{p(z)} = \frac{\bar{q}(z)p(z) - \bar{p}(z)q(z)}{\bar{p}(z)p(z)} = O(z^{K+1}).$$

Since $\bar{p}(0)p(0) = 1$, we must have that $(\bar{q}(z)p(z) - \bar{p}(z)q(z))/z^{K+1}$ is bounded as $z \to 0$. But since the degree of $\bar{q}p - \bar{p}q$ is at most $K = n_x + n_y$, this is possible only if it is identically zero. Hence $\bar{q}(z)p(z) = \bar{p}(z)q(z)$, which concludes the proof. □

4.2. Characterization of the solution. In this section we show three propositions that characterize solutions to (1.5) in terms of polynomials, coefficient vectors, and the column vectors of the A matrix in (3.6). We start by expressing the uniqueness properties of the solution in terms of its polyomial representation.

PROPOSITION 4.5. *Suppose the pairs (p,q) and $(\bar{p},\bar{q})$ are both polynomial solutions to* (1.5)*. Then the following hold.*

(i) $\mathrm{Deg}(p) - \mathrm{Deg}(q) = \mathrm{Deg}(\bar{p}) - \mathrm{Deg}(\bar{q})$.

(ii) *If $\mathrm{Deg}(\bar{p}) \le \mathrm{Deg}(p)$, and if there is no polynomial $r(z)$ such that $p = \bar{p}r$, then there is another solution $(\tilde{p},\tilde{q})$ with $\mathrm{Deg}(\tilde{p}) < \mathrm{Deg}(p)$. In particular, if $\mathrm{Deg}(p) = \mathrm{Deg}(\bar{p})$ but $p \ne \bar{p}$, there is such a lower degree solution.*

(iii) *If $\mathrm{Deg}(\bar{p}) \le \mathrm{Deg}(p)$, any polynomial pair $(\bar{p}r, \bar{q}r)$ is a solution if $r(z)$ is a polynomial satisfying $r(0) = 1$ and $\mathrm{Deg}(r) \le \mathrm{Deg}(p) - \mathrm{Deg}(\bar{p})$. In particular, if $\mathrm{Deg}(\bar{p}) \le m \le \mathrm{Deg}(p)$, there is a solution $(\tilde{p},\tilde{q})$ with $\mathrm{Deg}(\tilde{p}) = m$.*

Proof.

(i) The statement follows directly from Lemma 4.4, since $\bar{q}p = \bar{p}q$ implies that

$$\mathrm{Deg}(\bar{q}) + \mathrm{Deg}(p) = \mathrm{Deg}(\bar{p}) + \mathrm{Deg}(q).$$

(ii) We let

$$p(z) = r_p(z)\bar{p}(z) + s_p(z), \qquad q(z) = r_q(z)\bar{q}(z) + s_q(z)$$

be the unique polynomial decomposition of (p, q) such that r_p, r_q, s_p, s_q are polynomials, $\mathrm{Deg}(s_p) < \mathrm{Deg}(\bar{p})$, and $\mathrm{Deg}(s_q) < \mathrm{Deg}(\bar{q})$. Since $\bar{p}q = p\bar{q}$ by Lemma 4.4, we get

$$\bar{p}\bar{q}(r_q - r_p) = \bar{q}s_p - \bar{p}s_q.$$

Unless $r_q = r_p$, the degree of the left-hand side is at least $\mathrm{Deg}(\bar{p}) + \mathrm{Deg}(\bar{q})$, while the degree of the right-hand side is at most

$$\max\left(\mathrm{Deg}(\bar{q}) + \mathrm{Deg}(s_p),\ \mathrm{Deg}(\bar{p}) + \mathrm{Deg}(s_q)\right) < \mathrm{Deg}(\bar{q}) + \mathrm{Deg}(\bar{p}).$$

Hence, $r_q = r_p$ and $\bar{q}s_p = \bar{p}s_q$. Since $\bar{q}, \bar{p} \not\equiv 0$, it follows that s_p and s_q are either both zero or both nonzero. Suppose $s_p \not\equiv 0$ and $s_q \not\equiv 0$. Write $s_p(z) = z^{m_p}\tilde{s}_p(z)$ and $s_q(z) = z^{m_q}\tilde{s}_q(z)$, where $\tilde{s}_p(0) \neq 0$ and $\tilde{s}_q(0) \neq 0$. Since

$$z^{m_p}\tilde{s}_p(z)\bar{q}(z) = z^{m_q}\tilde{s}_q(z)\bar{p}(z)$$

and also $\bar{q}(0) = \bar{p}(0) = 1$, the lowest degree term in the left- and right-hand side polynomials are z^{m_p} and z^{m_q}, respectively, and therefore $m_p = m_q$. Consequently,

$$\tilde{s}_p(z)\bar{q}(z) = \tilde{s}_q(z)\bar{p}(z)$$

and $\tilde{s}_p(0) = \tilde{s}_q(0)$. We can then take $\tilde{p}(z) = \tilde{s}_p(z)/\tilde{s}_p(0)$ and $\tilde{q}(z) = \tilde{s}_q(z)/\tilde{s}_q(0)$. They satisfy

$$\tilde{p}(z)\bar{q}(z) = \tilde{q}(z)\bar{p}(z), \qquad \tilde{p}(0) = \tilde{q}(0) = 1,$$

while $\mathrm{Deg}(\tilde{p}) = \mathrm{Deg}(\tilde{s}_p) \leq \mathrm{Deg}(s_p) < \mathrm{Deg}(p)$ and similarly $\mathrm{Deg}(\tilde{q}) < \mathrm{Deg}(q) \leq n_y$. Hence $(\tilde{p}, \tilde{q})$ is a polynomial solution by Lemma 4.4. It has degree strictly less than (p, q), which shows the first statement in (ii). If $\mathrm{Deg}(p) = \mathrm{Deg}(\bar{p})$ and $p \neq \bar{p}$, then there is no $r(z)$ satisfying the requirements, showing the second statement in (ii).

(iii) We finally let $r(z)$ be any polynomial with $\mathrm{Deg}(r) \leq \mathrm{Deg}(p) - \mathrm{Deg}(\bar{p})$ and $r(0) = 1$. We then set $\tilde{p} = \bar{p}r$ and $\tilde{q} = \bar{q}r$. These polynomials trivially satisfy (3.4) and (4.5). Since $\mathrm{Deg}(\tilde{p}) = \mathrm{Deg}(r) + \mathrm{Deg}(\bar{p}) \leq \mathrm{Deg}(p) \leq n_x$ and

$$\mathrm{Deg}(\tilde{q}) = \mathrm{Deg}(r) + \mathrm{Deg}(\bar{q}) \leq \mathrm{Deg}(p) - \mathrm{Deg}(\bar{p}) + \mathrm{Deg}(\bar{q}) = \mathrm{Deg}(q) \leq n_y,$$

they also satisfy (3.3) and thus are a polynomial solution by Lemma 4.4. In particular, we can take $r(z)$ of degree m. □

A solution to (1.5) can also be characterized in terms of the coefficient vectors. We have the following proposition.

PROPOSITION 4.6. *The pair* $\boldsymbol{c} = (c_0, \ldots, c_{n_x})^T \in \mathbb{R}^{n_x+1}$ *and* $\boldsymbol{d} = (d_0, \ldots, d_{n_y})^T \in \mathbb{R}^{n_y+1}$ *is a coefficient solution to* (1.5) *if and only if*

(i) $c_0 = 1$,

(ii) $\boldsymbol{c}$ *is in the null-space of* A, *and*

(iii)

$$d_k = \sum_{j=0}^{\min(k,n_x)} c_j a_{k-j}, \qquad k = 0, \dots, n_y. \tag{4.6}$$

Proof. Suppose first that $\boldsymbol{c}$ is in the null-space of A, $c_0 = 1$, and $\{d_k\}$ is given by (4.6). Extend the coefficient sequences by setting $c_k = 0$ for $k > n_x$ and $d_k = 0$ for $k > n_y$. Since $\boldsymbol{c}$ is in the null-space of A, we get $\sum_{j=0}^{k} c_j a_{k-j} = 0$ when $n_y + 1 \leq k \leq n_x + n_x = K$, and in conclusion

$$d_k = \sum_{j=0}^{k} c_j a_{k-j}, \qquad k = 0, \dots, K. \tag{4.7}$$

Upon noting that $\{c_k\}_{k=0}^{\infty}$ and $\{d_k\}_{k=0}^{\infty}$ are the Taylor coefficients of P_c and P_d, and since $P_c(0) = c_0 = 1$, $P_d(0) = d_0 = a_0 c_0 = 1$, Lemma 4.2 shows that

$$P_d(z) = P_c(z) \left[a_0 + a_1 z + \cdots + a_K z^K + O\left(z^{K+1}\right) \right], \tag{4.8}$$

and by Lemma 4.4 we have that (P_c, P_d) is a solution to (1.5). Conversely, if (P_c, P_d) is a solution, then $c_0 = P_c(0) = 1$, and by Lemma 4.2 we get that (4.7) holds. For $k = n_y + 1, \dots, K$ this also implies that $\boldsymbol{c}$ is in the null-space of A. □

The final proposition of this section relates the degree of the solution to the column vectors of A and the linear spaces they span.

PROPOSITION 4.7. *Let* $V_j = \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_j\}$ *and* $V_j^0 = \operatorname{span}\{\boldsymbol{a}_0, \dots, \boldsymbol{a}_j\}$. *Set* $V_0 = V_{-1}^0 = \emptyset$. *Then the following hold.*

(i) *There is a solution if and only if* $\boldsymbol{a}_0 \in V_{n_x} = \operatorname{Range}(A_1)$.

(ii) *There is a solution of degree* $j \geq 0$ *if and only if*

$$\boldsymbol{a}_0 \in V_j \quad \textit{and} \quad \boldsymbol{a}_j \in V_{j-1}^0. \tag{4.9}$$

(iii) *When* $\boldsymbol{a}_0 \in V_{n_x}$, *then*

$$\boldsymbol{a}_0 \in V_d, \qquad V_d^0 = V_d$$

if and only if $d \geq D_{\max}$.

(iv) *When* $\boldsymbol{a}_0 \in V_{n_x}$, *the vectors*

$$\boldsymbol{a}_1, \dots, \boldsymbol{a}_{D_{\min}}$$

(when $D_{\min} > 0$*),*

$$\boldsymbol{a}_{D_{\max}+1}, \dots, \boldsymbol{a}_{n_x}$$

(when $D_{\max} < n_x$*) are all linearly independent. Moreover,*

$$\boldsymbol{a}_j \in V_{D_{\min}}, \quad V_j = V_{D_{\min}}, \qquad j = D_{\min}, \dots, D_{\max}.$$

Proof.

(i) By Proposition 4.6 there exists a solution to (1.5) if and only if there is a coefficient vector $\boldsymbol{c} = (1, \boldsymbol{c}')^T$ in the null-space of A, i.e.,

$$A\boldsymbol{c} = A_1\bar{\boldsymbol{c}} + \boldsymbol{a}_0 = 0.$$

But such a vector $\bar{\boldsymbol{c}}$ exists if and only if $\boldsymbol{a}_0$ is in the range of A_1. This shows (i).

(ii) Again by Proposition 4.6 there is a solution of degree j if and only if there is a vector $\boldsymbol{c} = (c_0, c_1, \ldots, c_j, 0, \ldots, 0)^T$ such that

$$0 = A\boldsymbol{c} = c_0\boldsymbol{a}_0 + c_1\boldsymbol{a}_1 + \cdots + c_j\boldsymbol{a}_j, \tag{4.10}$$

with $c_j \neq 0$ and $c_0 = 1$. For $j = 0$ this is clearly equivalent to $\boldsymbol{a}_0 = 0$ or $\boldsymbol{a}_0 \in V_0 = V^0_{-1}$. For $j > 0$ the existence of c_j-coefficients satisfying (4.10) is equivalent to the left condition in (4.9). Moreover, if $\boldsymbol{a}_j \neq V^0_{j-1} = \operatorname{span}\{\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{j-1}\}$, then we must have $c_j = 0$ to satisfy (4.10), and $\boldsymbol{c}$ cannot represent a solution of degree j. On the other hand, if $c_j = 0$ and $\boldsymbol{a}_j = c'_0\boldsymbol{a}_0 + \cdots + c'_{j-1}\boldsymbol{a}_{j-1}$ for some nonzero coefficients c'_k, then $\boldsymbol{a}_0 + c''_1\boldsymbol{a}_1 + \cdots + c''_{j-1}\boldsymbol{a}_{j-1} + \boldsymbol{a}_j = 0$, with $c''_k = (1 + c'_0)c_k - c'_k$, represents a solution of degree j. This shows (ii).

(iii) The statement is obvious in case $D_{\min} = 0$. If $D_{\min} > 0$, there are scalars such that

$$\boldsymbol{a}_0 = v_1\boldsymbol{a}_1 + \cdots + v_{D_{\min}}\boldsymbol{a}_{D_{\min}}, \tag{4.11}$$

by (3.7). Hence, $\boldsymbol{a}_0 \in V_{D_{\min}}$ and since the V_j spaces are nested, $V_j \subset V_{j+1}$, we have $\boldsymbol{a}_0 \in V_d$ for $d \geq D_{\min}$. Moreover, the minimal property of $D_{\min}$ ensures that $v_{D_{\min}} \neq 0$ in (4.11), so that $\boldsymbol{a}_0 \notin V_d$ when $d < D_{\min}$.

(iv) To show that when $D_{\min} > 0$ the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{D_{\min}}$ are linearly independent, we use (4.11) and note that $P_c(z)$ with $\boldsymbol{c} = (1, -v_1, \ldots, -v_{D_{\min}}, 0, \ldots, 0)^T$ is a polynomial solution to (1.5). Suppose now that there are nonzero coefficients c'_j such that

$$c'_1\boldsymbol{a}_1 + \cdots + c'_{D_{\min}}\boldsymbol{a}_{D_{\min}} = 0.$$

Then $P_{c'}$ with $\boldsymbol{c}' = (1, c'_1 - v_1, \ldots, c_{D_{\min}} - v_{D_{\min}}, 0, \ldots, 0)^T$ is another polynomial solution to (1.5). Moreover, by the minimality property of $D_{\min}$ we must have $c_{D_{\min}} - v_{D_{\min}} \neq 0$ and therefore $\operatorname{Deg}(P_c) = \operatorname{Deg}(P_{c'}) = D_{\min}$. But by (ii) in Proposition 4.5 this implies that there is yet another solution $P_{c''}$ of degree strictly less than $D_{\min}$. Hence, there are coefficients c''_j such that

$$\boldsymbol{a}_0 + c''_1\boldsymbol{a}_1 + \cdots + c''_d\boldsymbol{a}_d = 0,$$

with $d < D_{\min}$, contradicting (3.7). The vectors must therefore be linearly independent.

Suppose $D^* \geq D_{\min}$ is the highest degree of an existing solution. Since $P_c(z)$ is a solution of degree $D_{\min}$, we get from (iii) in Proposition 4.5 that there are solutions of all intermediate degrees $D_{\min}, \ldots, D^*$. Hence, from (ii), $\boldsymbol{a}_j \in V^0_{j-1}$ for $j = D_{\min}, \ldots, D^*$, and, from (iii), $\boldsymbol{a}_j \in V_{j-1}$ for $j = D_{\min}+1, \ldots, D^*$. Noting that if $\boldsymbol{a}_{j+1} \in V_j$, then $V_j = V_{j+1}$, we can conclude inductively that $V_{D_{\min}} = \cdots = V_{D^*}$ and $\boldsymbol{a}_j \in V_{D_{\min}}$ for $j = D_{\min}, \ldots, D^*$. We now have three different cases.

1. If $D^* = n_x$, then $V_{D_{\min}} = V_{n_x}$, and by (3.8) we get $D^* = \operatorname{rank} A_1 - D_{\min} + D_{\max} = \dim V_{n_x} - D_{\min} + D_{\max} = \dim V_{D_{\min}} - D_{\min} + D_{\max} = D_{\max}$ since either $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{D_{\min}}$ are linearly independent or $D_{\min} = 0$ and $V_{D_{\min}} = \emptyset$. This shows (iv) for $D^* = n_x$.
2. If $D^* < n_x$ and $D_{\min} = 0$, then $V_{D_{\min}} = V_{D^*} = \emptyset$, and

   $$V_{n_x} = \operatorname{span}\{\boldsymbol{a}_{D^*+1}, \dots, \boldsymbol{a}_{n_x}\}. \tag{4.12}$$

   Suppose there are nonzero coefficients α_k such that

   $$\alpha_{D^*+1}\boldsymbol{a}_{D^*+1} + \cdots + \alpha_{n_x}\boldsymbol{a}_{n_x} = 0,$$

   and let k^* be the highest index of all nonzero coefficients, $\alpha_{k^*} \neq 0$. Then $\boldsymbol{a}_{k^*} \in V^0_{k^*-1}$ and there is a solution of degree k^* by (ii), which is in contradiction to the definition of D^*. Hence, the vectors in (4.12) must be linearly independent and

   $$D^* = n_x - \dim V_{n_x} = D_{\min} + n_x - \operatorname{rank} A_1 = D_{\max},$$

   showing (iv) for this case.
3. If $D^* < n_x$ and $D_{\min} > 0$, we have

   $$V_{n_x} = \operatorname{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_{D_{\min}}, \boldsymbol{a}_{D^*+1}, \dots, \boldsymbol{a}_{n_x}\}. \tag{4.13}$$

   Suppose there are nonzero coefficients α_k such that

   $$\alpha_1\boldsymbol{a}_1 + \cdots + \alpha_{D_{\min}}\boldsymbol{a}_{D_{\min}} + \cdots + \alpha_{D^*+1}\boldsymbol{a}_{D^*+1} + \cdots + \alpha_{n_x}\boldsymbol{a}_{n_x} = 0.$$

   Since $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{D_{\min}}$ are linearly independent, at least one α_k with $k > D^*$ must be nonzero. By the same argument as above in case 2 we then get a contradiction, and the vectors in (4.13) must be linearly independent. Hence,

   $$D^* = D_{\min} + n_x - \dim V_{n_x} = D_{\min} + n_x - \operatorname{rank} A_1 = D_{\max},$$

   showing this final case. □

4.3. Proof of Theorem 4.1. To prove Theorem 4.1, we essentially have to combine the results from Propositions 4.5 and 4.7. The statement (i) is given directly by (i) in the latter. For the remaining points we have the following.

(ii) From (ii) in Proposition 4.7 we see that $\boldsymbol{a}_0 \in V_d$ and $\boldsymbol{a}_d \in V^0_{d-1}$. It follows from (iii) in Proposition 4.7 that $d \geq D_{\min}$. On the other hand, if $D_{\max} < n_x$ and $d > D_{\max}$, it says that $V^0_{d-1} = V_{d-1}$. Hence, $\boldsymbol{a}_d \in V_{d-1}$, which contradicts the linear independence of $\boldsymbol{a}_{D_{\max}}, \dots, \boldsymbol{a}_{n_x}$ established in point (iv) of Proposition 4.7.

(iii) We note that by (3.7) there are scalars $v_1, \dots, v_{D_{\min}}$ such that

$$\boldsymbol{a}_0 = v_1\boldsymbol{a}_1 + \cdots + v_{D_{\min}}\boldsymbol{a}_{D_{\min}}. \tag{4.14}$$

Hence, $\boldsymbol{a}_0 \in V_{D_{\min}}$, and since $v_{D_{\min}} \neq 0$, we also have $\boldsymbol{a}_{D_{\min}} \in V^0_{D_{\min}}$. By (ii) in Proposition 4.7 there is thus a solution of degree $D_{\min}$ which we denote (p^*, q^*). Since $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{D_{\min}}$ are linearly independent by (iii) in Proposition 4.7, the coefficients in (4.14) are unique, and therefore the $D_{\min}$-degree solution is also unique. Moreover, suppose that $x_j = y_i = x^* \neq 0$ for some i, j. Then p^* and q^* would have a common factor $(1-zx^*)$, and by Lemma 4.4 $\bar{p}(z) := p^*(z)/(1-zx^*)$ and $\bar{q}(z) := q^*(z)/(1-zx^*)$ would also be a solution. But this is impossible since $\operatorname{Deg}(\bar{p}) < \operatorname{Deg}(p^*) = D_{\min}$. By (iv), shown below, a solution is given by (p^*r, q^*r), where $r(0) = 1$ and $\operatorname{Deg}(r) = D_{\max} - D_{\min}$. Hence $n_y \geq \operatorname{Deg}(q^*r) = \operatorname{Deg}(q^*) + n_x - \operatorname{rank} A_1$. Suppose finally that $D_{\min} < \operatorname{rank} A_1$ and that $\operatorname{Deg}(q^*) < n_y - n_x + \operatorname{rank} A_1$. Let $\operatorname{Deg}(r) = D_{\max} + 1 - D_{\min}$. Then (p^*r, q^*r) is still a solution by Lemma 4.4 since (p^*, q^*) is a solution, $\operatorname{Deg}(p^*r) = D_{\max} + 1 = n_x + D_{\min} + 1 - \operatorname{rank} A_1 \leq n_x$, and

$$\operatorname{Deg}(q^*r) < n_y - n_x + \operatorname{rank} A_1 + D_{\max} + 1 - D_{\min} = n_y + 1.$$

This contradicts (ii) and therefore $\operatorname{Deg}(q^*) = n_y - n_x + \operatorname{rank} A_1$, concluding the proof of (iii).

(iv) We first note that there exists a solution of degree $D_{\max}$ by Proposition 4.7 since, if $D_{\max} > D_{\min}$, we have $\boldsymbol{a}_0 \in V^0_{D_{\max}-1}$ and $\boldsymbol{a}_{D_{\max}} \in V_{D_{\min}} = V_{D_{\max}-1} = V^0_{D_{\max}-1}$. Hence, (iii) in Proposition 4.5 shows that any polynomial pair of the stated type is a solution. On the other hand, if the polynomial solution is not of this type, then (ii) in Proposition 4.5 says there is a solution of degree strictly less than $D_{\min}$, contradicting (ii) above.

(v) We suppose first that A_1 is nonsingular. Then $\operatorname{rank} A_1 = n_x$ so that $D_{\min} = D_{\max}$ and the uniqueness is given by (iii) above. If, on the contrary, A_1 is singular, then $D_{\max} > D_{\min}$, and since we can then pick infinitely many polynomials $r(z)$ in (iv), we have infinitely many solutions.

(vi) This is a consequence of (iv). The solution can be represented by (p^*r, q^*r) for some polynomial $r(z)$ with $r(0) = 1$. Let $1/x_j$ for $j = 1, \ldots, D_{\min}$ and $1/y_j$ for $j = 1, \ldots, \operatorname{Deg}(q^*)$ be the roots of $p^*(z)$ and $q^*(z)$, respectively. Let $1/z_j$ for $j = 1, \ldots, \operatorname{Deg}(r)$ be the roots of $r(z)$. Then

$$m_k = \sum_{j=1}^{D_{\min}} x_j^k + \sum_{j=1}^{\operatorname{Deg}(r)} z_j^k - \sum_{j=1}^{\operatorname{Deg}(q^*)} y_j^k - \sum_{j=1}^{\operatorname{Deg}(r)} z_j^k = \sum_{j=1}^{D_{\min}} x_j^k - \sum_{j=1}^{\operatorname{Deg}(q^*)} y_j^k,$$

which is independent of $r(z)$ and uniquely determined because (p^*, q^*) is unique.

5. Proof of Theorem 2.1.

We can now use the results in section 4 to prove Theorem 2.1.

(i)–(ii) To show the statements about Algorithms 1 and 2 we consider the reduced problem

$$m_k = \sum_{j=1}^{\tilde{n}_x} \tilde{x}_j^k - \sum_{j=1}^{\tilde{n}_y} \tilde{y}_j^k, \qquad k = 1, \ldots, \tilde{K}, \tag{5.1}$$

where $\tilde{n}_x = \operatorname{rank} A_1 \leq n_x$, $\tilde{n}_y = n_y - n_x + \tilde{n}_x \leq n_y$, and $\tilde{K} = \tilde{n}_x + \tilde{n}_y \leq K$. The moments m_k in the left-hand side are the same as in (1.5). First, we consider the minimal solution (p^*, q^*) of (1.5). By (iv) in Proposition 4.7 we must have $\operatorname{Deg}(p^*) = D_{\min} \leq \operatorname{rank} A_1 = \tilde{n}_x$. Moreover, by (iii) in Theorem 4.1,

$$\operatorname{Deg}(q^*) \leq n_y - n_x + \operatorname{rank} A_1 = \tilde{n}_y.$$

It follows from Lemma 4.4 that (p^*, q^*) is also a solution to (5.1). Second, let $(\tilde{p}^*, \tilde{q}^*)$ be the minimal degree solution to (5.1). Then by (iv) in Theorem 4.1 there is a polynomial $r(z)$ with $r(0) = 1$ such that $p^* = \tilde{p}^* r$ and $q^* = \tilde{q}^* r$. But then $(\tilde{p}^*, \tilde{q}^*)$ is also a solution to (1.5) by Lemma 4.4. By the uniqueness of the minimal degree solution of (1.5) it follows that $r \equiv 1$ and $p^* = \tilde{p}^*$, $q^* = \tilde{q}^*$. Suppose now that there is another polynomial $r(z)$ with $r(0) = 1$, $\mathrm{Deg}(r) > 0$ such that $(p^* r, q^* r)$ is a solution to (5.1). Then $\mathrm{Deg}(p^* r) = D_{\min} + \mathrm{Deg}(r) \le \tilde{n}_x = \operatorname{rank} A_1$. Hence, $D_{\min} < \operatorname{rank} A_1$, and therefore by (iii) in Theorem 4.1 we have $\mathrm{Deg}(q^*) = n_y - n_x + \operatorname{rank} A_1 = \tilde{n}_y$. Thus, $\mathrm{Deg}(q^* r) > \tilde{n}_y$, which is impossible if $(p^* r, q^* r)$ is a solution. Hence, (p^*, q^*) is the unique solution to (5.1), and therefore $\tilde{A}_1$ is nonsingular by (v) in Theorem 4.1.

Since $\tilde{A}_1$ is invertible, the generalized eigenvalue problem (2.2) and $\boldsymbol{c}'$ are well defined. Moreover, we can construct $\tilde{A}_1^{-1} \tilde{A}_0$. By (2.3),

$$\tilde{A}_1^{-1} \tilde{A}_0 = \begin{pmatrix} -c_1 & 1 & 0 & \cdots & 0 \\ -c_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -c_{\tilde{n}_x - 1} & 0 & 0 & \ddots & 1 \\ -c_{\tilde{n}_x} & 0 & 0 & \cdots & 0 \end{pmatrix},$$

which is a companion matrix. It is well known that for those matrices the elements in the first column are the coefficients of its characteristic polynomial. This is shown as follows: let M_{ij} be the minor of $V := zI - \tilde{A}_1^{-1} \tilde{A}_0$, i.e., the determinant of the matrix obtained by removing row i and column j. Then, the determinant can be expanded by minors, for any j:

$$\det(V) = \sum_{i=1}^{\tilde{n}_x} (-1)^{i+j} v_{ij} M_{ij}, \qquad V = \{v_{ij}\}.$$

Taking $j = 1$, we get $M_{i,1} = \det(\mathrm{diag}(z, \ldots, z, -1, \ldots, -1))$ with $i - 1$ occurrences of -1, so that $M_{i,1} = z^{\tilde{n}_x - i} (-1)^{i-1}$. Therefore,

$$\begin{aligned} \det(V) &= (-1)^2 (c_1 + z) M_{1,1} + \textstyle\sum_{i=2}^{\tilde{n}_x} (-1)^{i+1} c_i M_{i,1} \\ &= c_1 z^{\tilde{n}_x - 1} + z^{\tilde{n}_x} + \textstyle\sum_{i=2}^{\tilde{n}_x} c_i z^{\tilde{n}_x - i} \\ &= P(z), \end{aligned}$$

which is exactly (2.6). This shows that the results of Algorithms 1 and 2 are identical, since the generalized eigenvalues in (2.2) are exactly the roots of $P(z)$.

It remains to show what the roots are. Let $\tilde{A} = [\tilde{a}_0 \ \tilde{A}_1]$ be the A matrix related to (5.1). Clearly, $\boldsymbol{c} = (1, \boldsymbol{c}'^T)^T$ is in the null-space of $\tilde{A}$, and hence $P_c(z)$ is the unique solution to (5.1). But for $z \neq 0$,

$$\begin{aligned} P(z) &= c_{\tilde{n}_x} + c_{\tilde{n}_x - 1} z + \cdots + c_1 z^{\tilde{n}_x - 1} + z^{\tilde{n}_x} \\ &= z^{\tilde{n}_x} \left(\frac{c_{\tilde{n}_x}}{z^{\tilde{n}_x}} + \frac{c_{\tilde{n}_x - 1}}{z^{\tilde{n}_x - 1}} + \cdots + \frac{c_1}{z} + 1 \right) \\ &= z^{\tilde{n}_x} P_c(1/z) \\ &= z^{\tilde{n}_x} (1 - x_1/z)(1 - x_2/z) \cdots (1 - x_{D_{\min}}/z) \\ &= z^{\tilde{n}_x - D_{\min}} (z - x_1)(z - x_2) \cdots (z - x_{D_{\min}}), \end{aligned}$$

which extends to $z = 0$ by continuity. This concludes the proof of points (i) and (ii).

(iii) Let (p, q) be a polynomial solution to (1.5) and $\boldsymbol{c}$ the corresponding coefficient solution. From Lemma 4.3 we have

$$q(z) = p(z)e^{m(z)},$$

where $m(z)$ is defined in (4.3). For the $(K+1)$th Taylor coefficient of the left- and right-hand sides we have by Lemmas 4.3 and 4.2

$$0 = \sum_{j=0}^{n_x} c_j a_{K+1-j} \quad \Rightarrow \quad a_{K+1} = -\sum_{j=1}^{n_x} a_{K+1-j} c_j, \tag{5.2}$$

since the kth Taylor coefficient of q and p is zero for $k > n_x$ and $k > n_y$, respectively. Finally, the last row of (2.1) extended to size $K+1$ gives

$$m_{K+1} = (K+1)a_{K+1} - \sum_{j=1}^{K} m_j a_{K+1-j}.$$

Together the last two equations show point (iii).

6. Properties of A_1 and Markov's theorem. We now look in more detail at the structure of the A_1 matrix. In particular, we look at the implications of $A_1 R$ being positive definite. Then we get an explicit simplified formula for the matrix, and our results also shed some light on the relationship of our results to the classical Markov theorem on the existence and uniqueness of solutions to the finite moment problem (1.1) discussed in the introduction. For this we need to define the matrix

$$R = \begin{pmatrix} & & 1 \\ & \cdots & \\ 1 & & \end{pmatrix}$$

and note that left (right) multiplication by R reverses the order of rows (columns) of a matrix. In our notation we can then formulate Markov's theorem as follows.

THEOREM 6.1 (Markov). *Suppose $K = 2n$ is even and $n = n_x = n_y$. There is a unique piecewise continuous function $f(x)$ satisfying*

$$m_k = k \int_{\mathbb{R}} x^{k-1} f(x) dx, \qquad 0 \le f \le 1, \qquad k = 1, \dots, K, \tag{6.1}$$

if $A_1 R$ is symmetric positive definite and the matrix

$$\begin{pmatrix} \boldsymbol{a}_0 & A_1 \\ a_{K+1} & \boldsymbol{a}_0^T \end{pmatrix} \tag{6.2}$$

is singular. This f is of the form in (1.2), (1.3).

REMARK 7. *The theorem does not rule out other forms of $f(x)$ a priori, and without the second condition in* (6.2) *such solutions are indeed possible. It considers only the case $n_x = n_y$, i.e., problem* (1.4), *and says nothing about the possibility of other solution types, e.g., when the $\{x_j\}$ and $\{y_j\}$ are not interlaced as in* (1.3).

We start by introducing some new notation that will be used throughout this section. If $\{x_j\}$ and $\{y_j\}$ are a solution of (1.5) and (p, q) is the corresponding polynomial solution as defined in (3.1), (3.2), we can introduce the new polynomials $p_r(z) = z^{n_x} p(1/z)$ and $q_r(z) = z^{n_y} q(1/z)$ to describe the solution. Defining them by continuity at $z = 0$, we have

$$p_r(z) = (z - x_1) \cdots (z - x_{n_x}), \qquad q_r(z) = (z - y_1) \cdots (z - y_{n_y}). \tag{6.3}$$

Furthermore, we assume that the number of *distinct* roots of p_r (x_j-branch values) is $\tilde{n}$. We also order the roots such that we can write

$$p_r(z) = (z - x_1)^{1+\eta_1} (z - x_2)^{1+\eta_2} \cdots (z - x_{\tilde{n}})^{1+\eta_{\tilde{n}}},$$

where $1 + \eta_j$ is the multiplicity of the root x_j, so that

$$n_x = \mathrm{Deg}(p_r) = \tilde{n} + \sum_{\ell=1}^{\tilde{n}} \eta_\ell.$$

We start the analysis with a lemma giving explicit expressions for the a_k-values.

LEMMA 6.2. *For $k \geq 0$,*

$$a_{n_y - n_x + 1 + k} = \sum_{j=1}^{\tilde{n}} \frac{1}{\eta!} \lim_{z \to x_j} \frac{d^{\eta_j}}{dz^{\eta_j}} \frac{(z - x_j)^{1+\eta_j} z^k q_r(z)}{p_r(z)}. \tag{6.4}$$

Proof. This result follows from an application of the residue theorem in complex analysis as follows. Let C_r be the circle in the complex plane with radius r. Since the roots of $p(z)$ are nonzero, the function q/p is analytic within and on C_ε if ε is taken small enough, and the Cauchy integral formula gives

$$a_k = \begin{cases} \frac{1}{k!} \frac{d^k}{dz^k} \frac{q(z)}{p(z)} \Big|_{z=0}, & k \geq 0, \\ 0, & k < 0, \end{cases} = \frac{1}{2\pi i} \oint_{C_\varepsilon} \frac{q(z)}{p(z) z^{k+1}} dz.$$

Setting

$$f(z) := \frac{q_r(z)}{p_r(z)} = \frac{z^{n_y - n_x} q(1/z)}{p(1/z)} \tag{6.5}$$

and changing variable $z \to 1/z$, we get

$$a_{n_y - n_x + 1 + k} = \frac{1}{2\pi i} \oint_{C_\varepsilon} \frac{q(z)}{p(z) z^{n_y - n_x + k + 2}} dz = \frac{1}{2\pi i} \oint_{C_\varepsilon} \frac{f(1/z)}{z^{k+2}} dz = \frac{1}{2\pi i} \oint_{C_{1/\varepsilon}} z^k f(z) dz.$$

Hence, $a_{n_y - n_x + 1 + k}$ is given by the sum of the residues of $z^k f(z)$ (assuming we take small enough ε). By (6.5) and the restriction $k \geq 0$ we see that its poles are located at the x_j-values and they have multiplicities $1 + \eta_j$ at x_j. Then (6.4) follows from the residue formula for a pole of a function $g(z)$ at z^* with multiplicity $\eta + 1$,

$$\mathrm{Res}(g, z^*) = \frac{1}{\eta!} \lim_{z \to z^*} \frac{d^\eta}{dz^\eta} (z - z^*)^{1+\eta} g(z). \qquad \square$$

When the branch values $\{x_j\}$ are *distinct*, the expression for the a_k elements simplifies. They can then be expressed as sums of the powers of $\{x_j\}$ in a way similar to the moments m_k, but with weights different from one. We can also give a more concise description of the matrices A_0 and A_1, which can be factorized into a product of Vandermonde and diagonal matrices. More precisely, we let V be the Vandermonde matrix

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{n_x} \\ x_1^2 & x_2^2 & \cdots & x_{n_x}^2 \\ \vdots & \cdots & \cdots & \vdots \\ x_1^{n_x-1} & x_2^{n_x-1} & \cdots & x_{n_x}^{n_x-1} \end{pmatrix}$$

and introduce the diagonal matrices

$$W = \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_{n_x} \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_{n_x} \end{pmatrix},$$

where w_j are the weights defined as

$$w_j = \frac{q_r(x_j)}{p_r'(x_j)}. \tag{6.6}$$

(Note that p_r has only simple roots when $\{x_j\}$ are distinct, so $p_r'(x_j) \neq 0$.) Then we can show the following.

PROPOSITION 6.3. *If $\{x_j\}$ are distinct, then for $k \geq 0$,*

$$a_{n_y-n_x+1+k} = \sum_{j=1}^{n_x} w_j x_j^k \tag{6.7}$$

and

$$A_1 R = VWV^T, \qquad A_0 R = VWXV^T. \tag{6.8}$$

Proof. When $\{x_j\}$ are distinct, $\eta_j = 0$ for all j and the expression (6.4) for the x_j-residue simplifies to

$$\lim_{z \to x_j} \frac{(z - x_j) z^k q_r(z)}{p_r(z)} = \frac{x_j^k q_r(x_j)}{p_r'(x_j)}.$$

This shows (6.7). For (6.8) we set $b_k = a_{n_y-n_x+1+k}$. Then

$$A_{1-r} R = \begin{pmatrix} b_r & b_{r+1} & \cdots & b_{r+n_x} \\ b_{r+1} & b_{r+2} & \cdots & b_{r+n_x+1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{r+n_x} & b_{r+n_x+1} & \cdots & b_{r+2n_x} \end{pmatrix} \in \mathbb{R}^{n_x \times n_x}, \qquad r = 0, 1.$$

From (6.7) we then have, for $k \geq 0$,

$$\begin{pmatrix} b_k \\ b_{k+1} \\ \vdots \\ b_{k+n_x} \end{pmatrix} = \sum_{j=1}^{n_x} w_j \begin{pmatrix} x_j^k \\ x_j^{k+1} \\ \vdots \\ x_j^{k+n_x} \end{pmatrix} = \sum_{j=1}^{n_x} w_j x_j^k \begin{pmatrix} 1 \\ x_j \\ \vdots \\ x_j^{n_x} \end{pmatrix} = V \begin{pmatrix} w_1 x_1^k \\ w_2 x_2^k \\ \vdots \\ w_{n_x} x_{n_x}^k \end{pmatrix} = VW \begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_{n_x}^k \end{pmatrix}.$$

Consequently,

$$A_{1-r}R = VW \begin{pmatrix} x_1^r & x_1^{r+1} & \dots & x_1^{r+n_x} \\ x_2^r & x_2^{r+1} & \dots & x_2^{r+n_x} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_x}^r & x_{n_x}^{r+1} & \dots & x_{n_x}^{r+n_x} \end{pmatrix} = VWX^rV^T,$$

which concludes the proof. □

We now consider the implications of a positive definite A_1R. It turns out that this is a necessary and sufficient condition to guarantee both distinct $\{x_j\}$-values and positive weights. We get the following theorem.

THEOREM 6.4. *The matrix A_1R is symmetric positive definite if and only if $\{x_j\}$ are distinct and the weights are strictly positive, $w_j > 0$ for $j = 1, \dots, n_x$.*

Proof. We use the same notation as in Lemma 6.2 and set

$$S_j(z) = \frac{1}{\eta_j!}(z - x_j)^{1+\eta_j}\frac{q_r(z)}{p_r(z)}.$$

We note that $S_j(z)$ is smooth and regular close to $z = x_j$. Then, by Lemma 6.2, for $k \geq 0$,

$$a_{n_y-n_x+1+k} = \sum_{j=1}^{\tilde{n}} \lim_{z\to x_j} \frac{d^{\eta_j}}{dz^{\eta_j}} z^k S_j(z).$$

Next, we let $\boldsymbol{v} = (v_1, \dots, v_{n_x})^T$ be an arbitrary vector in $\mathbb{R}^{n_x}$ and recall that $P_v(z)$ is the corresponding $n_x - 1$ degree polynomial

$$P_v(z) = v_1 + v_2 z + \cdots + v_{n_x} z^{n_x-1}.$$

Then

$$\begin{aligned} \boldsymbol{v}^T A_1 R \boldsymbol{v} &= \sum_{j=1}^{n_x}\sum_{k=1}^{n_x} v_j v_k a_{n_y-n_x+j+k-1} = \sum_{j=1}^{n_x}\sum_{k=1}^{n_x}\sum_{\ell=1}^{\tilde{n}} \lim_{z\to x_\ell} \frac{d^{\eta_\ell}}{dz^{\eta_\ell}} z^{j+k-2} S_\ell(z) v_j v_k \\ &= \sum_{\ell=1}^{\tilde{n}} \lim_{z\to x_\ell} \frac{d^{\eta_\ell}}{dz^{\eta_\ell}} S_\ell(z) \sum_{j=1}^{n_x}\sum_{k=1}^{n_x} z^{j+k-2} v_j v_k = \sum_{\ell=1}^{\tilde{n}} \lim_{z\to x_\ell} \frac{d^{\eta_\ell}}{dz^{\eta_\ell}} S_\ell(z) P_v(z)^2. \end{aligned} \tag{6.9}$$

If

$$\tilde{n} + \sum_{j=1}^{\tilde{n}} \lfloor \eta_j/2 \rfloor \leq n_x - 1, \tag{6.10}$$

then we can take

$$P_v(z) = (z - x_1)^{1+\tilde{\eta}_1}(z - x_2)^{1+\tilde{\eta}_2} \cdots (z - x_{\tilde{n}})^{1+\tilde{\eta}_{\tilde{n}}}, \qquad \tilde{\eta}_j = \lfloor \eta_j/2 \rfloor.$$

Since $2(1 + \tilde{\eta}_\ell) = 2 + 2\lfloor \eta_\ell/2 \rfloor \geq 2 + 2(\eta_\ell/2 - 1) > \eta_\ell$ and

$$\left.\left(\frac{d^\ell}{dz^\ell} f(z)(z - z^*)^k\right)\right|_{z=z^*} = 0, \qquad 0 \leq \ell < k,$$

for all smooth enough $f(z)$, we get $\boldsymbol{v}^T A_1 R \boldsymbol{v} = 0$, which contradicts the positivity of $A_1 R$. Hence,

$$\tilde{n} + \sum_{j=1}^{\tilde{n}} \lfloor \eta_j / 2 \rfloor > n_x - 1 = \tilde{n} + \sum_{\ell=1}^{\tilde{n}} \eta_\ell - 1.$$

Since for any integer $n > 0$ we have $\lfloor n/2 \rfloor \leq n - 1$, it follows that all $\eta_\ell = 0$ and $\tilde{n} = n_x$. Hence, if $A_1 R$ is positive definite, then $\{x_j\}$ are distinct.

To show the theorem it is now enough to show that, when $\{x_j\}$ are distinct, $A_1 R$ is positive if and only if the weights are positive. From (6.9) we then have

$$\boldsymbol{v}^T A_1 R \boldsymbol{v} = \sum_{\ell=1}^{n_x} S_\ell(x_\ell) P_v(x_\ell)^2 = \sum_{\ell=1}^{n_x} w_\ell P_v(x_\ell)^2.$$

Clearly, when all $w_\ell > 0$, this expression is positive for $\boldsymbol{v} \neq 0$, and $A_1 R$ is positive definite. To show the converse, we take $P_v(z)$ to be the Lagrange basis polynomials $L_j(z)$ of degree $n_x - 1$ defined as

$$L_j(x_i) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

If $A_1 R$ is positive, then

$$0 < \boldsymbol{v}^T A_1 R \boldsymbol{v} = \sum_{\ell=1}^{n_x} w_\ell L_j(x_\ell)^2 = w_j.$$

This can be done for each j, which concludes the proof. □

We can now relate our conclusions with those in Markov's theorem, Theorem 6.1. We consider all solutions to (1.5) instead of those given by the integral relation (6.1) with a piecewise continuous function $f(x)$. The extra condition (6.2) is then automatically satisfied, and we note that the positivity of $A_1 R$ guarantees a unique solution also in our space of density functions (1.6). We view this as a corollary of Theorems 4.1 and 6.4.

Corollary 6.5. *If there exists a solution to* (1.5)*, then the matrix in* (6.2) *is singular. When $n_x = n_y$, there is a unique solution to* (1.5) *of the form* (1.3) *if and only if $A_1 R$ is symmetric positive definite.*

Proof. We start by proving the singularity of (6.2). By (ii) in Proposition 4.6 a coefficient solution $\boldsymbol{c} = (c_0, \ldots, c_{n_x})^T = (c_0, \bar{\boldsymbol{c}}^T)^T$ satisfies $A\boldsymbol{c} = 0$. Since $A = (\boldsymbol{a}_0 \; A_1)$, it remains to prove that $c_0 a_{K+1} + \boldsymbol{a}_0^T \bar{\boldsymbol{c}} = 0$. This was already proved in (5.2).

Next, we prove the "if" part of the second statement. If $A_1 R$ is symmetric positive definite, it is nonsingular, and by (i), (iii), and (v) in Theorem 4.1, the minimal degree solution exists and is unique and $x_j \neq y_i$ for all i, j. (If $x_j = 0$ for some j, then there is no zero y_i-value since $\mathrm{Deg}(q^*) = n$ by point (iii).) By Theorem 6.4 the corresponding branch values $\{x_j\}$ are distinct. It remains to show that, upon some reordering, the $\{x_j\}$ and $\{y_j\}$ are interlaced as in (1.3).

Order the x_j-values in an increasing sequence and let m_k be the number of y_j-values such that $y_j < x_k$. Clearly, m_k is increasing and $0 \leq m_k \leq n_y$. Moreover, $\mathrm{sgn}(q_r(x_k)) = (-1)^{n_y - m_k}$, and since $\lim_{z\to\infty} p_r'(z) > 0$, we also have $\mathrm{sgn}(p_r'(x_k)) = (-1)^{n_x - k}$. Hence, by also using the fact that $n_y = n_x$,

$$\operatorname{sgn}(w_k) = (-1)^{n_y - m_k + n_x - k} = (-1)^{m_k + k}.$$

We conclude that $m_k + k$ is even, which implies that m_k is, in fact, *strictly* increasing. Then, for $k = 1, \dots, n_x - 1$, we have $m_{k+1} \geq m_k + 1$ and

$$n_x \geq m_{n_x} \geq m_k + n_x - k \quad \Rightarrow \quad m_k \leq k.$$

Similarly, $m_k \geq m_1 + k - 1 \geq k - 1$, so $k - 1 \leq m_k \leq k$, and therefore

$$2k - 1 \leq m_k + k \leq 2k.$$

Finally, since $m_k + k$ is even, we must have $m_k = k$, which implies that the values are interlaced.

We now consider the "only if" part. If there is a solution of the form (1.3), then the $\{x_j\}$-values are obviously distinct and $m_k = k$. By Proposition 6.3 the weights are then given by (6.6) and they are positive since, as above, $\operatorname{sgn}(w_k) = (-1)^{m_k+k} = 1$. It follows from Theorem 6.4 that $A_1 R$ is positive definite. □

7. Outlook.

Several interesting issues may be worth mentioning:

1. Computational complexity in a finite difference implementation: One can consult the article [14], where practical implementation issues and several examples of increasing complexity have been addressed in the context of geometric optics problems. In particular, comparisons with Lagrangian (ray-tracing) solutions are shown.
2. Extension to higher dimensions for the present problem: Nothing seems to exist in this direction at the time being; see, however, the last sections of [20] and the routines based on complex variables in [11, 9] for "shape from moments."
3. A very special case of the trigonometric moment problem can be solved by means of a slight variation of the algorithms presented here, in [14], and in section IV.A of [9]. That is to say, one tries to invert the following set of equations:

$$\sum_{j=0}^{n} \mu_j \exp(ik\lambda_j) = m_k, \qquad k = 0, \dots, n. \tag{7.1}$$

Let us state that in case the $n+1$ real frequencies λ_j are known, the set of complex amplitudes μ_j are found by solving a Vandermonde system:

$$\begin{pmatrix} 1 & \cdots & 1 \\ \exp(i\lambda_0) & \cdots & \exp(i\lambda_n) \\ \vdots & & \vdots \\ \exp(in\lambda_0) & \cdots & \exp(in\lambda_n) \end{pmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_n \end{pmatrix}.$$

The frequencies can be found through a byproduct of [9, 14] as we state now. Let us suppose n is odd (i.e., the number of equations is even); we form the two matrices

$$A_1 = \begin{pmatrix} m_0 & \cdots & m_{\frac{n-1}{2}} \\ \vdots & & \vdots \\ m_{\frac{n-1}{2}} & \cdots & m_{n-1} \end{pmatrix}, \qquad A_2 = \begin{pmatrix} m_1 & \cdots & m_{\frac{n+1}{2}} \\ \vdots & & \vdots \\ m_{\frac{n+1}{2}} & \cdots & m_n \end{pmatrix},$$

and then the frequencies can be obtained through a generalized eigenvalue problem, $A_1 \boldsymbol{v}_j = \lambda_j A_2 \boldsymbol{v}_j$, $j = 0, \dots, n$. This kind of algorithm can be used to check the accuracy of the classical FFT and will be studied in a forthcoming article.

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Oliver and Boyd, Edinburgh, 1965.

[2] N. I. Akhiezer and M. G. Krein, *Some Questions in the Theory of Moments*, Transl. Math. Monogr. 2, AMS, Providence, RI, 1962.

[3] Y. Brenier, *Équations de moment et conditions d'entropie pour des modèles cinétiques* (French) [Moment equations and entropy conditions for kinetic models], in Séminaire sur les Équations aux Dérivées Partielles, 1994–1995, École Polytech., Palaiseau, France, 1995, Exp. No. XXII.

[4] Y. Brenier and L. Corrias, *A kinetic formulation for multibranch entropy solutions of scalar conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 169–190.

[5] R. E. Curto and L. A. Fialkow, *The truncated complex K-moment problem*, Trans. Amer. Math. Soc., 352 (2000), pp. 2825–2855.

[6] D. Czarkowski, D. V. Chudnovsky, G. V. Chudnovsky, and I. W. Selesnick, *Solving the optimal PWM problem for single-phase inverters*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 465–475.

[7] P. Diaconis, *Application of the method of moments in probability and statistics*, in Moments in Mathematics (San Antonio, TX, 1987), Proc. Sympos. Appl. Math. 37, AMS, Providence RI, pp. 125–142.

[8] P. Diaconis and D. Friedman, *The Markov moment problem and de Finetti's theorem*, Math. Z., 247 (2004), pp. 183–199, 201–212.

[9] M. Elad, P. Milanfar, and G. H. Golub, *Shape from moments—an estimation theory perspective*, IEEE Trans. Signal Process., 52 (2004), pp. 1814–1829.

[10] F. Gamboa and L. V. Lozada-Chang, *Large-deviations for random power moment problems*, Ann. Probab., 32 (2004), pp. 2819–2837.

[11] G. H. Golub, P. Milanfar, and J. Varah, *A stable numerical method for inverting shape from moments*, SIAM J. Sci. Comput., 21 (1999), pp. 1222–1243.

[12] L. Gosse, *Using K-branch entropy solutions for multivalued geometric optics computations*, J. Comput. Phys., 180 (2002), pp. 155–182.

[13] L. Gosse, S. Jin, and X. Li, *Two moment systems for computing multiphase semiclassical limits of the Schrödinger equation*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1689–1723.

[14] L. Gosse and O. Runborg, *Finite moment problems and applications to multiphase computations in geometric optics*, Commun. Math. Sci., 3 (2005), pp. 373–392.

[15] L. Gosse and O. Runborg, *Resolution of the finite Markov moment problem*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 775–780.

[16] V. I. Korobov and G. M. Sklyar, *Time-optimality and the power moment problem*, Mat. Sb. (N.S.), 134 (176) (1987), pp. 186–206, 287 (in Russian); translation in Math. USSR-Sb., 62 (1989), pp. 185–206.

[17] M. G. Krein and A. A. Nudel'man, *The Markov Moment Problem and Extremal Problems*, Transl. Math. Monogr. 50, AMS, Providence, RI, 1977.

[18] A. S. Lewis, *Superresolution in the Markov moment problem*, J. Math. Anal. Appl., 197 (1996), pp. 774–780.

[19] D. T. Norris, *Optimal Solutions to the L^∞ Moment Problem with Lattice Bounds*, Ph.D. thesis, University of Colorado, Boulder, CO, 2002; available online from http://math.colorado.edu/~norrisdt/dougthesis.ps.

[20] M. Putinar, *A renormalized Riesz potential and applications*, in Advances in Constructive Approximation, M. Neamty and E. Saff, eds., Nashboro Press, Brentwood, TN, 2004, pp. 433–465.

[21] O. Runborg, *Some new results in multiphase geometrical optics*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1203–1231.

[22] M. I. Sezan and H. Stark, *Incorporation of a-priori moment information into signal recovery and synthesis problems*, J. Math. Anal. Appl., 122 (1987), pp. 172–186.

[23] B. Simon, *The classical moment problem as a self-adjoint finite difference operator*, Adv. Math., 137 (1998), pp. 82–203.

[24] G. M. Sklyar and L. V. Fardigola, *The Markov power moment problem in problems of controllability and frequency extinguishing for the wave equation on a half-axis*, J. Math. Anal. Appl., 276 (2002), pp. 109–134.

[25] G. Talenti, *Recovering a function from a finite number of moments*, Inverse Problems, 3 (1987), pp. 501–517.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1641–1664

INCIPIENT DYNAMICS OF SWELLING OF GELS*

HANG ZHANG[†] AND M. CARME CALDERER[†]

Abstract. In this article, we analyze a model of the incipient dynamics of gel swelling and perform numerical simulations. The governing system consists of balance laws for a mixture of nonlinear elastic solid and solvent yielding effective equations for the gel. We discuss the multiscale nature of the problem and identify physically realistic regimes. The mixing mechanism is based on the Flory–Huggins energy. We consider the case that the dissipation mechanism is the solid-solvent friction force. This leads to a system of weakly dissipative nonlinear hyperbolic equations. After addressing the Cauchy problem, we propose physically realistic boundary conditions describing the motion of the swelling boundary. We study the linearized version of the free boundary problem. Numerical simulations of solutions are presented too.

Key words. gel swelling, two-component mixture, polymer-solvent friction, type II diffusion, hyperbolic free boundary problem

AMS subject classifications. 35L45, 74B20, 74F10, 74F20

DOI. 10.1137/070680941

1. Introduction.

We present analysis and numerical simulations of a model of the incipient dynamics of polymer gel swelling. The system that we study is derived from the balance laws of a two-component mixture of solid polymer and solvent [3]. The free energy of the system consists of the elastic energy of deformation of the polymer together with the Flory–Huggins free energy of mixing. The dissipation is due to the friction force between polymer and solvent. The resulting balance laws of the mixture form a weakly dissipative hyperbolic system. We formulate the boundary conditions for the swelling boundaries and analyze the free boundary problem for the linearized equations. We discuss numerical simulations of the solutions of the Cauchy problem. The effective equations that we obtain clearly reveal the multiscale nature of the problem and the dynamics associated with the different time scales.

In dimensionless form, the elasticity, Flory–Huggins mixing, and dissipative mechanisms bring four time scales into the problem, with the largest one naturally associated with the friction mechanism: this is the time scale of relaxation to the equilibrium volume fraction, with diffusive dynamics. Indeed, many works on gels focus on the relaxation part of the process. In this article, we address the earlier time scale dynamics and investigate their physical and mathematical significances, with the goal of understanding their individual roles. This allows us to gain information on the start-up of the process and on how the evolution of the swelling surface occurs.

In our applications, we will refer mostly to two classes of materials, entangled linear polymers and polysaccharides. In terms of the physical parameters that characterize them, the dissipation coefficient and elasticity modulus of the latter are several orders of magnitude smaller than their polymeric counterparts.

Our study is motivated by polymeric applications to body implanted devices, such as bone replacement tissue and controlled drug release mechanisms. Predictions

*Received by the editors January 25, 2007; accepted for publication (in revised form) February 8, 2008; published electronically June 25, 2008. This work was supported in part by Medtronic Inc. and by National Science Foundation grant DMS-0456232.

http://www.siam.org/journals/siap/68-6/68094.html

[†] School of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church Street S.E., Minneapolis, MN 55455 (hzhang@math.umn.edu, mcc@math.umn.edu).

on change in shape from the dry state to saturation may help the manufacturing process. Another related application involving polysaccharides is the study of gel motility phenomena in myxobacteria.

Still a larger time scale would be present if polymer-polymer dissipation were taken into account. Since our numerical simulations focus on the dynamics at the smaller time scales, we neglect the latter dissipation source in the present work.

Another motivation to our work is to gain some understanding of the so-called type II diffusion phenomena [21] and [20]. It has been experimentally observed that the dynamics of interaction between the gel and its surrounding solvent, in the case that the polymer is dry, is very different from the case of partially swollen polymer. Indeed, the latter shows features of standard diffusion. In this work, we argue that type II diffusion is mostly a hyperbolic phenomena and therefore significantly different from standard diffusion.

The model that we study was developed in [3]. It consists of laws of balance of mass, momentum, and energy for two-component systems of nonlinear elastic solid and solvent. Fields of the problem include volume fractions ϕ_1 and ϕ_2, velocity fields $\mathbf{v}_1$ and $\mathbf{v}_2$ for solid and solvent, respectively, and pressure λ, a Lagrange multiplier corresponding to the constraint of $\phi_1 + \phi_2 = 1$. We observe that the continuum theory for a two-component mixture can be used as a tool to obtain governing equations for a third material, the gel, with properties that may be significantly different from those of the individual components. Moreover, the effective equations are formulated in terms of the center of mass velocity $\mathbf{V}$ and the diffusion velocity $\mathbf{U} = \mathbf{v}_1 - \mathbf{v}_2$. The model as formulated allows us to identify regimes associated with the different time scales, with the short times characterizing evolution of the interface between gel and solvent. The friction between polymer and solvent suggests existence of a purely diffusive regime with $\mathbf{V} = 0$. Indeed, considering initial conditions satisfying $\mathbf{V} = 0$, there exists a Lagrange multiplier function that maintains zero center of mass velocity for as long as the solution exists. If polymer-polymer friction is included in the model, either in the form of Newtonian dissipation or as given by a viscoelastic law, another time scale, larger by several orders of magnitude than the diffusive one, is added to the problem. Our model, then, suggests that, upon relaxation to equilibrium of the diffusive velocity and volume fraction, the material subsequently evolves as a viscoelastic fluid with uniform volume fraction, with respect to the transport velocity $\mathbf{V}$.

In one-dimensional geometries, the model reduces to a system of equations for the diffusion velocity of the mixture U and the volume fraction of the polymer ϕ_1. The dependence on the center of mass velocity can be eliminated by imposing transitional invariance of the solutions.

We formulate the one-dimensional problem in Eulerian coordinates, in which case it becomes a free boundary problem. First, we assume that the interface between dry polymer and solvent achieves a balance of force all the time, and that the interface is fully saturated. In our framework, this amounts to neglecting the shortest time scale of the system that causes very rapid saturation of the interface to an equilibrium volume fraction $\phi^* \in (0, 1)$. The value of ϕ^* can itself be determined by the pressure applied to the surrounding solvent [7]. We also assume that the interface moves at the speed of the polymer and formulate an ordinary differential equation for its dynamics. The remaining boundary conditions are formulated in terms of the symmetry of the domain with respect to $x = 0$, implying that $U(0, t) = 0$. We study the free boundary problem for the linearized equations and prove the global existence of solutions $(\phi(x, t), U(x, t), S(t)) \in C^1(\bar{Q}_S \times C^1(\bar{Q}_S \times C^2[0, t_1]))$, where $Q(S) := \{(x, t) : t \geq 0,\ 0 < x < S(t)\}$; we also prove a global bound for $S(t)$, $0 < |S(t) - L| < C$.

The analysis of the free boundary problem precludes the presence of shocks in the system. In a separate section, we consider the Cauchy problem and show that the system is weakly dissipative, as characterized by Dafermos [5]. The existence of an entropy-entropy pair flux allows us to show that the governing system is L^1-stable and, therefore, solutions of bounded variation follow as a consequence of the theorem in [5]. We show that the condition of hyperbolicity is satisfied for the material constants of a linear polymer as shown in Table 1. Consequently, the system remains hyperbolic and weakly dissipative for all time. However, this is not the case for polysaccharide data, where hyperbolicity is lost at a critical volume fraction ϕ_c that may be greater than the saturation value ϕ^*. Thus, the swelling interface may stop propagating before reaching saturation. We interpret such a phenomenon as the onset of deswelling. In the case of myxobacteria, this may suggest a reversal of direction perhaps achieved by deswelling. In this model, the break of hyperbolicity shown by the polysaccharide data occurs because of the small elastic modulus. This provides another motivation to study early dynamics. Indeed, in polysaccharide systems, the regime of relaxation dynamics may not be reached.

Works by Doi and coauthors address steady state solutions as well as relaxation regimes [23], [24], [25], [7], and [27]. Our modeling assumptions involving the free energy, which combines the Flory–Huggins contribution and the rubber elasticity, and the multiscale properties of the system are fully motivated by such works and those by Tanaka and Filmore [19].

This work is also partially inspired by the analysis in [15] of a flow with viscoelastic particles. From another point of view, the system of equations and free boundary problem share mathematical analogies with models of diffusion and transport aiming at including finite speed propagation effects in heat conduction [18].

In section 2 we explain the model, and in section 3, we derive the properties of the one-dimensional system. The Cauchy problem is studied in section 4, and the boundary conditions and the free boundary problem are formulated and studied in section 5. Finally, in section 6, we present numerical simulations for the regularized system.

2. The model.

We use the continuum theory of mixtures of an elastic solid and a solvent as the main tool to derive the governing equations of a gel [22, Chapter 5]. Since the free energy depends explicitly on the volume fraction of the components, the mixture modeling the gel turns out to be of immiscible type. Furthermore, since the intrinsic densities of the components are taken to be constant, the mixture is incompressible.

2.1. Balance of mass, transport and constitutive equations.

We assume that each component occupies a domain $\Omega_a \subset \mathbb{R}^3$, $a = 1, 2$, in the reference configuration (Lagrangian), with a reference volume fraction ϕ_a^R. Here the subindex 1 refers to the polymer, and 2 represents the fluid. In some applications, the reference configuration can be taken to be the initial state of the mixture. It is important to emphasize that the reference domains Ω_a, $a = 1, 2$, will, in general, be distinct. Both components occupy a common domain in the deformed (Eulerian) configuration.

The deformation of each component, polymer and fluid, respectively, is given by sufficiently smooth functions

$$\mathbf{x} = \mathcal{M}(\mathbf{X}, t), \quad \mathbf{X} \in \Omega_1,$$
$$\mathbf{x} = \mathcal{N}(\mathbf{X}, t), \quad \mathbf{X} \in \Omega_2,$$

with $F = \nabla_{\mathbf{X}}\mathcal{M}(\mathbf{X},t)$ denoting the gradient of deformation of the polymer. According to the theory of mixtures, both polymer and fluid may occupy the same region, with volume fractions $\phi_1(\mathbf{x},t)$, $\phi_2(\mathbf{x},t)$, respectively. Here $\mathbf{x} \in \Omega$ represents a point in a fixed region in space. We also assume that no other material or vacuum is present in the region; that is,

$$\phi_1(\mathbf{x},t) + \phi_2(\mathbf{x},t) = 1 \tag{2.1}$$

holds. We let ρ_1 and ρ_2 denote the mass densities of each component, respectively, per unit volume in space. These are related to the *true densities* $(\frac{\text{mass of component}}{\text{volume of component}})$ γ_1 and γ_2 as follows:

$$\rho_1 = \gamma_1\phi_1, \quad \rho_2 = \gamma_2\phi_2.$$

We assume that the mass densities of polymer and fluid are equal, and $\gamma_1 = \gamma_2 = 1$. In this case, the densities and volume fractions coincide:

$$\rho_1 = \phi_1, \quad \rho_2 = \phi_2. \tag{2.2}$$

We introduce the material velocities of polymer and solvent, respectively:

$$\begin{aligned} \tilde{\mathbf{v}}_1(\mathbf{X},t) &= \frac{\partial\mathcal{M}}{\partial t}(\mathbf{X},t), \quad \mathbf{X} \in \Omega_1, \\ \tilde{\mathbf{v}}_2(\mathbf{X},t) &= \frac{\partial\mathcal{N}}{\partial t}(\mathbf{X},t), \quad \mathbf{X} \in \Omega_2. \end{aligned}$$

We denote the corresponding velocity fields

$$\mathbf{v}_1(\mathbf{x},t) = \tilde{\mathbf{v}}_1(\mathcal{M}^{-1}(\mathbf{x},t),t), \quad \mathbf{v}_2(\mathbf{x},t) = \tilde{\mathbf{v}}_2(\mathcal{N}^{-1}(\mathbf{x},t),t), \quad \mathbf{x} \in \Omega. \tag{2.3}$$

We let $\mathcal{T}_1(\mathbf{x},t)$ and $\mathcal{T}_2(\mathbf{x},t)$ denote Cauchy stress tensor of polymer and fluid, respectively. Each one may consist of elastic and dissipative contributions, although in this work we emphasize the former. In addition, we take into account the friction forces $\mathbf{f}_a$, per unit volume, that the polymer exerts upon the fluid, and vice versa. The local forms of the laws of balance of mass and linear momentum are

$$\frac{\partial\phi_1}{\partial t} + (\mathbf{v}_1\cdot\nabla)\phi_1 + \phi_1\nabla\cdot\mathbf{v}_1 = 0, \tag{2.4}$$

$$\frac{\partial\phi_2}{\partial t} + (\mathbf{v}_2\cdot\nabla)\phi_2 + \phi_2\nabla\cdot\mathbf{v}_2 = 0, \tag{2.5}$$

$$\phi_1\frac{\partial\mathbf{v}_1}{\partial t} + \phi_1(\mathbf{v}_1\cdot\nabla)\mathbf{v}_1 = \nabla_{\mathbf{x}}\cdot\mathcal{T}_1 + \mathbf{f}_1, \tag{2.6}$$

$$\phi_2\frac{\partial\mathbf{v}_2}{\partial t} + \phi_2(\mathbf{v}\cdot\nabla)\mathbf{v}_2 = \nabla_{\mathbf{x}}\cdot\mathcal{T}_2 + \mathbf{f}_2. \tag{2.7}$$

Assuming that the second law of thermodynamics holds for all admissible processes [3], we have the following equations for the reversible parts of the stress tensors, $\mathcal{T}_1$ and $\mathcal{T}_2$, and an expression for the friction forces $\mathbf{f}_a$:

$$\mathcal{T}_1 = \phi_1\left\{\frac{\partial\psi_1}{\partial F}F - \left(\phi_1\frac{\partial\psi_1}{\partial\phi_1} + \phi_2\frac{\partial\psi_2}{\partial\phi_1} + \lambda\right)I\right\}, \tag{2.8}$$

$$\mathcal{T}_2 = -\phi_2\left\{\phi_1\frac{\partial\psi_1}{\partial\phi_2} + \phi_2\frac{\partial\psi_2}{\partial\phi_2} + \lambda\right\}I, \tag{2.9}$$

$$\mathbf{f}_1 = \lambda\nabla\phi_1 - \beta(\mathbf{v}_1 - \mathbf{v}_2) = -\mathbf{f}_2, \tag{2.10}$$

where λ is the Lagrange multiplier associated with the constraint (2.1) and $\beta(\phi_1, \phi_2)$ the polymer drag coefficient. The functions ψ_1 and ψ_2 represent the free energies of the polymer and solvent, respectively, giving the total free energy of the mixture, $\Psi \equiv \phi_1\psi_1 + \psi_2\psi_2$. According to the Flory theory of mixtures [8],

$$\Psi = \frac{K_BT}{V_m}\left(\frac{\chi}{2}\phi_1\phi_2 + \frac{1}{N}\phi_1 \log\phi_1 + \phi_2 \log\phi_2\right) + \phi_1 W(F). \tag{2.11}$$

Expressions of the component free energies [9] that yield (2.11) are

$$\psi_1 = \frac{K_BT}{2V_m}\chi\phi_2^2 + \frac{K_BT}{N_1V_m}\log\phi_1 + W(F), \tag{2.12}$$

$$\psi_2 = \frac{K_BT}{2V_m}\chi\phi_1^2 + \frac{K_BT}{N_2V_m}\log\phi_2, \tag{2.13}$$

where $W(F)$ represents the elastic deformation energy which we will assume to be neo-Hookean [1]; that is, $W(F) = \mu \text{trace} FF^T$, with $\mu > 0$.

With these, (2.8) and (2.9) become

$$\mathcal{T}_1 = \frac{K_BT}{N_xV_m}\phi_1\left((\det F)^{\frac{2}{3}} - \left(\frac{1}{2} + \frac{N_x}{N_1}\right) - \chi N_x\phi_1\phi_2\right)\mathbf{I} - \lambda\phi_1\mathbf{I} + 2\mu\phi_1 FF^T, \tag{2.14}$$

$$\mathcal{T}_2 = -\phi_2\left(\frac{K_BT}{N_2V_m} + \frac{K_BT}{V_m}\chi\phi_1\phi_2 + \lambda\right)\mathbf{I}. \tag{2.15}$$

We now list the parameters of the problem:

1. V_m is the volume occupied by one monomer;
2. K_B is the Boltzmann constant, and T is the absolute temperature;
3. N_1, N_2 denote the number of lattice sites occupied by the polymer and the solvent;
4. N_x is the number of monomers between entanglement points;
5. $\frac{\phi_1}{N_xV_m}$ represents the number of entanglement points per unit volume;
6. χ is the Flory interaction parameter;
7. β is the polymer drag coefficient;
8. μ is related to the elastic shear modulus.

Parameter values appropriate to semidry polymers are given in Table 1 [27], [16].

TABLE 1
Data for polymer and polysaccharide.

Parameter	Polymer	Polysaccharide
N_x	20	20
N_1	1000	1000
N_2	1	1
V_m	.1 nm^3	.1 nm^3
χ	.5	.5
T	300° K	300° K
μ	10^4 pNnm^{-2}	10^{-5} pNnm^{-2}
β	2.4×10^{10} pN$s$nm^{-4}	2.4×10^3 pN$s$nm^{-4}

Remark. The approach developed so far is also suitable to account for polymer-polymer friction by postulating a viscoelastic law for the total stress. Let us denote $\tau_1 = \mathcal{T}_1 + \phi_1\lambda$ and $\tau_2 = \mathcal{T}_2 + \phi_2\lambda$, the (reversible) extra stress of polymer and solvent, respectively. Let $\tau_1^{total} = \tau_1 + \tau_1^d$. In the case of Jeffrey's model [2], we have

$$\tau_1^{total} + \xi[\dot{\tau}_1^{total} - (\nabla\mathbf{v}_1)^T\tau_1^{total} - \tau_1^{total}(\nabla\mathbf{v}_1)] = \eta_0\mathbf{D}^1 + \tau_1, \tag{2.16}$$

where $\mathbf{D}_1$ is the strain, $\xi > 0$ denotes a relaxation constant, and $\eta_0 > 0$ is the Newtonian viscosity.

2.2. Governing equations of gels. The governing equations for the individual components give the governing system of the gel. The fields of the gel model consist of

$$\{\mathbf{V}, \mathbf{U}, F, \phi_1, \lambda\},$$

where $\mathbf{V} = \phi_1\mathbf{v}_1 + (1-\phi_1)\mathbf{v}_2$ represents the center of mass velocity and $\mathbf{U} = \mathbf{v}_1 - \mathbf{v}_2$ the diffusion velocity. The total stress $\mathcal{T}$ is defined by

$$\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2 - (1-\phi_1)\phi_1\mathbf{U}\otimes\mathbf{U}.$$

From (2.4)–(2.10), we derive the governing system for the new variables,

$$\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{V} = \nabla\cdot\mathcal{T}, \tag{2.17}$$

$$\begin{aligned}&\frac{\partial \mathbf{U}}{\partial t} + (1-2\phi_1)(\nabla\mathbf{U})\mathbf{U} - (\mathbf{U}\otimes\mathbf{U})\nabla\phi_1 + (\nabla\mathbf{V})\mathbf{U} + (\nabla\mathbf{U})\mathbf{V}\\ &= \frac{1}{\phi_1}\nabla\cdot\mathcal{T}_1 - \frac{1}{1-\phi_1}\nabla\cdot\mathcal{T}_2 - \frac{\beta}{\phi_1(1-\phi_1)}\mathbf{U} + \frac{\lambda\nabla\phi_1}{\phi_1(1-\phi_1)},\end{aligned} \tag{2.18}$$

$$F_t + (\mathbf{V} + (1-\phi_1)\mathbf{U})\cdot\nabla F = \nabla(\mathbf{V} + (1-\phi_1)\mathbf{U})F, \tag{2.19}$$

$$\nabla\cdot\mathbf{V} = 0, \tag{2.20}$$

$$\frac{\partial\phi_1}{\partial t} + ((\mathbf{V} + (1-\phi_1)\mathbf{U})\cdot\nabla)\phi_1 + \phi_1\nabla\cdot(\mathbf{V} + (1-\phi_1)\mathbf{U}) = 0. \tag{2.21}$$

Equation (2.19) is a version of the chain rule relating time derivatives of F with velocity gradients. This equation is required in mixed solid-fluid systems [14]. We note that the first equation gives the balance of linear momentum for the mixture, and the second can be interpreted as giving the evolution of the microstructure of the gel. We now introduce the following tensorial notation:

$$\begin{aligned}\hat{\mathcal{T}} &:= \frac{\mathcal{T}_1}{\phi_1} - \frac{\mathcal{T}_2}{1-\phi_1}\\ &= \frac{K_BT}{V_m}\left[\frac{1}{N_x}(\det F)^{\frac{2}{3}} - \left(\frac{1}{2N_x} + \frac{1}{N_1} - \frac{1}{N_2}\right)\right]I + 2\mu FF^T,\\ \hat{\mathcal{G}} &:= \frac{K_BT}{V_m}\left[\frac{1}{N_x}{\phi_1}^{-1}(\det F)^{\frac{2}{3}} - {\phi_1}^{-1}\left(\frac{1}{2N_x} + \frac{1}{N_1}\right) - \frac{1}{N_2}(1-\phi_1)^{-1} - \chi\right]I\\ &\quad + 2\mu\phi_1^{-1}FF^T.\end{aligned} \tag{2.22}$$

We point out that the first notation represents a relative stress, and the second plays the role of a body force, as indicated by the following calculations:

$$\begin{aligned}&\frac{1}{\phi_1}\nabla\cdot\mathcal{T}_1 - \frac{1}{1-\phi_1}\nabla\cdot\mathcal{T}_2\\ &= \nabla\cdot(\phi_1^{-1}\mathcal{T}_1 - (1-\phi_1)^{-1}\mathcal{T}_2) - (\phi_1^{-2}\mathcal{T}_1 + (1-\phi_1)^{-2}\mathcal{T}_2)\nabla\phi_1\\ &= \nabla\cdot\hat{\mathcal{T}} + \hat{\mathcal{G}}(\nabla\phi_1) - \frac{\lambda\nabla\phi_1}{\phi_1(1-\phi_1)}.\end{aligned}$$

This allows us to rewrite (2.18) as follows:

$$\begin{aligned}&\frac{\partial \mathbf{U}}{\partial t}+(1-2\phi_1)(\nabla\mathbf{U})\mathbf{U}-(\mathbf{U}\otimes\mathbf{U})\nabla\phi_1+(\nabla\mathbf{V})\mathbf{U}+(\nabla\mathbf{U})\mathbf{V}\\ &=\nabla\cdot\hat{\mathcal{T}}+\hat{\mathcal{G}}(\nabla\phi_1)-\frac{\beta}{\phi_1(1-\phi_1)}\mathbf{U}.\end{aligned}\tag{2.23}$$

The governing equations of the gel consist of (2.17), (2.19), (2.20), (2.21), and (2.23). We note that the latter equation does not involve λ explicitly. Indeed, it appears only in the balance of linear momentum of the center of mass (2.17).

We conclude this subsection by discussing two limiting regimes modeled by the previously obtained system. First, let us consider the system obtained by setting $\mathbf{V}=0$. This corresponds to fields initially satisfying $\mathbf{V}=0$ and such that λ solves the equilibrium equation resulting from (2.17). The governing system for $\mathbf{U}$ and ϕ_1 becomes

$$\begin{aligned}&\frac{\partial \mathbf{U}}{\partial t}+(1-2\phi_1)(\nabla\mathbf{U})\mathbf{U}-(\mathbf{U}\otimes\mathbf{U})\nabla\phi_1\\ &=\nabla\cdot\hat{\mathcal{T}}+\hat{\mathcal{G}}(\nabla\phi_1)-\frac{\beta}{\phi_1(1-\phi_1)}\mathbf{U},\\ &F_t+((1-\phi_1)\mathbf{U})\cdot\nabla F=\nabla((1-\phi_1)\mathbf{U})F,\\ &\frac{\partial\phi_1}{\partial t}+(((1-\phi_1)\mathbf{U})\cdot\nabla)\phi_1+\phi_1\nabla\cdot((1-\phi_1)\mathbf{U})=0.\end{aligned}\tag{2.24}$$

This corresponds to purely diffusive regimes where no net motion of the center of mass of the mixture takes place. Of course, this type of regime would not be compatible, for instance, with flow geometries with prescribed nonzero boundary velocity (e.g., shearing flow).

Another regime fully characterized by the single velocity $\mathbf{V}$ can also be obtained from the governing system. Indeed, setting $\mathbf{U}=0$ in (2.17), (2.19), (2.20), (2.21), and (2.23) and accounting for viscoelastic stress, we get

$$\frac{\partial\mathbf{V}}{\partial t}+(\mathbf{V}\cdot\nabla)\mathbf{V}=\nabla\cdot\tau-\nabla\lambda,\tag{2.25}$$

$$\tau+\xi[\dot{\tau}-(\nabla\mathbf{V})^T\tau-\tau(\nabla\mathbf{V})]=\eta_0\mathbf{D}+(\mathcal{T}_1+\mathcal{T}_2),\tag{2.26}$$

$$F_t+\mathbf{V}\cdot\nabla F=(\nabla\mathbf{V})F,\tag{2.27}$$

$$\nabla\cdot\mathbf{V}=0,\tag{2.28}$$

where $\mathcal{T}_1+\mathcal{T}_2$ denotes the total elastic stress of the system (2.14) and (2.15). In terms of dimensional analysis, including a dissipative stress in the system to account for polymer-polymer friction results in an additional time scale $t_v=\frac{\eta_0}{L_0^2\beta}t_0$ much greater than that governing the relaxation of the diffusive velocity $\mathbf{U}$:

$$t_0=\frac{\beta L_0^2V_mN_x}{K_BT}.\tag{2.29}$$

(Here L_0 denotes a typical macroscopic length scale of the problem. The dimensional analysis is presented in a later section.) This is due to the fact that the polymer-polymer viscosity coefficient represented by η_0 is much larger that the polymer-solvent friction coefficient β. Heuristically, we may argue that in a system where both velocities are initially present, $\mathbf{U}$ relaxes to 0 much faster than $\mathbf{V}$. In the largest time scale, the mixture is governed by equations (2.25)–(2.28) of viscoelastic flow, in a regime characterized by transport only.

2.3. Relaxation regimes. Many studies of gels address the relaxation regimes, beyond transient behavior. For instance, such an approach has been used by Doi, Tanaka, and other researchers in many pioneering studies of gel swelling [19], [7]. We now indicate how the proposed equations relate to these earlier models. For this we return to our original component formulation (2.1) and (2.4)–(2.7). Addition of (2.4)–(2.5), together with the constraint equation $\phi_1 + \phi_2 = 1$, yields

$$\nabla \cdot (\phi_1 \mathbf{v}_1 + (1-\phi_1)\mathbf{v}_2) = 0.$$

With the stress tensors of the form (2.8) and (2.9) and setting $\psi_1 = \phi_1 W(F)$ and $\psi_2 \equiv 0$, we get

$$\begin{aligned} \mathcal{T}_1 &= \phi_1 \frac{\partial W}{\partial F} F^T - \phi_1 \lambda, \\ \mathcal{T}_2 &= -\phi_2 \lambda. \end{aligned}$$

The equations of balance of linear momentum become

$$\frac{\partial \mathbf{v}_1}{\partial t} + (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_1 = \nabla \cdot \left(\phi_1 \frac{\partial \psi_1}{\partial F} F^T\right) - \phi_1 \nabla\lambda - \beta(\mathbf{v}_1 - \mathbf{v}_2), \tag{2.30}$$

$$\frac{\partial \mathbf{v}_2}{\partial t} + (\mathbf{v}_1 \cdot \nabla)\mathbf{v}_2 = -\phi_2 \nabla\lambda + \beta(\mathbf{v}_1 - \mathbf{v}_2). \tag{2.31}$$

Moreover, neglecting inertial terms, we get

$$\begin{aligned} &\nabla \cdot (\phi_1 \mathbf{v}_1 + (1-\phi_1)\mathbf{v}_2) = 0, \\ &\nabla \cdot \mathcal{T}_1 + \mathbf{f}_1 = 0, \\ &\nabla \cdot \mathcal{T}_2 + \mathbf{f}_2 = 0. \end{aligned}$$

Taking into account that $\mathbf{f}_1 = \lambda \nabla \phi_1 - \beta(\mathbf{v}_1 - \mathbf{v}_2) = -\mathbf{f}_2$, the previous equations yield

$$\nabla \cdot (\phi_1 \mathbf{v}_1 + (1-\phi_1)\mathbf{v}_2) = 0,$$

$$\nabla \cdot \left(\phi_1 \frac{\partial W(F)}{\partial F} F^T\right) + \phi_1 \nabla\lambda - \beta(\mathbf{v}_1 - \mathbf{v}_2) = 0, \tag{2.32}$$

$$-\phi_2 \nabla\lambda + \beta(\mathbf{v}_1 - \mathbf{v}_2) = 0. \tag{2.33}$$

Addition of (2.32) and (2.33) yields

$$\nabla \cdot \left(\phi_1 \frac{\partial W(F)}{\partial F} F^T - \lambda\right) = 0, \tag{2.34}$$

where $\frac{\partial W(F)}{\partial F}$ is the Piola–Kirchoff stress tensor [10]. Taking into account the balance of mass equation, $\phi_1 \det F = 1$, we rewrite (2.34) as

$$\nabla \cdot \left(\det F^{-1} \frac{\partial W(F)}{\partial F} F^T - \lambda\right) = 0.$$

Note that $\sigma = \det F^{-1} \frac{\partial W(F)}{\partial F} F^T$ is the Cauchy stress tensor. Summarizing,

$$\begin{aligned} &F_t + (\mathbf{v}_1) \cdot \nabla F = \nabla(\mathbf{v}_1)F, \\ &\frac{\partial \phi_1}{\partial t} + (\mathbf{v}_1 \cdot \nabla)\phi_1 + \phi_1 \nabla \cdot \mathbf{v}_1 = 0, \\ &\nabla \cdot (\sigma - \lambda) = 0, \\ &-(1-\phi_1)\nabla \cdot \lambda + \beta(\mathbf{v}_1 - \mathbf{v}_2) = 0, \\ &\nabla \cdot (\phi_1 \mathbf{v}_1 + (1-\phi_1)\mathbf{v}_2) = 0. \end{aligned}$$

We observe that the first equation gives the chain rule, the second corresponds to balance of mass, the third is the force balance, the fourth corresponds to Darcy's law, and the last is the incompressibility condition of the mixture.

Remark. Many analyses found in the literature consider additional linearization of the previous system [23], [24], [25], [26], and [27].

3. One-dimensional geometry. We consider the gel occupying a strip domain

$$\Omega = \{(x, y, z) : -L \leq x \leq L\}$$

in the form of a strip, with $L > 0$ fixed. For instance, this type of geometry may be appropriate in modeling gliding behavior of bacteria by polysaccharide swelling [11]. We seek solutions of the governing system with $x = M(X, t)$, $x = N(X, t)$ denoting the deformation map of the polymer and the fluid, respectively. The fields of the problem are taken as follows:

$$\mathbf{V} = (V(x, t), 0, 0), \quad \mathbf{U} = (U(x, t), 0, 0), \quad \phi_1 = \phi_1(x, t), \quad \lambda = \lambda(x, y, z, t). \tag{3.1}$$

The deformation gradient matrix is

$$F = \mathrm{diag}(g(x, t), 1, 1), \quad \text{with} \tag{3.2}$$

$$g(x, t) = \frac{\partial M(X, t)}{\partial X}|_{X=M^{-1}(x,t)} = \det F. \tag{3.3}$$

The equation of balance of mass for the polymer in Lagrangian form is

$$\phi_1(x, t) \det F(x, t) = \alpha, \tag{3.4}$$

$$g(x, t) = \alpha \phi_1(x, t)^{-1}, \tag{3.5}$$

where $0 \leq \alpha \leq 1$ is a parameter of the problem. It represents the volume fraction of dry polymer in the reference configuration. For the deformation gradient F given in (3.2) and (3.5), we calculate

$$\begin{aligned}\mathcal{T}_1 = \frac{K_B T}{N_x V_m}\left(\alpha^{\frac{2}{3}}\phi_1^{\frac{1}{3}} - \left(\frac{1}{2} + \frac{N_x}{N_1}\right)\phi_1 - \chi N_x \phi_1(1 - \phi_1)\right)\mathbf{I} - \lambda\phi_1\mathbf{I} \\ + 2\mu\phi_1 \,\mathrm{diag}(1, 1, \alpha^2\phi_1^{-2}).\end{aligned}$$

$\mathcal{T}_2$ is as in (2.15). The second and third component equations in (2.17) give $\lambda = \lambda(x, t)$ (independent of y and z). Moreover, the equation $\nabla \cdot \mathbf{V} = 0$ together with (3.1a) gives $V = V(t)$. Prescribing $V(0) = 0$, $V(t) = 0$, $t > 0$ follows, provided that $\nabla \cdot \mathcal{T} = 0$ holds. The latter determines λ in terms of ϕ_1 and U, up to a constant. Moreover, ϕ_1 and U satisfy the equations

$$\frac{\partial \phi_1}{\partial t} + \frac{\partial(\phi_1(1 - \phi_1)U)}{\partial x} = 0, \tag{3.6}$$

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}U^2(1 - 2\phi_1) - G(\phi_1)\right) = -\frac{\beta}{\phi_1(1 - \phi_1)}U, \tag{3.7}$$

where

$$\begin{aligned}G(\phi_1) = \frac{K_B T}{V_m N_x}\left(-\frac{1}{2}\alpha^{2/3}\phi_1^{-\frac{2}{3}} - \left(\frac{1}{2} + \frac{N_x}{N_1}\right)\log\phi_1\right) \\ + \mu\alpha^2\phi_1^{-2} - \frac{K_B T \chi}{V_m}\phi_1 + \frac{K_B T}{N_2 V_m}\log(1 - \phi_1).\end{aligned} \tag{3.8}$$

The sign of $G'(\phi_1)$ is very relevant to the forthcoming analysis. Indeed, the condition $G'(\phi_1) < 0$ will be needed to guarantee hyperbolicity of the governing system, and therefore is a requirement for the propagation of the swelling front towards the solvent region. It turns out that $G'(\phi_1) < 0$ holds for polymer data. However, in the case of polysaccharides with data as previously given, there is a quantity $\phi_c = \phi_c(\mu)$, $0 < \phi_c < 1$, such that $G'(\phi_c) = 0$. This may be interpreted in terms of the onset of deswelling, observed in bacteria motility phenomenon [11]; it may also be associated with volume phase transitions observed in systems with a small elastic shear modulus [13].

We assume that, initially, the polymer occupies the strip $-L < x < L$, and the solvent is in the region $|x| > L$. At a later time $t > 0$, the gel occupies the region $-S(t) < x < S(t)$, where $x = S(t)$ denotes the position of the interface between the gel and the pure solvent. We look for symmetric solutions about the origin, $x = 0$, i.e., $\phi_1(x) = \phi_1(-x)$ and $U(-x) = -U(x)$, $x \in (-S(t), S(t))$. Therefore, it is sufficient to solve the problem for $x > 0$ only. Thus, we assume that (3.6) and (3.7) hold for $x \in (0, S(t))$, $t > 0$, for the fields (ϕ_1, U). Equation (2.7) for the incompressible inviscid solvent $\phi_2 = 1$ holds in the region $x > S(t)$. This implies that

$$v = 0, \quad \lambda = -\frac{K_B T}{V_m N_2} + c, \tag{3.9}$$

where c is a constant. In addition, we choose c so that the pressure in the fluid region takes a prescribed value, p_0; that is, $\lambda = p_0$, $x > S(t)$.

The boundary conditions of the problem consist of symmetry conditions at $x = 0$ and balance of forces at the interface $x = S(t)$. The former reduce to

$$\frac{\partial \phi_1}{\partial x}(0, t) = 0, \quad U(0, t) = 0. \tag{3.10}$$

Letting $-$ and $+$ denote the left and right limit at $S(t)$, respectively, we formulate boundary conditions. We first establish balance of forces

$$(\mathcal{T}_1 + \mathcal{T}_2)^-_{11} = (\mathcal{T}_2)^+_{11}. \tag{3.11}$$

Also, following Yamaue and Doi [24], we propose the following constitutive equation that expresses the degree of permeability of the interface. For a given $P > 0$, we assume that

$$\lambda^- - \lambda^+ = P. \tag{3.12}$$

Using the expressions of $\mathcal{T}_i$, $i = 1, 2$, and substituting (3.12) into (3.11) yields

$$\begin{aligned} P = {} & \frac{K_B T}{N_x V_m}\left(\alpha^{\frac{2}{3}}(\phi_1^-)^{\frac{1}{3}} - \left(\frac{1}{2} + \frac{N_x}{N_1}\right)\phi_1^-\right) - \chi\frac{K_B T}{V_m}\phi_1^-(1 - \phi_1^-) - \frac{K_B T}{N_x V_2}(1 - \phi_1^-) \\ & + 2\mu\alpha^2(\phi_1^-)^{-1} + \frac{K_B T}{N_2 V_m}. \end{aligned} \tag{3.13}$$

If the interface is fully permeable, then the pressure is continuous and $P = 0$ holds. In the case of charged polymers the discontinuity of λ is related to the net surface charge. Here, we take the point of view of P being a parameter of the problem. In particular, we observe from (3.13) that prescribing P allows us to the determine the saturation value $\phi_1^- \equiv \phi^*$. This, in turn, motives the definition of the interface as the location $x = S(t)$ with $\phi_1(S(t), t) = \phi^*$ that moves with the speed of the polymer. Specifically, the dynamics of the interface are described by the following equations:

$$\frac{dS}{dt}(t) = (1 - \phi_1(S(t), t))U(S(t), t), \tag{3.14}$$

$$S(0) = L, \tag{3.15}$$

$$\phi_1(S(t), t) = \phi^*. \tag{3.16}$$

The problem reduces to three equations, (3.6), (3.7), and (3.14), for the unknowns (ϕ_1, U, S), with boundary conditions (3.10) and (3.16) and initial conditions (3.15) and

$$\phi_1(x, 0) = \phi^0(x), \quad U(x, 0) = 0, \quad x \in (0, L). \tag{3.17}$$

We conclude this section by listing the time scales of the problem. Let L_0 denote a typical length; in polymer experiments, this would be of the order of centimeters. We find the following time constants:

$$t_0 = \frac{\beta L_0^2 V_m N_x}{K_B T}, \quad t_1 = \frac{\beta L_0^2 V_m}{K_B T \chi},$$
$$t_2 = \frac{\beta L_0^2 V_m N_2}{K_B T}, \quad t_3 = \frac{\beta L_0^2}{\mu}.$$

Comparing the time constants, we observe that

$$t_0 = \frac{N_x}{\chi} t_1, \quad t_1 = \frac{1}{\chi N_2} t_2, \quad t_2 = \frac{\mu N_2 V_m}{K_B T} t_3.$$

For the data in (1), we have that

$$t_2 \sim 10^{-1} t_3.$$

The previous data reflect the relative orders of magnitude of the time scales in polymer applications. The largest time constant is t_0, and t_2 is the smallest. Many works on gels focus on the time scale t_0 corresponding to the relaxation regime. In our work, we study the dynamics at the time scale t_2. After (3.6) and (3.7) are scaled to make them nondimensional, G takes the form

$$G(\phi_1) = C_0 \left(-\frac{1}{2} \alpha^{2/3} \phi_1^{-\frac{2}{3}} - \left(\frac{1}{2} + \frac{N_x}{N_1} \right) \log \phi_1 \right)$$
$$+ C_3 \alpha^2 \phi_1^{-2} - C_2 \phi_1 + C_1 \log(1 - \phi_1),$$

with dimensionless parameters

$$C_0 = \frac{\beta^2 L_0^2 V_m N_2^2}{K_B T N_x}, \quad C_1 = \frac{\beta^2 L_0^2 V_m N_2^2 \chi}{K_B T},$$
$$C_2 = \frac{\beta^2 L_0^2 V_m N_2}{K_B T}, \quad C_3 = \frac{\mu \beta^2 L_0^2 V_m^2 N_2^2}{K_B^2 T^2}.$$

The scaled equations and coefficients are employed in the numerical simulations.

4. The Cauchy problem. In this section we consider the Cauchy problem for (3.6) and (3.7) with initial conditions

$$\begin{aligned} U(x,0) &= U_0(x), \\ \phi_1(x,0) &= \phi_0(x). \end{aligned} \tag{4.1}$$

First we show that for the range of physical parameters corresponding to a semidry polymer, the governing system is of hyperbolic type with dissipation. Let us denote

$$\begin{aligned} \mathbf{u} &= (\phi_1, U)^T, \\ \mathbf{F} &= \left[\phi_1(1-\phi_1)U,\ \frac{1}{2}U^2(1-2\phi_1) - G(\phi_1)\right]^T, \\ \mathbf{G} &= \left[0,\ \frac{\beta}{\phi_1(1-\phi_1)}U\right]^T. \end{aligned}$$

The governing system becomes

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x}(\mathbf{u}) + \mathbf{G}(\mathbf{u}) = 0. \tag{4.2}$$

The gradient matrix is

$$D\mathbf{F} = \begin{bmatrix} (1-2\phi_1)U & \phi_1(1-\phi_1) \\ -U^2 - G'(\phi_1) & (1-2\phi_1)U \end{bmatrix}.$$

Eigenvalues λ_i, $i = 1, 2$, of $D\mathbf{F}$ are

$$\begin{aligned} \lambda_1 &= (1-2\phi_1)U + \sqrt{-\phi_1(1-\phi_1)(U^2 + G'(\phi_1))}, \\ \lambda_2 &= (1-2\phi_1)U - \sqrt{-\phi_1(1-\phi_1)(U^2 + G'(\phi_1))}. \end{aligned}$$

They are real and distinct provided that

$$U^2 + G'(\phi_1) < 0$$

holds. The hyperbolic region in the space (ϕ_1, U) consists of the points between the graphs of $U = \pm\hat{U}(\phi_1)$, with $\hat{U}(\phi_1) = \sqrt{|G'(\phi_1)|}$. The right eigenvectors of $D\mathbf{F}$ are

$$\mathbf{r}_1 = \begin{bmatrix} \sqrt{\frac{\phi_1(1-\phi_1)}{|U^2+G'(\phi_1)|}} \\ 1 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} -\sqrt{\frac{\phi_1(1-\phi_1)}{|U^2+G'(\phi_1)|}} \\ 1 \end{bmatrix}.$$

Let

$$V = \begin{bmatrix} \sqrt{\frac{\phi_1(1-\phi_1)}{|U^2+G'(\phi_1)|}} & -\sqrt{\frac{\phi_1(1-\phi_1)}{|U^2+G'(\phi_1)|}} \\ 1 & 1 \end{bmatrix}, \quad V^{-1} = \begin{bmatrix} \frac{1}{2}\sqrt{\frac{|U^2+G'(\phi_1)|}{\phi_1(1-\phi_1)}} & \frac{1}{2} \\ -\frac{1}{2}\sqrt{\frac{|U^2+G'(\phi_1)|}{\phi_1(1-\phi_1)}} & \frac{1}{2} \end{bmatrix}.$$

The characteristic coordinates $\mathbf{w} = V^{-1}\mathbf{u}$ give

$$w_1 = \frac{1}{2}\left(\sqrt{\frac{\phi_1}{1-\phi_1}|U^2 + G'(\phi_1)|} + U\right), \quad w_2 = \frac{1}{2}\left(-\sqrt{\frac{\phi_1}{1-\phi_1}|U^2 + G'(\phi_1)|} + U\right).$$

Let us define the pair of functions,

$$\eta(\phi_1, U) = -\int_\phi G(\rho)\,d\rho + \frac{1}{2}\phi_1(1-\phi_1)U^2, \quad q = \phi_1(1-\phi_1)U\left[-G(\phi_1) + \frac{1}{2}U^2(1-2\phi_1)\right].$$

Lemma 4.1. *The functions* (η, q) *form an entropy-flux pair for the hyperbolic system.*

Proof. We need to find $B = (B_1(\phi_1, U, x, t), B_2(\phi_1, U, x, t))$ such that

$$B = D\eta,$$
$$BD\mathbf{F} = Dq,$$

where $D = (\partial_{\phi_1}, \partial_U)^T$. It is easy to verify that

$$\begin{aligned} D\eta &= \left(\frac{1}{2}\phi_1(1-\phi_1)U^2 - G, \phi_1(1-\phi_1)U\right), \\ Dq &= \left(\left(\frac{1}{2}\phi_1(1-\phi_1)U^2 - G\right)((1-2\phi_1)U - (U^2+G')\phi_1(1-\phi_1)U),\right. \\ &\qquad \left.\left(\frac{1}{2}\phi_1(1-\phi_1)U^2 - G\right)\phi_1(1-\phi_1) + \phi_1(1-\phi_1)(1-2\phi_1)U^2\right). \end{aligned}$$

Choosing $B = (\frac{1}{2}\phi_1(1-\phi_1)U^2 - G,\ \phi_1(1-\phi_1)U)$, a direct calculation gives

$$BD\mathbf{F} = Dq.$$

Hence, (η, q) is an entropy-flux pair. □

We now introduce the concept of L^1-stability [4]. The Cauchy problem (3.6)–(3.7), (4.1) is said to be L^1*-stable* at an equilibrium state $U = \hat{U}$ and $\phi_1 = \hat{\phi}$ if there are positive numbers r and b such that any admissible bounded variation (BV) solution $U(x,t)$ and $\phi_1(x,t)$ of (3.6)–(3.7), (4.1) defined on any time interval $[0,T)$, $0 < T \le \infty$, and taking values in the ball $B_r(\hat{\phi}, \hat{U})$ of R^2 satisfies the inequality

(4.3)
$$\int_\infty^\infty |\phi_1(x,t) - \hat{\phi}| + |U(x,t) - \hat{U}|\,dx \le b\int_\infty^\infty |U_0(x) - \hat{U}| + |\phi_0(x) - \hat{\phi}|\,dx, \quad 0 \le t < T.$$

Lemma 4.2. *The Cauchy problem* (3.6)–(3.7)*,* (4.1) *is* L^1*-stable at the equilibrium state* $U = 0$*,* $\phi = \phi^*$*.*

Proof. From Lemma 4.1 and the form of the entropy function, we see that $\eta(\phi_1, U)$ is C^1 near the equilibrium $U = 0$, $\phi = \phi^*$. Thus, there are constants $r > 0$ sufficiently small and $d > 0$ such that

$$d^{-\frac{1}{2}}(|\phi - \phi^*| + |U|) \le |\eta(\phi_1, U)| \le d^{\frac{1}{2}}(|\phi - \phi^*| + |U|)$$

holds for any $(\phi, U) \in B_r(\phi^*, 0)$. Moreover, the entropy production is nonnegative; i.e.,

$$B\mathbf{G} = \beta U^2 \ge 0.$$

Hence, by an argument similar to [4], the L^1-stability is readily established. So, there exist r and d such that any admissible BV solution $U(x,t)$ and $\phi_1(x,t)$ of (3.6)–(3.7), (4.1) satisfies (4.3) with $\hat{U} = 0$ and $\hat{\phi} = \phi^*$. □

We next derive the linearization of the governing system about the equilibrium solution, $U = 0$ and $\phi_1 = \phi^*$, where $\phi^* \in (0,1)$ denotes the saturation volume fraction. We calculate

$$\begin{aligned}
&D\mathbf{F}(\mathbf{u}_e) = \begin{bmatrix} 0 & \phi^*(1-\phi^*) \\ -G'(\phi^*) & 0 \end{bmatrix},\\
&2\lambda_1 = \sqrt{\phi^*(1-\phi^*)|G'(\phi^*)|}, \quad 2\lambda_2 = -\sqrt{\phi^*(1-\phi^*)|G'(\phi^*)|},\\
&\mathbf{r}_1 = \left[\sqrt{\frac{\phi^*(1-\phi^*)}{|G'(\phi^*)|}}, 1\right]^T, \quad \mathbf{r}_2 = \left[-\sqrt{\frac{\phi^*(1-\phi^*)}{|G'(\phi^*)|}}, 1\right]^T,\\
&V = \begin{bmatrix} \sqrt{\frac{\phi^*(1-\phi^*)}{|G'(\phi^*)|}} & -\sqrt{\frac{\phi^*(1-\phi^*)}{|G'(\phi^*)|}} \\ 1 & 1 \end{bmatrix}.
\end{aligned}$$

We also calculate

$$\begin{aligned}
&DG(\mathbf{u}^*) = \begin{bmatrix} 0 & 0 \\ 0 & \frac{\beta}{\phi^*(1-\phi^*)} \end{bmatrix},\\
&R = V^{-1}(DG)V = \frac{\beta}{2\phi^*(1-\phi^*)} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad (4.4)\\
&\mathbf{w} = V^{-1}\mathbf{u}.
\end{aligned}$$

The resulting linear diagonal system is

$$\frac{\partial \mathbf{w}}{\partial t} + \operatorname{diag}(\lambda_1, \lambda_2)\frac{\partial \mathbf{w}}{\partial x} + R\mathbf{w} = 0.$$

We recall that a matrix A is strictly diagonally dominant if

$$A_{ii} - \sum_{i \neq j} |A_{ij}| \geq 0, \quad i, j = 1, 2, \ldots, n.$$

It is easy to check that the matrix R defined by (4.4) has positive entries in the principal diagonal and is diagonally dominant. However, since it is not strictly diagonally dominant, the exponential decay property of $TV_x\phi_1$ and TV_xU established in Theorem 2 of [6] cannot be asserted here. This prevents us from obtaining asymptotic stability of solutions with respect to time.

Since the Cauchy problem (3.6)–(3.7) with initial condition (4.1) is L^1-stable and R is diagonally dominant, the results proved by Dafermos [5] on existence and decay of BV solutions of weakly dissipative hyperbolic systems apply as follows.

Theorem 4.3. *Let r, b be defined in* (4.3). *Consider integrable initial data (ϕ_0, U_0) taking values in $B_r(\phi^*, 0)$. Let*

$$\sigma = \int_{-\infty}^{\infty} |U_0(x)| + |\phi_0(x) - \phi^*|\, dx$$

and

$$\omega = TV_{(-\infty,\infty)}|U_0(x)| + TV_{(-\infty,\infty)}|\phi_0(x)|$$

over $(-\infty, \infty)$. Then there are positive constants σ_0, ω_0, a, and μ such that, when $\sigma < \sigma_0$ and $\omega < \omega_0$, there exists an admissible global BV solution $U(x,t)$, $\phi_1(x,t)$ to *the Cauchy problem* (3.6)–(3.7) *with* (4.1)*, taking values in* $B_r(\phi^*, 0)$*. Furthermore, for each fixed* $t \in (0, \infty)$*,* $(\phi_1(x,t), U(\cdot,t))$ *is integrable and has bounded variation over* $(-\infty, \infty)$*:*

$$\int_{-\infty}^{\infty} |\phi_1(x,t) - \phi^*| + |U(x,t)|\, dx \leq b\sigma,$$

$$TV_{(-\infty,\infty)}|U(\cdot,t)| + TV_{(-\infty,\infty)}|\phi_1(\cdot,t) - \phi^*| \leq a\omega e^{-\mu t} + a\sigma.$$

Next we discuss the free boundary problem for the linearized system.

5. Analysis of the free boundary problem.

We now analyze a free boundary problem for (3.6) and (3.7). We consider the boundary and initial conditions given by (3.10), (3.16), (3.14), (3.15), and (3.17).

Also, since $\phi_1(x,0)$ is the initial volume fraction, we have the bound $0 < \phi_1(x,0) < 1$ for any $0 \leq x \leq L$. We consider the linearization of the above equations and boundary conditions with respect to the equilibrium pair $(\phi^*, 0)$.

5.1. Linearization and hyperbolicity condition.

We linearize (3.6) and (3.7) with respect to $\phi = \phi^*$ and $U = 0$, set $\bar{\phi} = \phi_1 - \phi^*$, and write the resulting system as follows:

$$\begin{aligned} &\frac{\partial \bar{\phi}}{\partial t} + \phi^*(1-\phi^*)\frac{\partial \bar{U}}{\partial x} = 0, \\ &\frac{\partial \bar{U}}{\partial t} - G'(\phi^*)\frac{\partial \bar{\phi}}{\partial x} = -\frac{\beta}{\phi^*(1-\phi^*)}\bar{U}. \end{aligned} \tag{5.1}$$

The linearized free boundary conditions (3.10), (3.16), (3.14), (3.15), and (3.17) become

$$\begin{aligned} &S'(t) = \bar{U}(S(t),t)(1-\phi^*), \\ &S(0) = L, \\ &\bar{\phi}(S(t),t) = 0, \\ &\bar{U}(0,t) = 0, \qquad t \in [0,T], \\ &\bar{\phi}(x,0) = \phi_0(x) - \phi^*, \qquad x \in [0,L], \\ &\bar{U}(x,0) = \bar{U}_0(x), \qquad x \in [0,L]. \end{aligned} \tag{5.2}$$

Because $0 < \phi_0 < 1$, we have $-\phi^* < \bar{\phi} < 1 - \phi^*$. Let us recall that $G'(\phi^*) < 0$, which ensures hyperbolicity. Let $\Gamma = G'(\phi^*)$. We make the following change of variables:

$$\begin{aligned} &\bar{\phi} = \sqrt{-\Gamma}(p+q), \\ &\bar{U} = \sqrt{\phi^*(1-\phi^*)}(p-q). \end{aligned} \tag{5.3}$$

Then the system (5.1) and (5.2) changes to the following equivalent system of equations:

$$p_t + \lambda_v p_x = -\frac{\beta}{2\phi^*(1-\phi^*)}(p-q), \tag{5.4}$$

$$q_t - \lambda_v q_x = \frac{\beta}{2\phi^*(1-\phi^*)}(p-q) \tag{5.5}$$

with free boundary conditions

$$\begin{aligned}
&S'(t) = (1-\phi^*)\sqrt{\phi^*(1-\phi^*)}(p(S(t),t) - q(S(t),t)),\\
&S(0) = L,\\
&p(S(t),t) + q(S(t),t) = 0, \qquad t \in [0,T],\\
&p - q = 0, \qquad t \in [0,T],\\
&p(x,0) = p_0(x), \qquad x \in [0,L],\\
&q(x,0) = q_0(x), \qquad x \in [0,L],
\end{aligned} \tag{5.6}$$

where $\lambda_v = \sqrt{-\Gamma\phi^*(1-\phi^*)}$.

5.2. Free boundary problem for the linearized system. A free boundary problem analogous to the present one is studied in [28]. However, the proof of the theorem relies on the assumption that the speed U is strictly positive at $x = 0$. Here, we generalize the global existence result to the case $U = 0$ at $x = 0$. We first point out that the local existence of solution stated next follows from the theorem in [12].

Theorem 5.1. *Let $0 < \phi^* < 1$, $\Gamma < 0$ be constant. Suppose that $p_0(x), q_0(x) \in C^1[0,L]$ satisfy compatibility conditions at $(0,0)$ and $(L,0)$. Let*

$$|p_0(L)| < \frac{\sqrt{-\Gamma}}{2(1-\phi^*)}. \tag{5.7}$$

Then there is a $t_0 > 0$ such that the free boundary problem (5.4), (5.5), and (5.6) has a unique solution $(p(x,t), q(x,t), S(t)) \in C^1(\bar{Q}_{S,t_0}) \times C^1(\bar{Q}_{S,t_0}) \times C^2[0,t_0]$, where

$$Q_{S,t_0} = \{(x,t),\ 0 < x < S(t),\ 0 < t < t_0\},$$

and t_0 depends on $S(0)$, $S'(0)$, $\|p_0(x)\|_{C^1[0,L]}$, and $\|q_0(x)\|_{C^1[0,L]}$.

In order to prove global existence of solutions, we will use the lemmas stated next. The proof of Lemma 5.2 makes use of the approach presented in [28].

Lemma 5.2 (see [28]). *Let $(\bar{\phi}, \bar{U}, S)$ be a C^1 solution of (5.1) and (5.2), and define p, q as in (5.3). Also, suppose that p_0 satisfies (5.7). Then*

$$|S'| < \lambda_v, \quad |p(S(t),t)| < \frac{\sqrt{-\Gamma}}{2(1-\phi^*)}, \quad |q(S(t),t)| < \frac{\sqrt{-\Gamma}}{2(1-\phi^*)}$$

hold for $t \in (0,t_0)$.

Proof. We proceed by contradiction. First, since $S'(0) > \lambda_v$, we suppose that there exists $0 < \hat{t} \le t_0$ such that

$$\lim_{t\to\hat{t}^-} S'(t) = \lambda_v \quad \text{and} \quad S'(t) < \lambda_v \quad \text{for } 0 < t < \hat{t}$$

holds. This implies that $S'(t)$ reaches its maximum as t approaches $\hat{t}$, and, consequently,

$$\lim_{t\to\hat{t}^-} S''(t) \ge 0.$$

This together with boundary condition (5.2a) yields

$$\lim_{t\to\hat{t}^-} \frac{d}{dt}\bar{U}(S(t),t) = \lim_{t\to\hat{t}^-} \frac{\partial U}{\partial x}\lambda_v + \frac{\partial U}{\partial t} \ge 0 \quad \text{at } (S(\hat{t}),\hat{t}). \tag{5.8}$$

On the other hand, because of boundary condition (5.2c), $\bar{\phi} \equiv 0$ on $(S(t), t)$ follows, and therefore

$$\frac{d}{dt}\bar{\phi}(S(t), t) = \frac{\partial \bar{\phi}}{\partial x}\lambda_v + \frac{\partial \bar{\phi}}{\partial t} = 0 \quad \text{on } (S(t), t). \tag{5.9}$$

Multiplying (5.1a) by $\frac{\sqrt{-\Gamma}}{\sqrt{\phi^*(1-\phi^*)}}$, adding the result to (5.1b), and taking limits as $x \to S(t)$ and $t \to \hat{t}^-$, we get

$$\frac{\sqrt{-\Gamma}}{\sqrt{\phi^*(1-\phi^*)}}\left(\frac{\partial \bar{\phi}}{\partial x}\lambda_v + \frac{\partial \bar{\phi}}{\partial t}\right) + \frac{\partial U}{\partial x}\lambda_v + \frac{\partial U}{\partial t} + \frac{\beta \lambda_v}{\phi^*(1-\phi^*)^2} = 0. \tag{5.10}$$

By application of (5.9), this reduces to

$$\frac{\partial U}{\partial x}\lambda_v + \frac{\partial U}{\partial t} = -\frac{\beta \lambda_v}{\phi^*(1-\phi^*)^2} < 0,$$

which contradicts inequality (5.8). We can follow the analogous argument in the case that

$$\lim_{t \to \hat{t}^-} S'(t) = -\lambda_v \quad \text{and} \quad S'(t) > -\lambda_v \quad \text{for } 0 < t < t_0.$$

Therefore,

$$|S'(t)| < \lambda_v = \sqrt{-\Gamma \phi^*(1-\phi^*)}$$

holds. This inequality together with boundary conditions (5.6a) and (5.6c) yields that the two remaining conclusions of the lemma hold. □

Lemma 5.3 (see [28]). *Let* $(\bar{\phi}, \bar{U}, S) \in C^1(Q_{S,t_0}) \times C^1(Q_{S,t_0}) \times C^2[0, t_0]$ *be a solution of the system* (5.1) *satisfying boundary conditions* (5.2). *Also, suppose that* $\max\{\|p_0\|_{L^\infty}, \|q_0\|_{L^\infty}\} \leq C_0 < \frac{\sqrt{-\Gamma}}{2(1-\phi^*)}$. *Then* $|p(x,t)| \leq C_0$ *and* $|q(x,t)| \leq C_0$ *for* $x \in Q_{S,t_0}$, *where* Q_{S,t_0} *is defined in Theorem* 5.1.

Proof. Arguing by contradiction, suppose otherwise; that is, there exist $\epsilon > 0$, $(x^*, t^*) \in Q_{S,t_0}$, such that

$$\begin{aligned} &|p(x^*, t^*)| = \max\{|p(x^*, t^*)|, |q(x^*, t^*)|\} = C_0 + \epsilon, \\ &\max\{|p(x,t)|, |q(x,t)|\} < C_0 + \epsilon \quad \text{for any fixed } t < t^*. \end{aligned}$$

First, notice that $S(t^*) > 0$, because, otherwise, if $S(t^*) = 0$, we can apply the boundary conditions at $x = 0$ and $x = S(t)$ to conclude that $p = q = 0 \leq C_0$. For $0 < x^* \leq S(t)$, there exists $\delta > 0$ such that $x^* - \lambda_v \delta > 0$ and $t^* - \delta > 0$. Integrating on characteristics, we have

$$\begin{aligned} p(x^*, t^*) = {}& e^{-\frac{\beta}{2\phi^*(1-\phi^*)}\delta} p(x^* - \lambda_v \delta, t^* - \delta) \\ &+ \int_0^\delta \frac{\beta}{2\phi^*(1-\phi^*)} e^{-\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - \delta)} q(x^* + (\tau - \delta), t^* + (\tau - \delta))\, \mathrm{d}\tau. \end{aligned}$$

With the help of the mean value theorem, we get the estimate

$$\begin{aligned} |p(x^*, t^*)| \leq {}& e^{-\frac{\beta}{2\phi^*(1-\phi^*)}\delta} |p(x^* - \lambda_v \delta, t^* - \delta)| \\ &+ |q(x^* + (\theta - \delta), t^* + (\theta - \delta))|(1 - e^{-\frac{\beta}{2\phi^*(1-\phi^*)}\delta}) \\ \leq {}& \max\{|p(x^* - \lambda_v \delta, t^* - \delta)|, |q(x^* + (\theta - \delta), t^* + (\theta - \delta))|\} \\ < {}& C_0 + \epsilon. \end{aligned}$$

This is a contradiction to the statement $|p(x^*, t^*)| = C_0 + \epsilon$. Now, if $x^* = 0$, there exist $0 < \delta < t^*$ such that $0 < \lambda_v \delta < S(t)$; applying the boundary condition $p(0, t^*) = q(0, t^*)$ yields

$$\begin{aligned}|p(0,t^*)| = |q(0,t^*)| &\le \max\{|q(\lambda_v\delta, t^* - \delta)|, |p(\lambda_v(\delta-\theta), t-(\delta-\theta))|\}\\ &< C_0 + \delta,\end{aligned}$$

where $0 < \theta < \delta$. This is again a contradiction to $|p(x^*, t^*)| = C_0 + \epsilon$. Hence $|p(x,t)| \le C_0$. We can follow an analogous argument in the case that $|q(x^*, t^*)| = \max\{|p(x^*, t^*)|, |q(x^*, t^*)|\} = C_0 + \epsilon$. □

Remark. Note that we can choose C_0 so that $\max\{\|\phi_0\|_{L^\infty}, \|U_0\|_{L^\infty}\} < \min\{\phi^*, 1 - \phi^*\}$. Hence $\bar{\phi}$ is always bounded by $\min\{\phi^*, 1-\phi^*\}$.

LEMMA 5.4. *Under the assumption of Lemma* 5.3 *and* $\int_0^L \bar{\phi}_0\,dx + \phi^* L > 0$, *there exist* $C > 0$ *and* $\eta > 0$ *such that*

$$C \ge S(t) \ge \eta > 0,$$

where η *depends on* C_0, p_0, q_0, ϕ^*, *and* Γ.

Proof. Integrating $\frac{\partial \bar{\phi}}{\partial t} + \phi^*(1-\phi^*)\frac{\partial \bar{U}}{\partial x} = 0$ with respect to x at fixed t, we get

$$\int_0^{S(t)} \frac{\partial \bar{\phi}}{\partial t}\,\mathrm{d}x + \phi^*(1-\phi^*)\bar{U}(S(t),t) = \phi^*(1-\phi^*)\bar{U}(0,t).$$

By applying boundary conditions (5.2a) and (5.2d), we get

$$\int_0^{S(t)} \frac{\partial \bar{\phi}}{\partial t}\,\mathrm{d}x + \phi^* S'(t) = 0.$$

Moreover, using the boundary condition $\bar{\phi}(S(t),t) = 0$, we get

$$\frac{d}{dt}\left\{\int_0^{S(t)} \bar{\phi}\,\mathrm{d}x + \phi^* S\right\} = 0.$$

Integration with respect to t gives

$$\int_0^{S(t)} \bar{\phi}\,\mathrm{d}x + \phi^* S - \int_0^L \bar{\phi}_0\,\mathrm{d}x - \phi^* L = 0.$$

Following Lemma 5.3, we can choose C_0 such that

$$\|\bar{\phi}\|_{L^\infty} < 2\sqrt{-\Gamma}C_0 \le \min\{\phi^*, 1-\phi^*\} - \epsilon.$$

Hence

$$S(t) \ge \frac{1}{2\sqrt{-\Gamma}C_0 + \phi^*}\left(\int_0^L \bar{\phi}_0\,\mathrm{d}x + \phi^* L\right) > 0,$$

and

$$S(t) \le \frac{1}{-2\sqrt{-\Gamma}C_0 + \phi^*}\left(\int_0^L \bar{\phi}_0\,\mathrm{d}x + \phi^* L\right). \qquad \square$$

LEMMA 5.5. *Then under the assumption of Lemma* 5.3,

$$|p_x(x,t)| \le C_3, \quad |q_x(x,t)| \le C_3$$

for $0 < t < t_0$*, where* C_3 *depends on* $\|p_0\|_{C^1[0,L]}$, $\|p_0\|_{C^1[0,L]}$*, and* C_0.

Proof. First, let us suppose that $t_0 < t_1 = \frac{L}{2\lambda_v}$ as in Figure 1. Differentiating (5.4) and (5.5) with respect to x yields

$$p_{xt} + \lambda_v p_{xx} = -\frac{\beta}{2\phi^*(1-\phi^*)}(p_x - q_x), \tag{5.11}$$

$$q_{xt} - \lambda_v q_{xx} = \frac{\beta}{2\phi^*(1-\phi^*)}(p_x - q_x) \tag{5.12}$$

and boundary condition

$$p_x + q_x = 0 \quad \text{at } x = 0, \tag{5.13}$$

$$(2(1-\phi^*)p - \sqrt{-\Gamma})p_x + (2(1-\phi^*)p + \sqrt{-\Gamma})q_x = 0 \quad \text{at } x = S(t). \tag{5.14}$$

Now, for fixed $0 < t < t_0$, define

$$\begin{aligned} A(t) &= \|p_x(x,t)\|_{L^\infty[0,S(t)]}, \\ B(t) &= \|q_x(x,t)\|_{L^\infty[0,S(t)]}, \\ C(t) &= \max\{A(t), B(t)\}, \\ C_1 &= \max\left\{\|p_{0x}\|_{L^\infty[0,L]}, \|q_{0x}\|_{L^\infty[0,L]}\right\}. \end{aligned}$$

If $(x,t) \in \Omega_3$, integrating along the characteristics of (5.11) and (5.12) yields

$$\begin{aligned} q_x(x,t) &= e^{-\frac{\beta}{2\phi^*(1-\phi^*)}(t-t_0)} q_x(S(t_0), t_0) \\ &\quad + \int_{t_0}^{t} \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} p_x(S(t_0) - \lambda_v(\tau - t_0), \tau)\, \mathrm{d}\tau, \end{aligned} \tag{5.15}$$

$$\begin{aligned} p_x(S(t_0), t_0) &= e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t_0} p_x(S(t_0) - \lambda_v t, 0) \\ &\quad + \int_{0}^{t_0} \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t_0)} q_x(S(t_0) - \lambda_v(t_0 - \tau), \tau)\, \mathrm{d}\tau. \end{aligned} \tag{5.16}$$

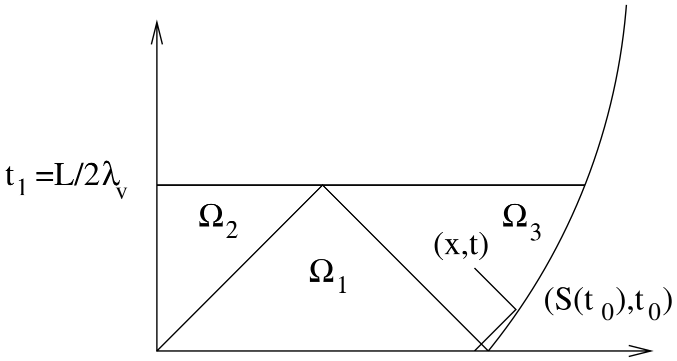


FIG. 1.

From the boundary condition (5.14), we have

$$q_x = \frac{\sqrt{-\Gamma} - 2(1-\phi^*)p}{2(1-\phi^*)p + \sqrt{-\Gamma}} p_x.$$

It follows from Lemma 5.3 that $|p(x,t)| \leq C_0 <$ const. Hence

$$\begin{aligned} |q_x| &\leq \left| \frac{\sqrt{-\Gamma} - 2(1-\phi^*)p}{2(1-\phi^*)p + \sqrt{-\Gamma}} \right| |p_x| \\ &\leq \left| \frac{\sqrt{-\Gamma} + 2(1-\phi^*)C_0}{\sqrt{-\Gamma} - 2(1-\phi^*)C_0} \right| |p_x| = C_2 |p_x|, \end{aligned}$$

where $C_2 := |\frac{\sqrt{-\Gamma}+2(1-\phi^*)C_0}{\sqrt{-\Gamma}-2(1-\phi^*)C_0}| > 1$. Combining (5.15) and (5.16) gives

$$|q_x(x,t)| \leq C_1 C_2 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + C_2 \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau.$$

For $(x,t) \in \Omega_1 \cup \Omega_2$, we have the estimate

$$|q_x(x,t)| \leq C_1 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau.$$

Hence

$$B(t) \leq C_1 C_2 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + C_2 \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau$$

holds. Now, for $(x,t) \in \Omega_2$, we have the following relations:

$$\begin{aligned} p_x(x,t) &= e^{-\frac{\beta}{2\phi^*(1-\phi^*)}\frac{x}{\lambda_v}} p_x\left(0, t - \frac{x}{\lambda_v}\right) \\ &\quad + \int_{t-\frac{x}{\lambda_v}}^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} q_x\left(\lambda_v\left(\tau - t + \frac{x}{\lambda_v}\right), \tau\right) \mathrm{d}\tau, \\ q_x\left(0, t - \frac{x}{\lambda_v}\right) &= e^{-\frac{\beta}{2\phi^*(1-\phi^*)}(t - \frac{x}{\lambda_v})} q_x\left(\lambda_v\left(t - \frac{x}{\lambda_v}\right), 0\right) \\ &\quad + \int_0^{t-x/\lambda_v} \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - (t - \frac{x}{\lambda_v}))} p_x\left(\lambda_v\left(t - \frac{x}{\lambda_v} - \tau\right), \tau\right) \mathrm{d}\tau. \end{aligned}$$

Application of the boundary condition (5.13) to the previous expression yields the estimate

$$|p_x(x,t)| \leq C_1 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau.$$

For $(x,t) \in \Omega_1 \cup \Omega_3$, we get

$$|p_x(x,t)| \leq C_1 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau.$$

Hence,

$$C(t) \leq C_1 C_2 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} + C_2 \int_0^t \frac{\beta}{2\phi^*(1-\phi^*)} e^{\frac{\beta}{2\phi^*(1-\phi^*)}(\tau - t)} C(\tau)\, \mathrm{d}\tau.$$

By Gronwall's inequality, we have

$$C(t) \leq C_1 C_2 e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t} e^{C_2(1-e^{-\frac{\beta}{2\phi^*(1-\phi^*)}t})} =: C_3.$$

Note that C_3 has a bound depending only on C_1, C_2, and t_0. Now for $t_0 > t_1$, we take the value of the solution at $t = t_1$ as initial condition and extend the estimate up to $t_1 + \frac{S(t_1)}{2\lambda_v}$. Since $S(t_1) > \eta > 0$, we can extend the estimate up to t_0. This completes the proof of the lemma. □

LEMMA 5.6. *Under the assumption of Lemma* 5.3, *there exists* $\epsilon > 0$ *such that*

$$|S'(t)| \leq \lambda_v(1-\epsilon).$$

Proof. Suppose that for any positive constant $\epsilon > 0$, there exist $0 < t^* \leq \min\{t_0, t_1\}$ and $0 < \delta < \epsilon$ such that $|S'(t^*)| = \lambda_v(1-\delta)$. Thus, we have

$$\begin{aligned} &\frac{\partial\bar{\phi}}{\partial x}\lambda_v(1-\delta) + \frac{\partial\bar{\phi}}{\partial t} = 0 \quad \text{on } (S(t), t), \\ &\lim_{t\to t^*} \frac{\partial U}{\partial x}\lambda_v(1-\delta) + \frac{\partial U}{\partial t} \geq 0. \end{aligned} \tag{5.17}$$

By following arguments analogous to those in the derivation of (5.10) and using (5.17), we have

$$\lambda_v\delta\left(\frac{\lambda_v}{\phi^*(1-\phi^*)}\frac{\partial\bar{\phi}}{\partial x} + \frac{\partial\bar{U}}{\partial x}\right) \leq -\frac{\beta}{\phi^*(1-\phi^*)}\frac{\lambda_v(1-\delta)}{1-\phi^*}. \tag{5.18}$$

By Lemma 5.5, $|p_x| \leq C_3$ and $|q_x| \leq C_3$, and thus $\bar{\phi}_x$ and $\bar{U}_x$, as linear combinations of p_x and q_x, are also bounded by a constant. Hence, $\frac{\lambda_v}{\phi^*(1-\phi^*)}\frac{\partial\bar{\phi}}{\partial x} + \frac{\partial\bar{U}}{\partial x} > -C$ holds, where $C = (\frac{\lambda_v}{\phi^*(1-\phi^*)} + 1)C_3$. Thus, letting $(x,t) \to (x^*, t^*)$ and using (5.18) and (5.17), we get

$$-\delta\lambda_v C + \frac{\beta(1-\delta)}{1-\phi^*} \leq 0,$$

which implies

$$\delta \geq \frac{\beta}{C(1-\phi^*)\lambda_v + \beta} > 0.$$

This contradicts inequality (5.18). □

We now state the following theorem on global existence. It also states that the interface remains bounded and cannot collapse to a point.

THEOREM 5.7. *Let the assumptions of Theorem* 5.1 *hold. Let* $\delta > 0$ *be such that*

$$\max\{\|p_0\|_{L^\infty}, \|q_0\|_{L^\infty}\} < \delta.$$

Then for any $t > 0$, *there exists a unique solution*

$$(U(x,\tilde{t}), \phi_1(x,\tilde{t}), S(\tilde{t})) \in C^1(\bar{Q}_{S,t}) \times C^1(\bar{Q}_{S,t}) \times C^2[0,t],$$

where $Q_{S,t} = \{(x,\tilde{t}),\ 0 < x < S(t),\ 0 < \tilde{t} < t\}$. *Moreover, there exist* η *and* μ *such that*

$$\mu_1 > S(\tilde{t}) > \eta > 0.$$

Proof. Because the transformation (5.3) is nonsingular, there exists $\delta > 0$ such that if

$$\max\{|p(x,t)|, |q(x,t)|\} < \delta < \frac{\sqrt{-\Gamma}}{2(1-\phi^*)},$$

then $|\phi| < \max\{\phi^*, 1-\phi^*,\}$ holds. Now, define t_{max} to be the maximum time of the local solution to (5.1) with (5.2). First, let us suppose that $t_{max} \leq t_1 = \frac{L}{2\lambda_v}$. By Lemma 5.6, $|S'(t)| \leq \lambda_v(1-\epsilon)$, with $\epsilon > 0$ depending on $\max\{\|p_0\|_{C^1}, \|q_0\|_{C^1}\}$. From Lemmas 5.3 and 5.5, we have that

$$\lim_{t\to t_{max}} \|p(x,t)\|_{C^1} < C_3,$$
$$\lim_{t\to t_{max}} \|q(x,t)\|_{C^1} < C_3.$$

On the other hand, by Lemma 5.4, there exists $\eta > 0$ such that $S(t) \geq \eta > 0$ for $0 < t < t_{max}$. Then by the local existence theorem, we can extend the solution beyond t_{max}; hence we have $t_{max} \geq t_1$. Now, taking $\{\bar{\phi}(x,t_1), U(x,t_1), S(t_1)\}$ as initial condition, we can extend the solution up to time $\frac{L}{2\lambda_v} + \frac{S(t_1)}{2\lambda_v}$. Note that $S(t) > \eta$, and therefore we can further extend the solution up to $\frac{L}{2\lambda_v} + m\frac{S(t_1)}{2\lambda_v}$ for any $m \in \mathbb{Z}^+$. Hence, the solution exists for all $t > 0$. □

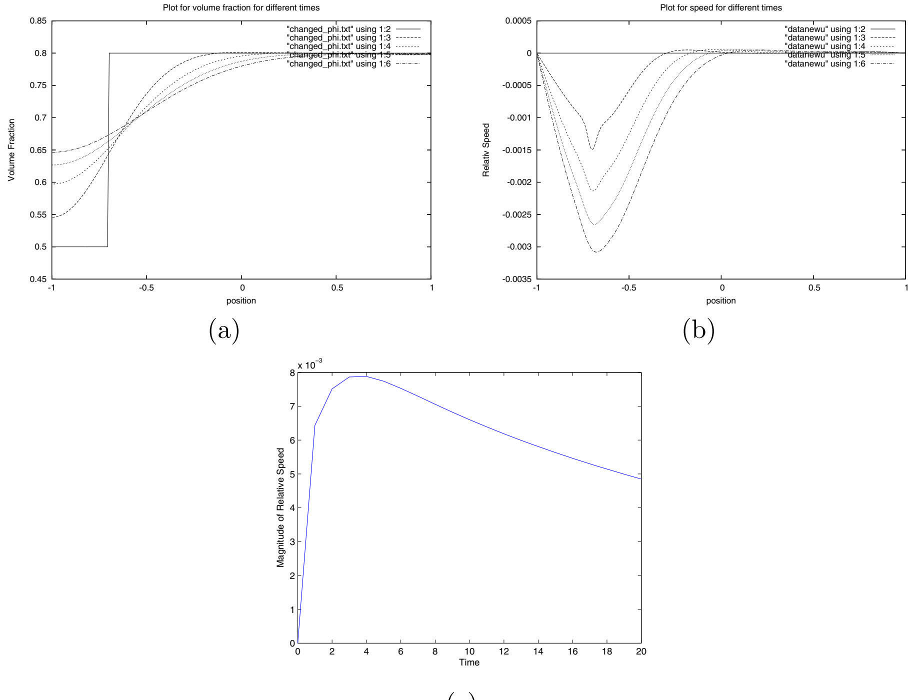
6. Numerical simulations. Based on the theoretical study, the following simulations are carried out for the Cauchy problem, with periodic boundary conditions. In order to achieve numerical stability, an artificial dissipation of form $\gamma\triangle^2\phi_1$ is added to the balance of mass equation, with $\gamma > 0$ small. Data are taken from Table 1.

It is well known that the value of β may be very sensitive to the volume fraction of the polymer. We consider the following expression for β [20]:

$$\beta(\phi_1) = (1-\phi_1)\phi_1^{\frac{2\gamma}{(3\gamma-1)}},$$

where $\gamma = 1/2$ for a Θ-solvent and $\gamma = 3/5$ for a good solvent [27]. Since we are interested in initial behavior where ϕ_1 jumps from 0 to 1, near the initial location of the interface between solvent and dry polymer, effectively, β is considerably smaller than the constant value in Table 1. The time and length scales of the calculations are taken to be $10^{-7}s$ and $10^{-5}m$, respectively. We take the polymer domain as the strip $(-1,3)$ centered at $x = 1$. Because of the symmetry of the problem, we show only the strip $(-1,1)$. We use the spectral method [17] to solve the nonlinear equation recursively; the result is shown in Figure 2. We observe that initially, the diffusive velocity builds up quickly, and the volume fraction changes rapidly. However, after a short initial time interval, the diffusive velocity decays, and the polymer volume fraction tends to an equilibrium saturation value.

7. Conclusion. We have analyzed the model developed in [3] for a two-component mixture of elastic solid and solvent and obtained effective equations for gels. We investigate the multiple time scales of the system and characterize the corresponding dynamics. We argue that studying early dynamics provides information on the evolution of swelling fronts and also gives a mathematical characterization of type II diffusion in polymers. We consider one-dimensional geometries and study the corresponding Cauchy problem, by applying the theory of Dafermos [5] on weakly dissipative hyperbolic systems. This allows us to establish existence and asymptotic



(a)

(b)

(c)

FIG. 2. (a) *polymer profile,* (b) *speed profile,* (c) *maximum relative velocity U versus time.*

properties of the solution of the Cauchy problem. We interpret the breakdown of the hyperbolicity condition, occurring at critical volume fractions for polysaccharide data, as an onset of deswelling. We formulate and study the free boundary problem for the linearized system and prove existence and uniqueness of solutions. This provides direct information on interface evolution; in particular, we show that the strip domain cannot collapse to a point. Follow-up studies address the nonlinear free boundary problem by combining estimates of the Cauchy problem with the information on the solution of the linearized problem.

Acknowledgments. The authors wish to thank Hans Weinberger, Suping Lyu, and Brandon Chabaud for many fruitful discussions. Both authors also thank Jie Shen for help and advice on the numerical study of the problem.

REFERENCES

[1] S. S. ANTMAN, *Nonlinear Problems of Elasticity*, 2nd ed., Springer, New York, 2005.

[2] R. B. BIRD, R. C. ARMSTRONG, AND O. HASSAGER, *Dynamics of Polymeric Liquids*, Wiley-Interscience, New York, 1987.

[3] B. CHABAUD, H. ZHANG, AND M. C. CALDERER, *Modeling of Viscoelastic Solid and Fluid Mixture with Applications to Gel*, Preprint 2189, Institute for Mathematics and Its Applications, 2008.

[4] C. M. DAFERMOS, *A system of hyperbolic conservation laws with frictional damping*, Z. Angew. Math. Phys., 46 (1995), pp. S294–S307.

[5] C. M. DAFERMOS, *Hyperbolic systems of balance laws with weak dissipation*, J. Hyperbolic Differ. Equ., 3 (2006), pp. 505–527.

[6] C. M. Dafermos and L. Hsiao, *Hyperbolic systems of balance laws with inhomogeneity and dissipation*, Indiana Univ. Math. J., 31 (1982), pp. 471–491.

[7] M. Doi and A. Onuki, *Dynamic coupling between stress and composition in polymer solutions and blends*, J. Phys. II France, 2 (1992), pp. 1631–1656.

[8] P. J. Flory, *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, NY, 1953.

[9] D. R. Gaskell, *Introduction to the Thermodynamics of Materials*, Taylor & Francis, Washington, DC, 1995.

[10] M. E. Gurtin, *An Introduction to Continuum Mechanics*, Math. Sci. Engrg. 158, Academic Press, New York, London, 1981.

[11] E. Hoiczyk, *Gliding motility in cyanobacteria: Observations and possible explanations*, Arch. Micro., 174 (2000), pp. 11–17.

[12] D. Li and W. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Mathematics Department, Duke University, Durham, NC, 1985.

[13] Y. Li and T. Tanaka, *Phase transitions of gels*, Annu. Rev. Mater. Sci., 22 (1992), pp. 243–277.

[14] F. Lin, C. Liu, and P. Zhang, *On hydrodynamics of viscoelastic fluids*, Comm. Pure Appl. Math., 58 (2005), pp. 1437–1471.

[15] C. Liu and N. J. Walkington, *An Eulerian description of fluids containing visco-hyperelastic particles*, Arch. Ration. Mech. Anal., 159 (2001), pp. 229–252.

[16] H. F. Mark and J. I. Kroschwitz, *Encyclopedia of Polymer Science and Engineering*, Wiley, New York, 1980.

[17] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[18] R. E. Showalter and N. J. Walkington, *Micro-structure models of diffusion in fissured media*, J. Math. Anal. Appl., 155 (1991), pp. 1–20.

[19] T. Tanaka and D. J. Filmore, *Kinetics of swelling of gels*, J. Chem. Phys., 70 (1979), pp. 1214–1218.

[20] N. L. Thomas and A. H. Windle, *A theory of case* II *diffusion*, Polymer, 23 (1982), pp. 529–542.

[21] N. L. Thomas and A. H. Windle, *A deformation model for case* II *diffusion*, Polymer, 21 (1980), pp. 613–619.

[22] C. Truesdell, *Rational Thermodynamics*, Springer, New York, 1984.

[23] T. Yamaue and M. Doi, *Swelling dynamics of constrained thin-plate gels under an external force*, Phys. Rev. E (3), 70 (2004), 011401.

[24] T. Yamaue and M. Doi, *Theory of one-dimensional swelling dynamics of polymer gels under mechanical constraint*, Phys. Rev. E (3), 69 (2004), 041402.

[25] T. Yamaue and M. Doi, *The stress diffusion coupling in the swelling dynamics of cylindrical gels*, J. Chem. Phys., 122 (2005), 084703.

[26] T. Yamaue, H. Mukai, K. Asaka, and M. Doi, *Electrostress diffusion coupling model for polyelectrolyte gels*, Macromolecules, 38 (2005), pp. 1349–1356.

[27] T. Yamaue, T. Taniguchi, and M. Doi, *The simulation of the swelling and deswelling dynamics of gels*, Molecular Phys., 102 (2004), pp. 167–172.

[28] T. Yang and F. Yi, *Global existence and uniqueness for a hyperbolic system with free boundary*, Discrete Contin. Dynam. Systems, 7 (2001), pp. 763–780.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1665–1687

A MATHEMATICAL MODEL FOR THE CONTROL AND ERADICATION OF A WOOD BORING BEETLE INFESTATION*

STEPHEN A. GOURLEY† AND XINGFU ZOU‡

Abstract. We propose a mathematical model for an infestation of a wooded area by a beetle species in which the larva develop deep in the wood of living trees. Due to the difficulties of detection, we presume that only a certain proportion of infested trees will be detected and that detection, if it happens, will occur only after some delay, which could be long. An infested tree once detected is immediately cut down and burned. The model is stage structured and contains a second time delay, which is the development time of the beetle from egg to adult. There is a delicate interplay between the two time delays due to the possibility in one case for a larva to mature even in a tree destined for destruction. We present conditions sufficient for infestation eradication and discuss the significance of the conditions, particularly in terms of the proportion of infested trees that need to be detected and removed. If the infestation is successfully eradicated, there are always a number of trees that completely escape infestation, and we compute lower bounds and an approximation for this number. Finally, we present the results of some numerical simulations.

Key words. delay, age-structure, infestation, eradication

AMS subject classifications. 34K25, 34K60, 92D30

DOI. 10.1137/060674387

1. Introduction.

In this paper we present a mathematical model of a possible strategy for the control of an infestation of wood boring beetles in which the larvae are burrowed deep in the wood of trees so that they are well protected from natural enemies but still have some intrinsic death rate. Our model also incorporates removal of trees that have been diagnosed as infested. Our work has been motivated in large part by recent infestations in parts of the US and Canada of *Anoplophora glabripennis*, commonly known as the Asian longhorned beetle (ALB), which attacks hardwood trees. Maple, willow, and elm constitute especially good hosts. Birch, ash, poplar, and numerous other tree species can also host this pest. The ALB has been intercepted at ports and warehouses all over North America, and it is believed that the pest entered the US (later spreading to Canada) in wooden packing crates used for imports from China. The beetle is native to China and Korea, and in China has caused major damage to poplar plantations with significant economic loss [6]. So far in North America the pest has affected only urban and suburban areas, but the potential impact of the ALB on the millions of acres of hardwood forests in the US and Canada could be devastating. It has been estimated that 1.2 billion trees could be at risk if the ALB were to become established in North America (see Nowak et al. [13] for a further discussion of the potential impact). Since the ALB species is not native to North America, it has no known natural enemies there (in fact, even in China it has few natural enemies).

*Received by the editors November 8, 2006; accepted for publication (in revised form) March 17, 2008; published electronically June 25, 2008. This research was partially supported by NSERC and NCE-MITACS of Canada and by PREA of Ontario.

http://www.siam.org/journals/siap/68-6/67438.html

†Department of Mathematics, University of Surrey, Guildford, Surrey GU2 7XH, UK (s.gourley@surrey.ac.uk).

‡Department of Applied Mathematics, University of Western Ontario, London, ON, Canada N6A 5B7 (xzou@uwo.ca).

The first time the ALB was found in North America seems to have been in Brooklyn, NY in 1996. Since then, the species has been found infesting trees in several US cities, including New York and Chicago, and it was found in Toronto, Canada for the first time in 2003.

Currently, the only known effective method of control of the ALB is to cut down, chip, and burn infested trees. Chemical controls are of limited potential because the larvae are deep within the tree and no effective chemical controls are yet available, though the effectiveness of certain insecticides is being investigated. It is current practice to establish quarantines around known infested areas and to monitor potential host trees within a certain distance of an infested area. Adult beetles are poor fliers but can fly short distances up to a few hundred yards to other neighboring trees, though in fact an adult often remains on the same host tree from which it emerged. ALB infested trees once removed are always replaced with a nonhost species. Other measures currently used for control include inspection of imports and the imposition of regulations on wooden packing material used for imports.

The ALB is a large beetle (up to 1.5 in long) which is easy to recognize. Adults are active from late spring to fall, when they perish. However, a large proportion of the life cycle of the beetle is spent in the larval stage deep within a tree, and this makes the detection of ALB activity more difficult. ALB larval activity on a tree is usually spotted either by inspectors or by members of the public. Warning signs of a tree being infested include exit holes (typically the diameter of a dime), oozing sap, sawdust accumulation, and unseasonable yellowing or drooping of leaves. The females prefer to lay their eggs in the upper canopy of a tree, though the lower trunk and branches can become affected if the upper canopy has been damaged by previous ALB activity. Preference for the upper canopy means that detection is more likely if inspectors are able to inspect it, for example, by climbing the tree. This slows down and increases the expense of systematically searching for ALB in a wooded area with the consequence of trees potentially missing detection. We shall aim to include these factors in the models we present in this paper. Since at present ALB affects only urban or suburban areas of North America, an infested tree probably stands a reasonable chance of being diagnosed as such, though possibly not until some time after the laying of eggs. An infested tree, if detected, will always be cut down and burned, but if not detected, the tree may survive several more years. Its death in this case will be due to the weakening of the tree and disruption of sap flow caused by the tunneling due to the larvae.

Adult ALB of both sexes are promiscuous, mating repeatedly and with different partners, according to greenhouse experiments reported in Morewood et al. [11]. The female will chew through the bark on the upper trunk and lay an egg. A single female can lay from 35 to 90 eggs during her lifetime of one season. The egg hatches after 1–2 weeks and the larva burrows deep into the tree, where it is very well protected from natural enemies, so there is a high probability of survival to maturity (this probability does depend to some extent on the tree species; see Morewood et al. [12]). Later in its development the larva enters the pupa stage, and finally the adult emerges from the tree. The whole duration from egg to adult lasts about one year but can be as long as 18 months. Adult emergence creates visible exit holes, which can sometimes be seen with binoculars, though the holes are not the only or necessarily the earliest sign of tree infestation.

In our models we shall deal with the issue of infested tree detection by supposing that there is a time delay σ between the time a tree becomes infested and the subsequent detection of ALB activity on the tree. To allow for the difficulties of detection we assume that only a certain fraction λ of infested trees is detected and then immediately removed. Such a fraction might well be close to 1, if the infestation is confined to a small suburban wooded area, but is likely to be much smaller if ALB infestation were to develop in a wilderness area. The models include a second time delay τ which models the developmental time of the beetle from egg to adult. As we shall see, there is a delicate interplay between the two delays σ and τ for reasons that will be explained early in the next section.

Subsection 2.1 deals with the case when $\sigma < \tau$ and presents the model for this case together with a detailed derivation. Positivity of solutions is established, which is not at all obvious from the appearance of the model equations. Then, sufficient conditions are presented for infestation eradication and a lower bound is given for the final number of susceptible trees.

Subsection 2.2 addresses the case $\sigma > \tau$. The model equations for this case look similar to those for the $\sigma < \tau$ case, but there are subtle differences, and a different strategy is required to establish positivity of solutions. For this case we again present an inequality that is sufficient for infestation eradication. Subsection 2.3 deals with the case when $\lambda = 1$. In subsection 2.4 we present a Laplace transform analysis that enables us to calculate analytically the final number of susceptible trees in the case when the number of adult beetles is small throughout the course of the infestation.

In some parameter regimes the infestation is not eradicated but instead all trees become infested, with the number of susceptible trees tending to zero (we shall see, however, that if the infestation is eradicated, there are always some trees that escape infestation). In situations in which eradication has not been achieved, the beetle numbers typically evolve to a periodic cycle in a forest in which all trees end up being infested. In reality the goal, of course, is to prevent this from happening and aim for eradication, but without necessarily requiring the detection and removal of every single infested tree.

2. Model derivation and analysis.

Let $T_s(t)$ and $T_i(t)$ denote, respectively, the numbers of susceptible and infested trees. Trees can survive about 4 years of infestation before they die; this is somewhat longer than the timescale on which we would want to remove infested trees, so disease-induced death of infested trees is neglected. It is also reasonable to neglect natural mortality of trees which occurs on an even longer timescale (e.g., of 100 years or more for maple trees). The quantities $L(t)$ and $A(t)$ denote the numbers of larval and adult beetles.

The model we shall develop involves two time delays. We shall let σ denote the amount of time that elapses between the instant that a tree becomes infested and the subsequent instant at which there is a probability λ of its being removed and burnt as a consequence (i.e., a fraction $\lambda \in [0, 1]$ of trees that become infested are removed σ time units later). We shall let τ denote the time it takes between the laying of an egg and subsequent emergence of an adult beetle, i.e., the duration of the larval stage, which in this paper is understood to include all pre-adult stages. It will be clear that the cases $\sigma < \tau$ and $\sigma > \tau$ have to be dealt with separately. For example, if $\sigma < \tau$, then the period between time of infection of a tree and its subsequent removal (if it is removed) is not long enough to allow any larva to mature; however, a larva can still mature if it is fortunate enough to be in a host tree that is not removed. On the other hand if $\sigma > \tau$, then it is possible for larvae to complete their development into maturity and escape as adult beetles even if all infested trees are removed.

2.1. The case $\sigma < \tau$. For the case when $\sigma < \tau$, we propose the following model:

$$\frac{dT_s(t)}{dt} = -\beta A(t)T_s(t), \tag{2.1}$$

$$\frac{dT_i(t)}{dt} = \beta A(t)T_s(t) - \lambda\beta A(t-\sigma)T_s(t-\sigma), \tag{2.2}$$

$$\begin{aligned}\frac{dL(t)}{dt} &= T_i(t)B(A(t)) - \mu_L L(t) - \lambda\beta A(t-\sigma)T_s(t-\sigma)\int_0^\sigma B(A(t-a))e^{-\mu_L a}\,da\\ &- e^{-\mu_L\tau}B(A(t-\tau))\left[T_i(t-\tau) - \lambda\beta\int_0^\sigma A(\tilde{a}+t-\tau-\sigma)T_s(\tilde{a}+t-\tau-\sigma)\,d\tilde{a}\right],\end{aligned} \tag{2.3}$$

$$\begin{aligned}\frac{dA(t)}{dt} &= e^{-\mu_L\tau}B(A(t-\tau))\left[T_i(t-\tau) - \lambda\beta\int_0^\sigma A(\tilde{a}+t-\tau-\sigma)T_s(\tilde{a}+t-\tau-\sigma)\,d\tilde{a}\right]\\ &\quad - \mu_A A(t).\end{aligned} \tag{2.4}$$

Here, all parameters are nonnegative with $\lambda \in [0,1]$. We justify each equation in (2.1)–(2.4) below.

Susceptible trees are converted to infested trees via contact with adult beetles, and it is assumed that the rate at which this occurs is given by the law of mass action (equation (2.1)). There is no term reflecting regeneration of trees, partly because this would occur on a relatively slow timescale and partly because tree replanting would be of some nonsusceptible species and might not take place at all while the infestation is still present.

The second term in the right-hand side of (2.2) represents the cutting down (and subsequent burning) of infested trees. It is assumed that when a tree becomes infested, it may be recognized and diagnosed as such but only after some time delay σ, which models the time taken for the tree to begin exhibiting telltale signs. A fraction $\lambda \in [0,1]$ of trees which become infested are later cut down, so that at time t the rate of cutting down of infested trees is λ times the infection rate at the earlier time $t-\sigma$.

The first term in the right-hand side of (2.3) is the birth rate, assumed proportional to the total number of infested hosts (recall that a tree is considered infested after contact with an adult beetle) and also to $B(A(t))$, where the function $B(\cdot)$ is the number of eggs laid per unit time per tree. We assume that all eggs hatch successfully, but some of the larvae may die in the tree at a rate μ_L. Of course, larvae may also die due to trees being cut down and burned. The rate at which this happens is evidently related to the total cutting down rate of infested and dead trees and is computed as follows:

$$\underbrace{\lambda\beta A(t-\sigma)T_s(t-\sigma)}_{\text{rate of tree removal}}\underbrace{\int_0^\sigma B(A(t-a))e^{-\mu_L a}\,da}_{\text{larvae per tree}}.$$

The last term in (2.3), which also appears in (2.4), is the rate at time t at which larvae mature into adult beetles. We next provide a rigorous derivation of this term, which is essentially the birth rate at the earlier time $t-\tau$ (τ being the length of the maturation period), modified to allow for natural mortality and mortality due to tree removal. Death of adult beetles is modeled by the last term in (2.4).

To derive the maturation term for the case when $\sigma < \tau$, let $b(t,a)$ denote the density of beetles at time t of age a. Larval beetles and adult beetles are, respectively, those of age less than τ and greater than τ so that

$$L(t) = \int_0^\tau b(t,a)\,da, \qquad A(t) = \int_\tau^\infty b(t,a)\,da. \tag{2.5}$$

It is larval beetles that are affected by removal of trees, but we must note that, since we assume $\sigma < \tau$ here, the larvae that are removed due to tree removal will have age up to at most σ. Any older larvae will necessarily be in trees that escaped removal. We model this as follows using von Foerster age-structured equations:

$$\frac{\partial b}{\partial t} + \frac{\partial b}{\partial a} = -\mu_L b(t,a) - \lambda\beta A(t-\sigma)T_s(t-\sigma)B(A(t-a))e^{-\mu_L a}, \qquad a < \sigma, \tag{2.6}$$

$$\frac{\partial b}{\partial t} + \frac{\partial b}{\partial a} = -\mu_L b(t,a), \qquad \sigma < a < \tau. \tag{2.7}$$

The explanation for the last term in the right-hand side of (2.6) is as follows. It is the rate at which larvae of age a are removed due to tree removal, and is therefore the rate of tree removal $\lambda\beta A(t-\sigma)T_s(t-\sigma)$, times the larvae density of age a per tree that is thus removed, which will be the birth rate per tree at time $t-a$ times the probability of survival to age a, i.e., $B(A(t-a))e^{-\mu_L a}$.

For adult beetles,

$$\frac{\partial b}{\partial t} + \frac{\partial b}{\partial a} = -\mu_A b(t,a), \qquad a > \tau. \tag{2.8}$$

Differentiating the expression for $A(t)$ in (2.5) gives

$$\frac{dA}{dt} = b(t,\tau) - \mu_A A, \tag{2.9}$$

assuming that $b(t,\infty) = 0$. We shall find $b(t,\tau)$ in terms of the birth rate $b(t,0)$ by integrating (2.6) and (2.7) along characteristics. Since we previously defined $B(A(t))$ as the number of eggs laid per unit time per tree, the birth rate $b(t,0)$ is given by

$$b(t,0) = T_i(t)B(A(t)).$$

Define

$$b_\zeta(a) = b(a+\zeta, a).$$

Then, for $a \le \sigma$,

$$\begin{aligned}\frac{db_\zeta(a)}{da} &= \left[\frac{\partial b}{\partial t} + \frac{\partial b}{\partial a}\right]_{t=a+\zeta}\\ &= -\mu_L b_\zeta(a) - \lambda\beta A(a+\zeta-\sigma)T_s(a+\zeta-\sigma)B(A(\zeta))e^{-\mu_L a}.\end{aligned}$$

Solving this for $b_\zeta(a)$ leads to

(2.10)

$$b(t,a) =$$

$$e^{-\mu_L a}B(A(t-a))\left[T_i(t-a) - \lambda\beta\int_0^a A(\tilde a + t - a - \sigma)T_s(\tilde a + t - a - \sigma)\,d\tilde a\right], \quad a \le \sigma.$$

For ages a between σ and τ, an easier calculation involving (2.7) shows that

$$b(t,a) = b(t-(a-\sigma),\sigma)e^{-\mu_L(a-\sigma)},$$

and $b(t-(a-\sigma),\sigma)$ can be found from (2.10) giving that, for $\sigma \le a \le \tau$,

$$b(t,a) = e^{-\mu_L a}B(A(t-a))\left[T_i(t-a) - \lambda\beta\int_0^\sigma A(\tilde{a}+t-a-\sigma)T_s(\tilde{a}+t-a-\sigma)\,d\tilde{a}\right]. \tag{2.11}$$

This expression looks rather like the corresponding one for $a \le \sigma$ (expression (2.10)), but note that the upper limit on the integral is now σ rather than a. This difference is very important. Putting $a = \tau$ into (2.11) gives an expression for $b(t,\tau)$, which we insert into (2.9), thereby completing the derivation of (2.4).

The expression for $b(t,\tau)$ is those larvae of age τ and represents the rate at which larval beetles become adult beetles (the adult recruitment rate). Expression (2.11) with $a=\tau$ shows that this is basically the birth rate at the earlier time $t-\tau$ (corrected for larval mortality) minus those larvae that would have made it to adulthood but were removed and destroyed with their host tree (the integral term represents accumulated removal of trees that could have hosted the larvae we are discussing, i.e., trees that became infested at times between $t-\tau-\sigma$ and $t-\tau$). An alternative viewpoint is that the term in square brackets in (2.11) (with $a=\tau$) is the "effective" number of host trees at time $t-\tau$ when the eggs are laid since, in this $\sigma<\tau$ regime, those trees that are removed might as well not have been there in the first place.

The derivation of the larval equation (2.3) is by differentiation of the expression for $L(t)$ in (2.5), breaking the integral up into the $a<\sigma$ and $a \in (\sigma,\tau)$ contributions, and using (2.6) and (2.7).

As we are considering the $\sigma<\tau$ situation, in which it is impossible for a larva to complete its development in a tree that gets removed, we should expect that when $\lambda = 1$ (i.e., every tree that becomes infested is later removed), the maturation rate $b(t,\tau)$ should be zero. When (and only when) $\lambda=1$, the number of infested trees at time t is given by

$$T_i(t) = \int_{t-\sigma}^t \beta A(\xi)T_s(\xi)\,d\xi.$$

Therefore, indeed, $b(t,\tau)=0$ when $\lambda=1$, for $\sigma<\tau$.

2.1.1. Initial data and positivity. The initial data for system (2.1), (2.2), (2.3), (2.4) has the form

$$\begin{aligned}
T_s(t) &= T_s^0(t) \ge 0, \qquad t\in[-\tau-\sigma,0],\\
A(t) &= A^0(t) \ge 0, \qquad t\in[-\tau-\sigma,0],\\
T_i(t) &= \beta\int_{-\sigma}^0 A^0(t+\xi)T_s^0(t+\xi)\,d\xi, \qquad t\in[-\tau,0],\\
L(0) &= \int_{-\tau}^0 B(A^0(\xi))e^{\mu_L\xi}T_i(\xi)\,d\xi\\
&\quad -\lambda\beta\int_0^\sigma\int_{-\tau}^{-a} B(A^0(\xi))A^0(a+\xi-\sigma)T_s^0(a+\xi-\sigma)e^{\mu_L\xi}\,d\xi\,da,
\end{aligned} \tag{2.12}$$

where $T_s^0(t)$ and $A^0(t)$ are prescribed continuous functions. The last two conditions in (2.12) are compatibility conditions, by which we mean that the initial data for T_i

and L is not arbitrary but is computed from the prescribed nonnegative initial data for T_s and A. For example, the larvae present at time $t = 0$ are the offspring of the adults at earlier times, and the modeling described thus far in this paper leads to the expression in (2.12) for $L(0)$. The initial data (2.12) including the compatibility conditions is the only ecologically relevant initial data.

PROPOSITION 1. *Let $\sigma < \tau$ and $\lambda \in [0, 1]$, and let the initial data for system* (2.1), (2.2), (2.3)*,* (2.4) *satisfy* (2.12)*. Let the birth function be bounded and satisfy $B(0) = 0$ and $B(A) > 0$ for $A > 0$. Then all variables in* (2.1)–(2.4) *are defined for all $t > 0$ and are bounded and remain nonnegative for $t > 0$.*

Proof. Existence of solutions follows by the method of steps which, since $\sigma < \tau$ here, is carried out successively on the steps $t \in [0, \sigma]$, $t \in [\sigma, 2\sigma]$, etc. It is easily seen that this works on the subsystem consisting of (2.1), (2.2), and (2.4). For example when $t \in [\sigma, 2\sigma]$, all arguments of the delayed variables in (2.4) remain less than σ. So, local existence is assured for $T_s(t)$, $T_i(t)$, and $A(t)$ (global existence will be shown later). It turns out that $L(t)$ has an explicit expression in terms of these variables (see (2.17) below). Next we shall show that all variables remain nonnegative for as long as they are defined.

It is obvious that $T_s(t) \geq 0$ for all $t > 0$. Next we shall show nonnegativity of $A(t)$. This will be achieved by jointly showing nonnegativity of $A(t)$ and the function $f(t)$ defined by

$$f(t) = T_i(t) - \lambda\beta \int_{t-\sigma}^{t} A(\xi)T_s(\xi)\, d\xi. \tag{2.13}$$

Using (2.2) we find that

$$\frac{df}{dt} = \beta(1 - \lambda)A(t)T_s(t). \tag{2.14}$$

Also, (2.4) can be rewritten in a form involving $f(t)$,

$$\frac{dA(t)}{dt} = e^{-\mu_L \tau} B(A(t - \tau))f(t - \tau) - \mu_A A(t). \tag{2.15}$$

As regards initial conditions for $f(t)$, note that when $t \in [-\tau, 0]$, from (2.12),

$$\begin{aligned} f(t) &= \beta \int_{-\sigma}^{0} A(t + \xi)T_s(t + \xi)\, d\xi - \lambda\beta \int_{t-\sigma}^{t} A(\xi)T_s(\xi)\, d\xi \\ &= (1 - \lambda)\beta \int_{-\sigma}^{0} A(t + \xi)T_s(t + \xi)\, d\xi \geq 0, \end{aligned} \tag{2.16}$$

so $f(t)$ is nonnegative initially. The functions $f(t)$ and $A(t)$ can be viewed as satisfying (2.14) and (2.15), considered here as a coupled system in which $T_s(t)$ is some known nonnegative function. The assumptions on $B(\cdot)$ and the nonnegativity of $f(t)$ and $A(t)$ for $t \leq 0$ allow us to deduce, from Theorem 2.1 on page 81 of Smith [16], that $f(t) \geq 0$ and $A(t) \geq 0$ for all $t > 0$.

From (2.2),

$$\begin{aligned} \frac{dT_i(t)}{dt} &= \beta A(t)T_s(t) - \lambda\beta A(t - \sigma)T_s(t - \sigma) \\ &\geq \beta A(t)T_s(t) - \beta A(t - \sigma)T_s(t - \sigma) = \frac{d}{dt} \int_{t-\sigma}^{t} \beta A(\xi)T_s(\xi)\, d\xi. \end{aligned}$$

Thus $T_i(t) - \int_{t-\sigma}^{t} \beta A(\xi) T_s(\xi)\, d\xi$ is an increasing function of t which, by (2.12), is zero when $t = 0$. Hence

$$T_i(t) \geq \int_{t-\sigma}^{t} \beta A(\xi) T_s(\xi)\, d\xi \geq 0$$

for $t > 0$.

Finally, we address positivity of $L(t)$. The solution of (2.3) subject to the compatibility condition in (2.12) is most easily found from (2.5) with (2.10) and (2.11) and turns out to be

$$\begin{aligned} L(t) \quad = \quad & \int_{t-\tau}^{t} B(A(\xi)) e^{-\mu_L (t-\xi)} T_i(\xi)\, d\xi \\ & - \lambda\beta \int_0^{\sigma} \int_{t-\tau}^{t-a} B(A(\xi)) A(a + \xi - \sigma) T_s(a + \xi - \sigma) e^{-\mu_L (t-\xi)}\, d\xi\, da. \end{aligned} \tag{2.17}$$

Indeed, the most general solution of (2.3), which is linear in $L(t)$, is the above expression plus $C \exp(-\mu_L t)$ for some constant C, and the latter term would have to be set to zero to satisfy (2.12). The state space of initial data in (2.12) is forward invariant in this sense.

Using (2.13) and nonnegativity of $f(t)$, it can be shown that

$$L(t) \geq \int_0^{\sigma} \int_{t-\bar{\xi}}^{t} B(A(\xi)) A(\bar{\xi} + \xi - \sigma) T_s(\bar{\xi} + \xi - \sigma) e^{-\mu_L (t-\xi)}\, d\xi\, d\bar{\xi}.$$

Hence $L(t) \geq 0$. Having shown nonnegativity of each solution variable while it is defined, we may now establish global existence. This can be done by establishing a priori bounds. Indeed, nonnegativity of the variables and (2.1) imply that $T_s(t)$ is decreasing, so that $T_s(t)$ is bounded above and below. Since $df/dT_s = -(1 - \lambda)$, it follows that $f(t)$ is also bounded above and below. It then follows from (2.15), using boundedness of $B(\cdot)$, that $A(t)$ is bounded. Then (2.13) implies that $T_i(t)$ is bounded. Then expression (2.17) shows that $L(t)$ is bounded and the proof is complete. □

2.1.2. Infestation eradication. In this section we present conditions on the parameters which guarantee that for initial data satisfying (2.12) in section 2.1.1 the infestation is eradicated.

THEOREM 1. *Let* $\sigma < \tau$ *and* $\lambda \in [0, 1]$*, and let the initial data for system* (2.1), (2.2), (2.3), (2.4) *satisfy* (2.12)*. Let the birth function satisfy* $B(0) = 0$ *and* $0 < B(A) \leq B'(0)A$ *for* $A > 0$*. Assume further that*

$$(1 - \lambda) e^{-\mu_L \tau} B'(0) \left(T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right) < \mu_A. \tag{2.18}$$

Then the solution of the system satisfies $A(t) \to 0$ *and* $L(t) \to 0$ *as* $t \to \infty$*, so that the infestation is eradicated.*

Furthermore, the final number $T_s(\infty)$ *of susceptible trees is not less than*

$$T_s(0) \exp\left(-\beta \left\{ \frac{A^0(0) + (1-\lambda) e^{-\mu_L \tau} B'(0) \left(T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right) \int_{-\tau}^{0} A^0(\xi)\, d\xi}{\mu_A - (1-\lambda) e^{-\mu_L \tau} B'(0) \left(T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right)} \right\} \right). \tag{2.19}$$

Remark 1. Inequality (2.18) essentially arises from the worst imaginable (but not actually attainable) scenario in which the entire forest becomes infested before the infestation is eradicated. If the infestation is successfully eradicated, there will always be some trees that escape infestation, and (2.19) gives a lower bound for this number of escaped trees. However, in the proof of the theorem we are faced with the difficulty that the number $T_s(\infty)$ is not known exactly. We need an upper bound for the function $f(t)$ defined by (2.13) involving only known quantities, and in achieving this we are forced to use $T_s(\infty) \geq 0$ so that we are effectively considering an extreme but unattained situation. The quantity $T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi$ is the total initial number of trees (susceptible and infested), all of which would end up infested in this worst case scenario. But a fraction $1 - \lambda$ of them is *not* removed. The left-hand side of (2.18) is the per capita maturation rate at large times, being the per capita egg laying rate per tree $B'(0)$, multiplied by the number of infested trees at large times corrected for tree removal, multiplied by the survival probability $e^{-\mu_L \tau}$.

Proof of Theorem 1. From (2.1) and nonnegativity of solutions, $T_s(t)$ is a decreasing nonnegative function which therefore approaches a nonnegative limit as $t \to \infty$.

Recall the function $f(t)$ defined by (2.13). From (2.1) and (2.14) note that

$$\frac{df}{dT_s} = -(1 - \lambda).$$

Hence

$$f(t) = -(1 - \lambda) T_s(t) + T_i(0) - \lambda \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi + (1 - \lambda) T_s(0).$$

Since $T_s(t) \geq 0$,

$$\begin{aligned} f(t) &\leq T_i(0) - \lambda \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi + (1 - \lambda) T_s(0) \\ &= (1 - \lambda) \left[T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right] \qquad \text{using (2.12).} \end{aligned}$$

Using the form of (2.4) involving $f(t)$ (i.e., (2.15)) and the above upper bound for $f(t)$, we obtain

(2.20)
$$\begin{aligned} \frac{dA(t)}{dt} &\leq (1 - \lambda) e^{-\mu_L \tau} B(A(t - \tau)) \left(T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right) - \mu_A A(t) \\ &\leq (1 - \lambda) e^{-\mu_L \tau} B'(0) A(t - \tau) \left(T_s(0) + \beta \int_{-\sigma}^{0} A^0(\xi) T_s^0(\xi)\, d\xi \right) - \mu_A A(t). \end{aligned}$$

Since the right-hand side of this is increasing as a function of the delayed variable $A(t - \tau)$, we may say that $A(t)$ is bounded above by the solution of the differential equation obtained by replacing "$\leq$" by "$=$" and satisfying the same initial data as that for A (see, e.g., Theorem 1.1 on page 78 of Smith [16]). By a straightforward and standard argument involving the characteristic equation of the resulting linear delay equation, utilizing Theorem 5.1 on page 92 of Smith [16] to assure ourselves that the dominant eigenvalue is real, we conclude that $A(t) \to 0$ as $t \to \infty$ under the hypothesis (2.18). The proof that $L(t) \to 0$ follows from (2.3) and the theory of asymptotically autonomous systems (see, e.g., Castillo-Chavez and Thieme [2]). Furthermore since the convergence of $A(t)$ to zero will be exponential, we are assured that $\int_0^\infty A(t)\,dt < \infty$, which is necessary for what follows.

Integrating (2.21) from 0 to ∞ and rearranging gives

$$\int_0^\infty A(t)\,dt \le \frac{A^0(0) + (1-\lambda)e^{-\mu_L\tau}B'(0)\left(T_s(0) + \beta\int_{-\sigma}^0 A^0(\xi)T_s^0(\xi)\,d\xi\right)\int_{-\tau}^0 A^0(\xi)\,d\xi}{\mu_A - (1-\lambda)e^{-\mu_L\tau}B'(0)\left(T_s(0) + \beta\int_{-\sigma}^0 A^0(\xi)T_s^0(\xi)\,d\xi\right)}.$$

Inserting this estimate into

$$T_s(\infty) = T_s(0)\exp\left(-\beta\int_0^\infty A(t)\,dt\right),$$

which follows from (2.1), gives the estimate (2.19). The proof is complete. □

2.2. The case $\boldsymbol{\sigma > \tau}$. If $\sigma > \tau$, then the maturation time for a larva is less than the time that elapses between a tree becoming infested and its possible subsequent removal σ time units later. Thus, if an egg is laid on a particular tree just after that tree became infested, then that larva is not at risk of having its host tree removed and burned. If an egg is laid on a tree that became infested some time ago, such that the tree now has less than τ time units to go before the time at which there is a probability λ of its being removed, that larva could still survive to maturation if its host is not actually removed. These considerations lead us to the following different model equations:

$$\frac{dT_s(t)}{dt} = -\beta A(t)T_s(t), \tag{2.21}$$

$$\frac{dT_i(t)}{dt} = \beta A(t)T_s(t) - \lambda\beta A(t-\sigma)T_s(t-\sigma), \tag{2.22}$$

(2.23)
$$\frac{dL(t)}{dt} = T_i(t)B(A(t)) - \mu_L L(t) - \lambda\beta A(t-\sigma)T_s(t-\sigma)\int_0^\tau B(A(t-a))e^{-\mu_L a}\,da$$
$$- e^{-\mu_L\tau}B(A(t-\tau))\left[T_i(t-\tau) - \lambda\beta\int_0^\tau A(\tilde{a}+t-\tau-\sigma)T_s(\tilde{a}+t-\tau-\sigma)\,d\tilde{a}\right],$$

(2.24)
$$\frac{dA(t)}{dt} = e^{-\mu_L\tau}B(A(t-\tau))\left[T_i(t-\tau) - \lambda\beta\int_0^\tau A(\tilde{a}+t-\tau-\sigma)T_s(\tilde{a}+t-\tau-\sigma)\,d\tilde{a}\right]$$
$$- \mu_A A(t).$$

All parameters are again nonnegative with $\lambda \in [0,1]$.

This system looks very similar to the corresponding system for $\sigma < \tau$ described in subsection 2.1, but there is an important difference: the upper limits in the integrals in (2.23) and (2.24) are τ rather than σ. This is because in this case, we need only break down $b(t,a)$ into two cases: (i) when $a < \tau$, $b(t,a)$ is governed by the PDE in (2.6); and (ii) for $a > \tau$, $b(t,a)$ is governed by the PDE in (2.8). The derivation of (2.23) and (2.24) is similar to but even simpler than that of (2.3) and (2.4).

The first goal for this model is again to prove positivity of the solutions corresponding to the initial compatibility conditions (2.12). First, (2.21) gives $T_s(t) = T_s(0)\exp(-\beta\int_0^t A(\theta)\,d\theta) > 0$ for all $t \geq 0$. In order to obtain the positivity of other variables, we introduce a new variable,

$$\begin{aligned} g(t) &= T_i(t) - \lambda\beta\int_0^{\tau} A(\tilde{a}+t-\sigma)T_s(\tilde{a}+t-\sigma)\,d\tilde{a} \\ &= T_i(t) - \lambda\beta\int_{t-\sigma}^{t+\tau-\sigma} A(\xi)T_s(\xi)\,d\xi. \end{aligned} \tag{2.25}$$

Then we have

$$\frac{dg(t)}{dt} = \beta A(t)T_s(t) - \lambda\beta A(t+\tau-\sigma)T_s(t+\tau-\sigma), \tag{2.26}$$

and the adults equation (2.24) can be rewritten as

$$\frac{dA(t)}{dt} = e^{-\mu_L\tau}B(A(t-\tau))g(t-\tau) - \mu_A A(t). \tag{2.27}$$

Since $g(t)$ does not behave as nicely as the function $f(t)$ in subsection 2.1, we have to tackle positivity via another strategy.

Note that, for $t \in [0, \sigma-\tau]$,

$$\frac{dg(t)}{dt} \geq \beta A(t)T_s(t) - \beta A(t-(\sigma-\tau))T_s(t-(\sigma-\tau)) = \frac{d}{dt}\int_{t-(\sigma-\tau)}^{t} \beta A(\xi)T_s(\xi)\,d\xi,$$

implying that $g(t) - \int_{t-(\sigma-\tau)}^{t} \beta A(\xi)T_s(\xi)\,d\xi$ is increasing on $[0, \sigma-\tau]$. Thus, for $t \in [0, \sigma-\tau]$, we have

$$\begin{aligned} &g(t) - \int_{t-(\sigma-\tau)}^{t} \beta A(\xi)T_s(\xi)\,d\xi \\ &\quad\geq g(0) - \int_{-(\sigma-\tau)}^{0} \beta A(\xi)T_s(\xi)\,d\xi \\ &\quad= T_i(0) - \lambda\int_{-\sigma}^{-(\sigma-\tau)} \beta A(\xi)T_s(\xi)\,d\xi - \int_{-(\sigma-\tau)}^{0} \beta A(\xi)T_s(\xi)\,d\xi \\ &\quad\geq T_i(0) - \int_{-\sigma}^{-(\sigma-\tau)} \beta A(\xi)T_s(\xi)\,d\xi - \int_{-(\sigma-\tau)}^{0} \beta A(\xi)T_s(\xi)\,d\xi \\ &\quad= T_i(0) - \int_{-\sigma}^{0} \beta A(\xi)T_s(\xi)\,d\xi = 0. \end{aligned} \tag{2.28}$$

Hence,

$$g(t) \geq \int_{t-(\sigma-\tau)}^{t} \beta A(\xi)T_s(\xi)\,d\xi \quad \text{for} \quad t \in [0, \sigma-\tau]. \tag{2.29}$$

This implies that if $A(0) > 0$ and $T_s(0) > 0$ (recalling that $A^0(\theta)$ and $T^0(\theta)$ are continuous in (2.12)), then $g(0) > 0$. Let $\delta = \min\{\tau, \sigma-\tau\}$; then either both $A(t)$ and $g(t)$ remain positive on $[0, \delta]$ or $A(t)$ will become negative before $g(t)$. In the latter case, there is a $t_0 \in (0,\delta]$ such that $A(t_0)=0$, $A(t)>0$, and $g(t)>0$ for $t<t_0$. It follows from (2.27) that

$$A'(t_0) = e^{-\mu_L \tau} B(A(t_0-\tau))g(t_0-\tau) > 0,$$

which is impossible. This contradiction shows that $A(t)>0$ and $g(t)>0$ for all $t\in[0,\delta]$. Repeating this process, we can obtain the positivity of $A(t)$ and $T_s(t)$ on $[0,2\delta]$ and, by induction, on $[0,n\delta]$ for all positive integers n, giving the positivity of $A(t)$ and $T_s(t)$ for all $t>0$.

Once we have obtained the positivity of $A(t)$ and $T_s(t)$, the positivity of $T_i(t)$ and $L(t)$ can be obtained in precisely the same way as in subsection 2.1. Therefore, we have obtained the following positivity result for (2.21)–(2.24), parallel to Proposition 1 for (2.1)–(2.4).

Proposition 2. *Let $\sigma>\tau$ and $\lambda\in[0,1]$, and let the initial data for system* (2.21)–(2.24) *satisfy* (2.12) *with $A(0)>0$ and $T_s(0)>0$. Let the birth function satisfy $B(0)=0$ and $B(A)>0$ for $A>0$. Then all variables in* (2.21)–(2.24) *remain nonnegative for $t>0$.*

We next seek conditions under which the infestation will be eradicated. Adding (2.21) and (2.26) gives

$$\frac{d}{dt}[g(t)+T_s(t)] = -\lambda\beta A(t-(\sigma-\tau))T_s(t-(\sigma-\tau)) \le 0.$$

Thus,

$$\begin{aligned} g(t) &\le g(t)+T_s(t) \le g(0)+T_s(0)\\ &= T_i(0) - \lambda\beta\int_{-\sigma}^{-(\sigma-\tau)} A(\xi)T_s(\xi)\,d\xi + T_s(0)\\ &= \beta\int_{-\sigma}^{0} A(\xi)T_s(\xi)\,d\xi - \lambda\beta\int_{-\sigma}^{-(\sigma-\tau)} A(\xi)T_s(\xi)\,d\xi + T_s(0)\\ &= (1-\lambda)\beta\int_{-\sigma}^{-(\sigma-\tau)} A(\xi)T_s(\xi)\,d\xi + \beta\int_{-(\sigma-\tau)}^{0} A(\xi)T_s(\xi)\,d\xi + T_s(0). \end{aligned} \tag{2.30}$$

Therefore, in the case $\sigma>\tau$, if (2.18) is replaced by

(2.31)
$$e^{-\mu_L\tau}B'(0)\left[(1-\lambda)\beta\int_{-\sigma}^{-(\sigma-\tau)} A(\xi)T_s(\xi)\,d\xi + \beta\int_{-(\sigma-\tau)}^{0} A(\xi)T_s(\xi)\,d\xi + T_s(0)\right] < \mu_A,$$

then by an argument similar to that in the proof of Theorem 1, we can conclude that the solution of system (2.21)–(2.24) with the initial compatibility conditions (2.12) satisfies $A(t)\to 0$ and $L(t)\to 0$ as $t\to\infty$; that is, the infestation will be eradicated. We have proved the following theorem.

Theorem 2. *Let $\sigma>\tau$ and $\lambda\in[0,1]$, and let the initial data for system* (2.21), (2.22), (2.23), (2.24) *satisfy* (2.12) *with $T_s(0)>0$ and $A(0)>0$. Let the birth function satisfy $B(0)=0$ and $0<B(A)\le B'(0)A$ for $A>0$. Assume further that* (2.31) *holds. Then the solution of the system satisfies $A(t)\to 0$ and $L(t)\to 0$ as $t\to\infty$ so that the infestation is eradicated.*

As in section 2.1, under the assumptions in Theorem 2 and based on the $T_s(t)$ and $A(t)$ equations (i.e., (2.21) and (2.27)) and the estimate (2.30) for the function

$g(t)$, we can also establish a lower bound for $T_s(\infty)$, the final number of susceptible trees. Indeed, if we denote by M the right-hand side of (2.30), then by an argument similar to that for obtaining the estimate (2.19), we can derive the following lower bound for $T_s(\infty)$:

$$T_s(\infty) \geq T_s(0) \exp \left(-\beta \left\{ \frac{A^0(0) + e^{-\mu_L \tau} B'(0) M \displaystyle\int_{-\tau}^{0} A^0(\xi)\, d\xi}{\mu_A - e^{-\mu_L \tau} B'(0) M} \right\} \right). \tag{2.32}$$

2.3. The case $\boldsymbol{\lambda = 1}$. In this subsection we discuss the situation when $\lambda = 1$, which means that every tree that becomes infested is removed σ time units later. We shall deal with both of the cases $\sigma < \tau$ and $\sigma > \tau$. In either case, the expression

$$T_i(t) = \int_{t-\sigma}^{t} \beta A(\xi) T_s(\xi)\, d\xi \tag{2.33}$$

is available to us, though the implications of this fact for the cases $\sigma < \tau$ and $\sigma > \tau$ are different. Expression (2.33) gives us the total number of infested trees at time t, which, when $\lambda = 1$, is simply the accumulation of all new infestations over the previous σ units of time (the corresponding expression for $T_i(t)$ when $\lambda \neq 1$ appears later (expression (2.39)) and in this case includes trees that became infested before time $t - \sigma$ but escaped detection).

In the case $\sigma < \tau$ the use of expression (2.33) in (2.4) yields $A'(t) = -\mu_A A(t)$ so that $A(t) \to 0$, and the infestation is eradicated. This is hardly surprising since if every infested tree is removed σ time units after the time of infestation, and $\sigma < \tau$, then no larva is being given enough time to mature.

If $\sigma > \tau$, the equation for $A(t)$ is (2.24), and the use of (2.33) leads to

$$\frac{dA(t)}{dt} = e^{-\mu_L \tau} B(A(t-\tau)) \int_{t-\sigma}^{t-\tau} \beta A(\xi) T_s(\xi)\, d\xi - \mu_A A(t). \tag{2.34}$$

We will use this equation to show that the infestation is always eradicated when $\lambda = 1$, regardless of the values of σ and τ. The truth of this result even in the $\sigma > \tau$ case is a little surprising, since this case offers the possibility of some larvae maturing before their host tree is destroyed. However, numerical simulations do show that even though eradication is still the final outcome, there may be a long transient in which the infestation grows worse for a while before dying out.

THEOREM 3. *Let $\lambda = 1$, and consider the system consisting of either* (2.1)–(2.4) *or* (2.21)–(2.24)*, with initial data satisfying* (2.12) *with $T_s(0) > 0$ and $A(0) > 0$. Let the birth function satisfy $B(0) = 0$ and $0 < B(A) \leq B'(0)A$ for $A > 0$. Then the infestation is eradicated, that is, $(T_i(t), L(t), A(t)) \to (0, 0, 0)$ as $t \to \infty$.*

Proof. It is sufficient to show that $A(t) \to 0$. Then $L(t) \to 0$ follows trivially from (2.3) or (2.23) as appropriate, and $T_i(t) \to 0$ follows from (2.33). We have already commented above that if $\sigma < \tau$, then $A(t) \to 0$ trivially; this can be extended to $\sigma = \tau$. So it remains to consider the case $\sigma > \tau$, and it is here that we shall make use of (2.34), which has to be coupled to (2.21). The latter equation, together with nonnegativity of solutions, implies that $T_s(t)$ must decay monotonically to some nonnegative limit as $t \to \infty$. If $T_s(t) \to 0$, then the asymptotic limit of (2.34) is just $A'(t) = -\mu_A A(t)$, and so $A(t) \to 0$. So it remains to consider the case that

$\lim_{t\to\infty} T_s(t) > 0$. If this is so, then from (2.21) it follows that

$$\int_0^\infty A(t)\,dt < \infty. \tag{2.35}$$

If we can show that, additionally,

$$\int_0^\infty |A'(t)|\,dt < \infty, \tag{2.36}$$

then a result from integration theory assures us that $\lim_{t\to\infty} A(t) = 0$. The integral in (2.34) can be expressed in terms of $T_s(t-\tau)$ and $T_s(t-\sigma)$ using (2.21), and the monotonicity properties of $T_s(t)$ therefore assure us of the existence of a finite C such that

$$\left|\int_{t-\sigma}^{t-\tau} \beta A(\xi) T_s(\xi)\,d\xi\right| \le C \quad \text{for all } t \ge 0.$$

Therefore, integration of (2.34) and using the estimate $0 < B(A) \le B'(0)A$ lead to

$$\int_0^\infty |A'(t)|\,dt \le Ce^{-\mu_L\tau}B'(0)\int_{-\tau}^0 A^0(t)\,dt + (Ce^{-\mu_L\tau}B'(0) + \mu_A)\int_0^\infty A(t)\,dt.$$

Hence $\int_0^\infty |A'(t)|\,dt < \infty$. The proof is complete. □

2.4. Approximation of $T_s(\infty)$. In subsections 2.1 and 2.2 we have seen that, under the eradication condition (2.18) for system (2.1)–(2.4), or condition (2.31) for system (2.21)–(2.24), the infestation is eradicated and the beetle does not affect the entire forest, since $A(t) \to 0$ and $T_s(t) \to T_s(\infty) =: T_s^* > 0$. Lower bounds for T_s^* are provided by inequalities (2.19) and (2.32) for models (2.1)–(2.4) and (2.21)–(2.24), respectively. In this subsection we provide an approach for obtaining an approximation for T_s^*. We present our analysis only for (2.21)–(2.24); the analysis for (2.1)–(2.4) is similar.

It is easily seen using (2.26) that the function

$$t \to g(t) - \int_{t-(\sigma-\tau)}^{t} \beta A(\xi) T_s(\xi)\,d\xi$$

is increasing for all $t \ge 0$. But (2.30) implies that this function is bounded from above. Therefore, it has a limit as $t \to \infty$, and thus $g(t)$ also has a limit as $t \to \infty$, since we consider the situation in which $A(t) \to 0$. It follows from (2.25) that $T_i^* := \lim_{t\to\infty} T_i(t)$ also exists and $T_i^* = \lim_{t\to\infty} g(t)$. Now, when t is sufficiently large, $A(t)$ becomes very small, and hence we may study the linearized approximation of (2.27) for small $A(t)$, and also replace $g(t-\tau)$ by its (as yet undetermined) limiting value T_i^* as $t \to \infty$ to obtain

$$\frac{dA(t)}{dt} = e^{-\mu_L\tau}B'(0)T_i^* A(t-\tau) - \mu_A A(t). \tag{2.37}$$

Letting p be the transform variable, the Laplace transform $\bar{A}(p)$ of $A(t)$ is

$$\bar{A}(p) = \frac{A^0(0) + e^{-\mu_L\tau}T_i^* B'(0)e^{-p\tau}\displaystyle\int_{-\tau}^0 A^0(\xi)e^{-p\xi}\,d\xi}{p - e^{-\mu_L\tau}T_i^* B'(0)e^{-p\tau} + \mu_A}.$$

The structure of (2.37) assures us that the dominant eigenvalue of its characteristic equation is real (see Smith [16]). Furthermore, this dominant eigenvalue is negative, since we consider the situation when $A(t) \to 0$, so let it be $-p^*(T_i^*)$, to emphasize the dependence on T_i^*, with $p^*(T_i^*) > 0$. This dominant eigenvalue is also the singularity of $\bar{A}(p)$ of greatest real part. By the inversion formula for Laplace transforms, $A(t)$ can be expressed as a contour integral, which by Cauchy's residue formula can be evaluated as a sum of residues of the poles of $\bar{A}(p)$. We shall include in this calculation only the pole of greatest real part, which is located at $p = -p^*(T_i^*)$, to give

$$\begin{aligned} A(t) &\approx \operatorname{res}\left(\bar{A}(p)e^{pt},\ p = -p^*(T_i^*)\right) \\ &= \frac{e^{-p^*(T_i^*)\,t}\left[A^0(0) + e^{-\mu_L\tau}T_i^*B'(0)e^{p^*(T_i^*)\tau}\displaystyle\int_{-\tau}^{0} A^0(\xi)e^{p^*(T_i^*)\xi}\,d\xi\right]}{1 + \tau e^{-\mu_L\tau}T_i^*B'(0)e^{p^*(T_i^*)\tau}}. \end{aligned}$$

Now, from (2.21),

$$T_s^* = T_s(0)\exp\left(-\beta\int_0^\infty A(t)\,dt\right),$$

giving

$$T_s^* \approx T_s(0)\exp\left(-\frac{\beta\left[A^0(0) + e^{-\mu_L\tau}T_i^*B'(0)e^{p^*(T_i^*)\tau}\displaystyle\int_{-\tau}^{0} A^0(\xi)e^{p^*(T_i^*)\xi}\,d\xi\right]}{p^*(T_i^*)(1 + \tau e^{-\mu_L\tau}T_i^*B'(0)e^{p^*(T_i^*)\tau})}\right). \tag{2.38}$$

The solution of (2.22), subject to the initial value formula for $T_i(0)$ from (2.12), is

$$T_i(t) = \int_{t-\sigma}^{t} \beta A(\xi)T_s(\xi)\,d\xi + (1-\lambda)\int_{-\sigma}^{t-\sigma} \beta A(\xi)T_s(\xi)\,d\xi, \tag{2.39}$$

which states that the number of infested trees at time t is the accumulated total of newly infested trees since time $t-\sigma$, plus the accumulated total over all times prior to $t-\sigma$ that escaped being cut down. Since we consider the case when $A(t) \to 0$, we can use (2.39) to find T_i^* in terms of T_s^* as follows:

$$\begin{aligned} T_i^* &= (1-\lambda)\int_{-\sigma}^{\infty} \beta A(\xi)T_s(\xi)\,d\xi \\ &= (1-\lambda)\left[\int_{-\sigma}^{0} \beta A^0(\xi)T_s^0(\xi)\,d\xi - \int_0^\infty \frac{dT_s(\xi)}{d\xi}\,d\xi\right], \end{aligned}$$

giving

$$T_i^* = (1-\lambda)\left[\int_{-\sigma}^{0} \beta A^0(\xi)T_s^0(\xi)\,d\xi + T_s(0) - T_s^*\right]. \tag{2.40}$$

Recall that $p^*(T_i^*) > 0$ has been defined such that $p = -p^*(T_i^*)$ is the singularity of $\bar{A}(p)$ of greatest real part. Therefore, $p^*(T_i^*)$ satisfies

$$-p^*(T_i^*) - e^{-\mu_L\tau}T_i^*B'(0)e^{p^*(T_i^*)\tau} + \mu_A = 0. \tag{2.41}$$

Equation (2.41) defines $p^*(T_i^*)$ as a function of T_i^*, and then (2.38) and (2.40) are solved simultaneously for T_s^* and T_i^*.

3. Numerical simulations. We have carried out some numerical simulations of our model for both the $\sigma < \tau$ and $\sigma > \tau$ situations. Fortunately, the alternative formulations of the two systems in terms of the functions $f(t)$ and $g(t)$ make the systems easily amenable to simulation using standard software tools for delay equations including those found within MATLAB.

For the $\sigma < \tau$ situation described in subsection 2.1, we simulate the system of three equations consisting of (2.1), (2.14), and (2.15). Note that one of the delays, σ, is not explicitly present in this system. However, σ plays a role through the formula (2.16), which is used to compute the initial data for the variable f from that for A and T_s. We measure time in months, so $\tau = 12$, corresponding to the maturation time of one year. We chose the birth rate function $B(A)$ to have the form $B(A) = b_m A e^{-aA}$, a common choice in the mathematical study of insect infestations, because it reflects the decreasing per capita egg laying rate due to crowding. In this formula the quantity b_m is the egg laying rate per female adult beetle per tree at lower densities without the effects of crowding. We assumed that a single female lays on average about 60 eggs during her life (though estimates vary considerably). She is active only in summer, but we average over a year to arrive at a figure of 5 eggs per female per month. Good data on the survival probability for the larvae are not available, but by being deep inside the trees the larvae are well protected from predators, so if we assume a survival probability of about 0.8, this leads to $\mu_L = 0.0186$. Values for other parameters are shown in the figures. Figures 1 and 2 illustrate two situations in which $\sigma < \tau$. In Figure 1 the infestation is eradicated, whereas in Figure 2 (in which we used a lower number for the probability λ of an infested tree being detected and removed) the number of susceptible trees tends to zero and the entire forest ends up infested, with the number of adult beetles tending to a constant. Other simulations showed a similar outcome but with the number of adults evolving to a periodic cycle. We also noted from our numerical experiments that (2.18) does not appear to be the best possible condition for eradication (i.e., it is sufficient but not necessary).

For the $\sigma > \tau$ situation of subsection 2.2, the appropriate system to simulate is that consisting of (2.21), (2.26), and (2.27). This system involves both delays σ and τ explicitly, with initial data for g calculated from that for A and T_s using (2.25) and the initial data formula for T_i in (2.12). Figure 3 shows a simulation for the $\sigma > \tau$ situation in which the time between tree infestation and tree removal is a little longer than the time taken for a larva to complete its development and mature as an adult beetle. One expects that it will be more difficult to achieve eradication. The simulation shows that it is still possible to do so, but only by detecting and removing 95% of infested trees. Figure 4 shows a situation with $\lambda = 1$, i.e., every infested tree is later destroyed, but with σ chosen to be considerably larger than τ, so that many larvae can complete their maturation and emerge as adults even though their host tree is doomed. With $\lambda = 1$, eradication is the final outcome (Theorem 3) even though $\sigma > \tau$, but the simulation shows a large and destructive transient, with very few susceptible trees remaining after the infestation has died out.

4. Discussion. We have derived a mathematical model to describe the infestation of wood boring beetles with the Asian longhorned beetle (ALB) as a prototype. Two delays are needed for the model. One of these delays, denoted by σ, is the average duration between the time a tree becomes infested and the time the infestation in the tree is detected and the tree removed. The second delay is the maturation delay τ for the beetle. Since the purpose is to examine whether or not the cut-burn

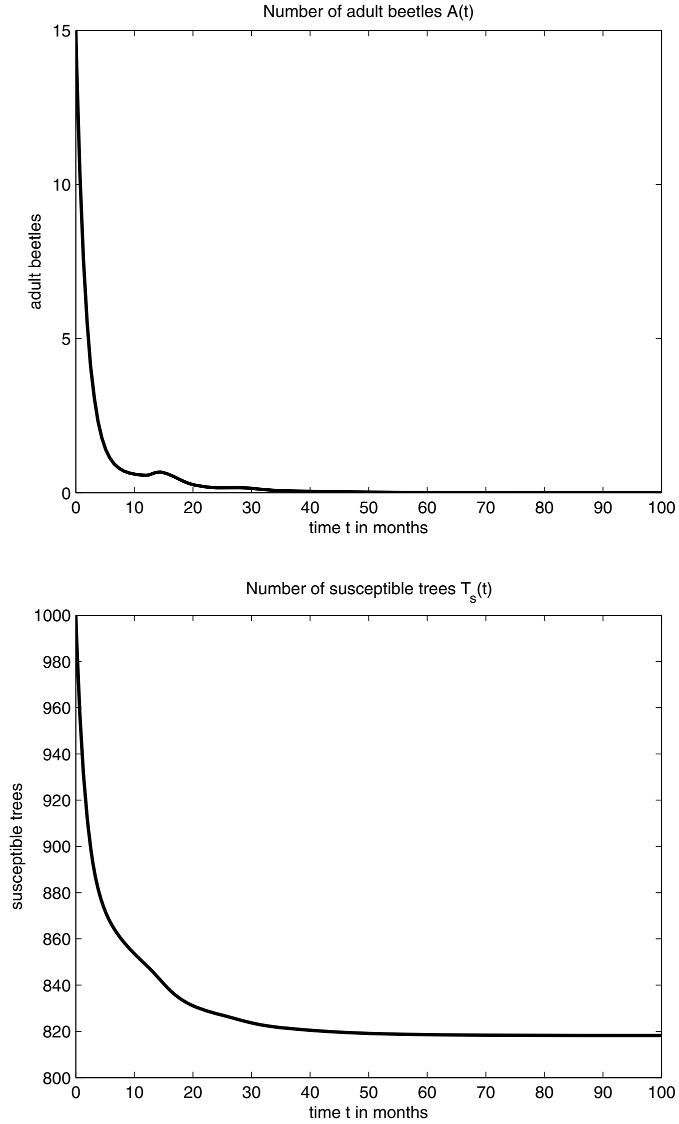


FIG. 1. *The* $\sigma < \tau$ *case: a simulation of* (2.1), (2.14), *and* (2.15). *Parameter values were* $\tau = 12$, $\lambda = 0.9$, $\beta = 0.005$, $\sigma = 3$, $b_m = 0.005$, $a = 0.1$, $\mu_L = 0.0186$, $\mu_A = 0.55$. *For these values,* (2.18) *is satisfied (by a fairly narrow margin) and the infestation is eradicated.*

(or removal) control strategy is successful, we have also incorporated a removal rate $\lambda \in [0, 1]$ into the model. The parameter λ accounts for the possibility that some infested trees might escape detection, which is very likely indeed if ALB activity were to reach wilderness areas, and also permits us to explore the possibility that infestation eradication might be possible without necessarily cutting down and burning every single infested tree. The model assumes different forms depending on whether $\sigma < \tau$ or $\sigma > \tau$, given by (2.1)–(2.4) and (2.21)–(2.24), respectively.

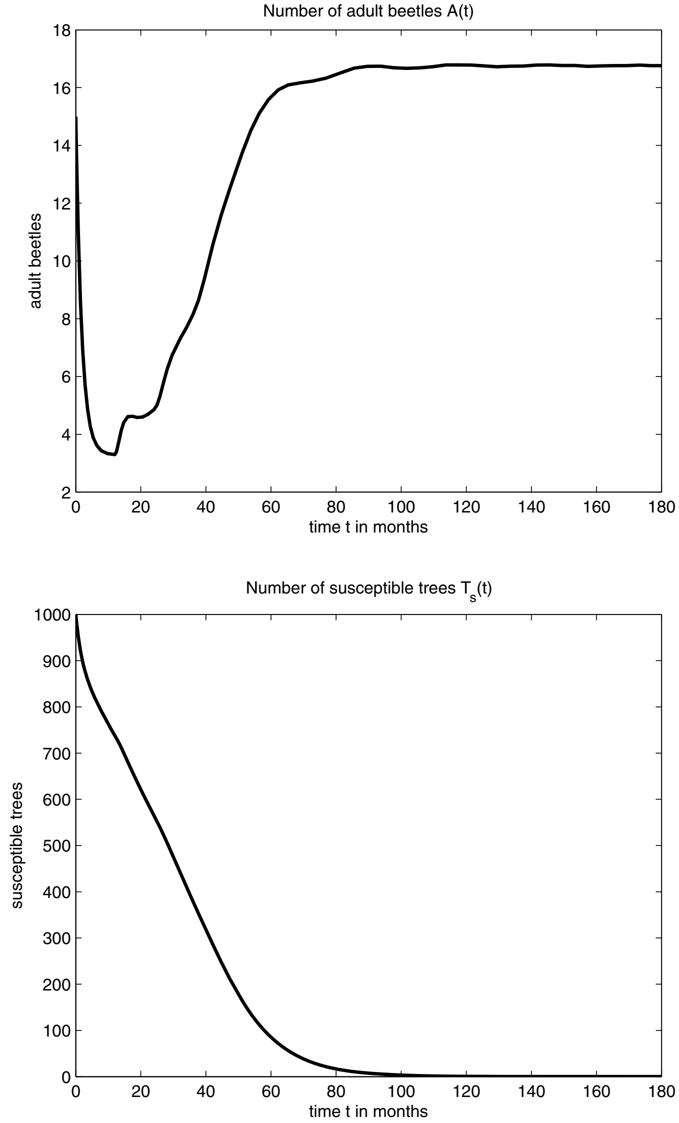


FIG. 2. *The $\sigma < \tau$ case: a simulation of* (2.1), (2.14), *and* (2.15). *Parameter values were $\tau = 12$, $\lambda = 0.4$, $\beta = 0.005$, $\sigma = 3$, $b_m = 0.005$, $a = 0.1$, $\mu_L = 0.0186$, $\mu_A = 0.55$. For these values the infestation takes over the whole forest and the number of adult beetles evolves to a constant.*

By applying the comparison method for delay differential equations, we have obtained some conditions for each of the two model systems that are sufficient for infestation eradication. We have also established lower bounds and even approximations for the remaining number of susceptible trees after eradication of the infestation, and this number has economic significance. We have also conducted some numerical simulations which confirm all the theoretical results.

If $\sigma < \tau$, then a beetle larva cannot complete its maturation in a host tree destined for detection and removal. The eradication condition for this case is (2.18), from which we conclude that if the proportion λ of attacked trees that are detected and removed is sufficiently close to 1, then the control strategy succeeds. Solving (2.18) for λ gives an explicit requirement on λ. In the case $\sigma > \tau$, detection of infestation in a tree is not happening quickly enough and it may be possible for a beetle larva to complete its maturation even in a tree destined for removal, especially if the larva hatched from an egg that was laid soon after the tree became infested. The corresponding model (2.21)–(2.24) is more difficult to analyze, but nevertheless a condition (namely, (2.31)) for eradication of the infestation can be obtained. However, it is more difficult to satisfy the condition. Indeed, if

$$e^{-\mu_L \tau} B'(0) \left[\beta \int_{-(\sigma-\tau)}^{0} A(\xi) T_s(\xi)\, d\xi + T_s(0) \right] < \mu_A, \tag{4.1}$$

then (2.31) will hold if $\lambda \in [0,1]$ is chosen sufficiently close to 1. But if (4.1) does not hold, that is,

$$e^{-\mu_L \tau} B'(0) \left[\beta \int_{-(\sigma-\tau)}^{0} A(\xi) T_s(\xi)\, d\xi + T_s(0) \right] \geq \mu_A, \tag{4.2}$$

then condition (2.31) is never satisfied regardless of the value of $\lambda \in [0,1]$, although we do know from Theorem 3 that the infestation is nevertheless eradicated if $\lambda = 1$. To understand the difference between conditions (2.18) and (2.31), note that condition (2.31) relates to the $\sigma > \tau$ situation, in which the timescale for infestation detection in a tree is longer than the maturation time for the beetle. Naturally, we should expect that infestation eradication should be more difficult in the $\sigma > \tau$ situation than in the $\sigma < \tau$ situation in which, if a tree is found to be infested, then its destruction happens sufficiently quickly so that a larva cannot mature in it. In the $\sigma > \tau$ situation it will be possible for some larvae to mature even in a host tree destined for removal.

Note that σ and λ are the only parameters in the model that are within our control. For example, σ could be decreased by the use of high technology acoustic detectors that can detect larval activity in a tree, and λ, which effectively measures the likelihood of ALB activity being detected in a tree, can be raised by increasing public awareness of the telltale signs of tree infestation. Another related beetle species in Japan has been controlled by the use of fungal bands which contain cultures of insect pathogenic fungi, and there have been trials of the technique on the ALB in Anhui, China (see Hajek et al. [7]). The fungal bands are placed at an approximate height of 2–2.5 m around the trees and infect the adult beetles, which can then transfer the infection during mating, leading to a reduced number of viable eggs. The effectiveness of the technique can be augmented by the use of a chemical attractant. On the use of entomopathogenic nematodes, see also Qin et al. [14].

None of the abovementioned measures is likely to be of much value if ALB were to take hold in wilderness areas. This, of course, highlights the importance of ensuring that the ALB does not become established in North America. Some studies of Keena [8] on the dependence of ALB activity on temperature suggests that the lower 48 states should be able to support beetle survival and reproduction. The numerical simulation work reported in this paper highlights the importance of rapid detection and removal of as many infested trees as possible, and moreover, the simulations demonstrate that if the detection of infestation timescale σ is significantly larger than

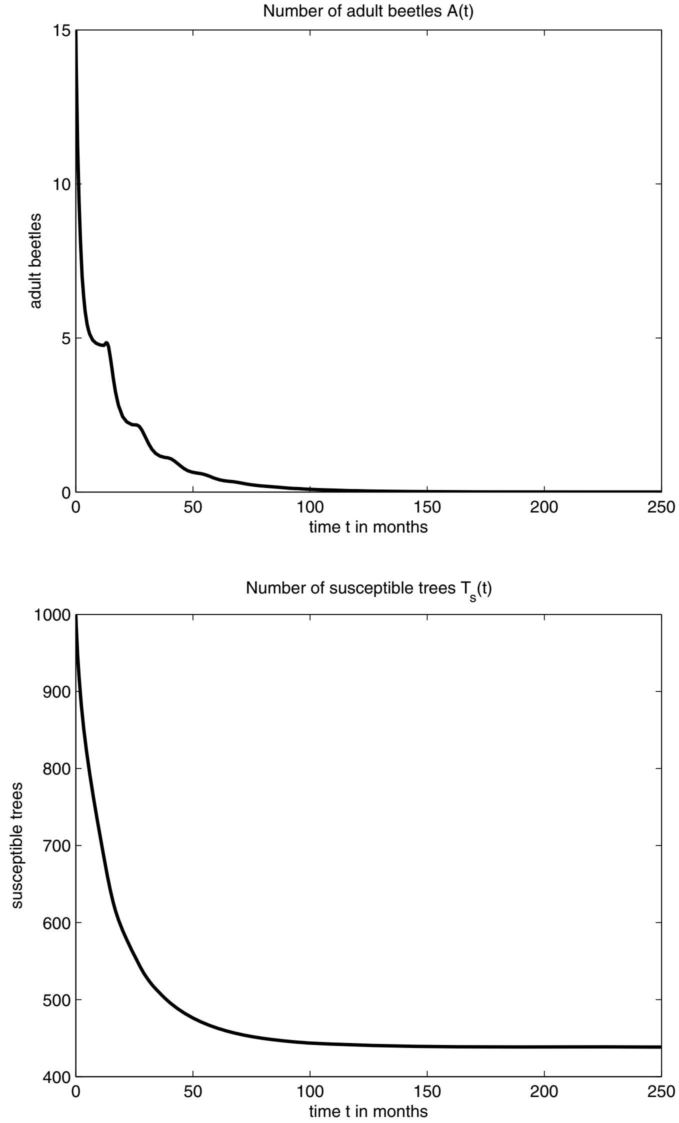


FIG. 3. *The $\sigma > \tau$ case: a simulation of* (2.21), (2.26), *and* (2.27). *Parameter values were $\tau = 12$, $\lambda = 0.95$, $\beta = 0.005$, $\sigma = 14$, $b_m = 0.005$, $a = 0.1$, $\mu_L = 0.0186$, $\mu_A = 0.55$. In this situation it is possible for a larva to complete its maturation in a tree destined for removal. Nevertheless it is possible to achieve eradication, but only with a large proportion of infested trees being detected and removed.*

the beetle maturation delay τ, then even though the infestation can be eradicated, the large transients will result in decimation of the forest.

We point out that although the eradication conditions (2.18) and (2.31) are obtained under the hypothesis $B(A) \leq B'(0)A$ for $A > 0$, this assumption is not crucial. Indeed, if $B(A)$ is continuously differentiable and $B(0) = 0$ (which always holds for all birth functions), then by the mean value theorem one can write

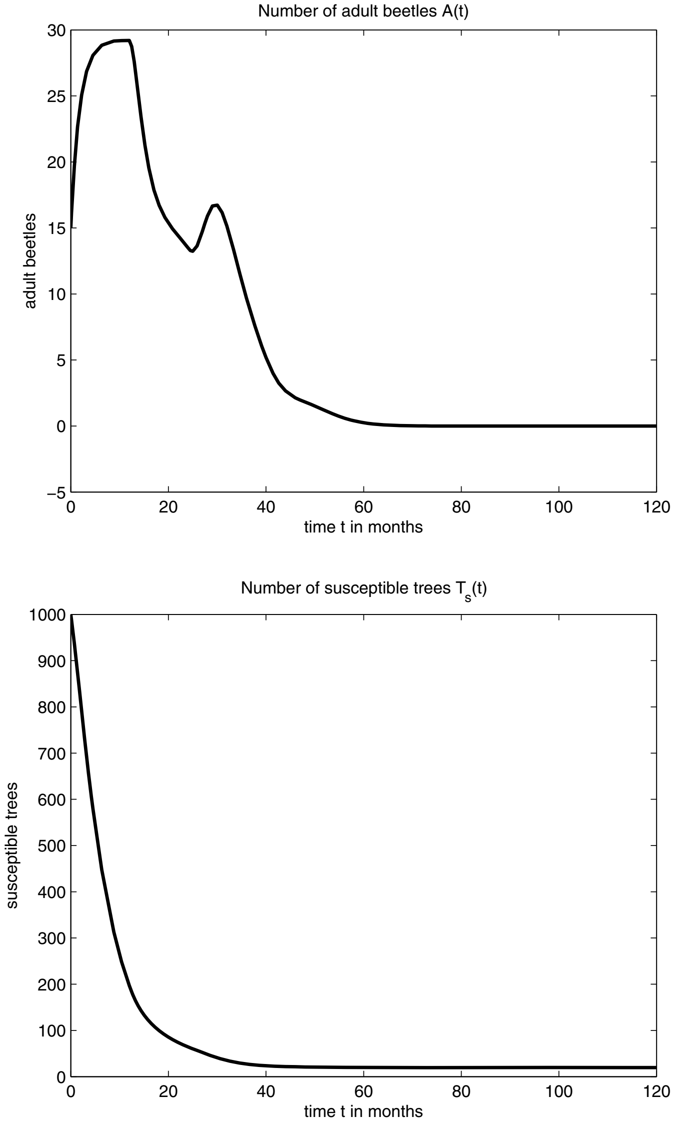


Fig. 4. *The $\sigma > \tau$ case: a simulation of* (2.21), (2.26), *and* (2.27). *Parameter values were $\tau = 12$, $\lambda = 1$, $\beta = 0.005$, $\sigma = 28$, $b_m = 0.005$, $a = 0.1$, $\mu_L = 0.0186$, $\mu_A = 0.55$. In this situation every tree that becomes infested is later destroyed, but with σ considerably larger than τ, many larvae can mature and escape as adults before their host tree is destroyed. Nevertheless the final outcome is still eradication, though only after a very destructive transient in which the infestation gets worse. Note that the final number of susceptible trees is very low, indicating severe destruction of the forest during the course of the infestation.*

$B(A) = B'(\theta)A \le B_m A$, where $B_m = \sup_{\theta \ge 0} B'(\theta)$. Therefore, the results in Theorems 1 and 2 remain true, with $B'(0)$ being replaced by B_m. Obviously the corresponding eradication conditions become more demanding on λ.

Spatial spread is another important issue that we have not considered here in this initial work. It is known that the adults of the ALB can fly, but only short distances between neighboring trees. Of course, the larvae do not disperse at all. The invasion of the ALB from Asia to North America is believed to have been via the use of wooden packing crates and wood products which may contain individuals in premature stages (eggs, larvae, or pupae). This no doubt accounts for the long range transport between continents and between different districts. Short range dispersal of adult beetles could possibly be modeled, as a first approximation, by the incorporation of Fickian diffusion. However, the actual dispersal behavior of the ALB is not so straightforward, and some experimental work reported in Bancroft and Smith [1] suggests a dependence on beetle density, weather conditions, beetle size and tree size, and that a thorough knowledge of dispersal behavior will be beneficial to eradication efforts. Release of the ALB is prohibited in the US, so there is little experimental work documented, but there have been detailed mark and release studies in Gansu Province, China (Smith et al. [17]) at a site chosen for its landscape similarities to those of urban infestations in the US. These studies suggest that dispersal depends on the spacing of suitable host trees, the age of the beetles, availability of host material, and crowding on suitable trees.

A great deal of mathematical work has been carried out on the dispersal behavior of beetles and insect species more generally (see Shigesada and Kawasaki [15], and Kot, Lewis, and van den Driessche [9]), but not specifically on the ALB. In fact, in our future work on the ALB we are considering the use of integrodifference equations that are continuous in space but discrete in time, since the ALB seems to follow a predictable pattern in nonnative habitats of having one generation per year. Discrete time models have been commonly used in the past to model beetle populations. The flour beetle *tribolium* has been particularly well studied due to its high rates of reproduction, short life cycle (4 to 6 weeks from egg to adult), ease of culture, and strong cannibalistic tendencies, which make the beetle suitable for laboratory studies (see Costantino and Desharnais [3]). These characteristics do imply that *tribolium* is very different from the ALB. Cannibalism in particular can affect all pre-adult stages of *tribolium*, including even some callow (not fully sclerotized) adults (see Mertz [10]).

Costantino et al. [4] compared the results of laboratory studies with the predictions of a discrete time model incorporating, unlike the present study, two pre-adult compartments. In the laboratory studies, adult mortality could be manipulated by removing or adding adults, and recruitment by removing or adding younger adults. Under conditions of high adult mortality, quasi-periodic cycles and chaos were observed in the laboratory populations and in the model predictions.

It seems unlikely that a continuous time model such as the one we present here could work satisfactorily for beetle species which have complex dynamics. Discrete time models might in principle be able to address the issue of complex dynamics, but these models have their limitations too. For example, western pine beetle populations can be asynchronous and indistinguishably overlapped because of differential brood development rates in different trees (see DeMars et al. [5]).

When it comes to modeling the eradication of infestation by removal strategy, we may have to consider diffusion of adults which depends on the removal strength (i.e., the parameter $\lambda \in [0,1]$), since when an infected tree is cut, the adult beetles in that tree will all fly to the neighboring trees. In other words, the removal of infested trees will enhance the diffusion of the adults. This will make modeling a more interesting yet more challenging job. We leave such problems for future investigation.

Acknowledgments. We are grateful to the referees for their very useful comments and suggestions.

REFERENCES

[1] J. S. Bancroft and M. T. Smith, *Dispersal and influences on movement for Anoplophora glabripennis calculated from individual mark-recapture*, Entomologia Experimentalis et Applicata, 116 (2005), pp. 83–92.

[2] C. Castillo-Chavez and H. R. Thieme, *Asymptotically autonomous epidemic models*, in Mathematical Population Dynamics: Analysis of Heterogeneity, I. Theory of Epidemics, O. Arino et al., eds., Wuerz, Canada, 1995, pp. 33–50.

[3] R. F. Costantino and R. A. Desharnais, *Population Dynamics and the Tribolium Model: Genetics and Demography.* Springer-Verlag, New York, 1991.

[4] R. F. Costantino, R. A. Desharnais, J. M. Cushing, and B. Dennis, *Chaotic dynamics in an insect population*, Science, 275 (1997), pp. 389–391.

[5] C. J. DeMars, G. W. Slaughter, W. D. Bedard, N. X. Norick, and B. Roettgering, *Estimating western pine beetle-caused tree mortality for evaluating an attractive pheromone treatment,* J. Chem. Ecol., 6 (1980), pp. 853–866.

[6] R. Gao, X. Qin, D. Chen, and W. Chen, *A study on the damage of poplar caused by Anoplophora glabripennis*, Forest Research, 6 (1993), pp. 189–193.

[7] A. E. Hajek, B. Huang, T. Dubois, M. T. Smith, and Z. Li, *Field studies of control of Anoplophora glabripennis (Coleoptera: Cerambycidae) using fiber bands containing the entomopathogenic fungi Metarhizium anisopliae and Beauveria brongniartii*, Biocontrol Science and Technology, 16 (2006), pp. 329–343.

[8] M. A. Keena, *Effects of temperature on Anoplophora glabripennis (Coleoptera: Cerambycidae) adult survival, reproduction, and egg hatch*, Environ. Entomol., 35 (2006), pp. 912–921.

[9] M. Kot, M. A. Lewis, and P. van den Driessche, *Dispersal data and the spread of invading organisms*, Ecology, 77 (1996), pp. 2027–2042.

[10] D. B. Mertz, *The tribolium model and the mathematics of population growth*, Ann. Rev. Ecol. Syst., 3 (1972), pp. 51–78.

[11] W. D. Morewood, P. R. Neiner, J. C. Sellmer, and K. Hoover, *Behavior of adult Anoplophora glabripennis on different tree species under greenhouse condition*, Journal of Insect Behavior, 17 (2004), pp. 215–226.

[12] W. D. Morewood, K. Hoover, P. R. Neiner, J. R. McNeil, and J. C. Sellmer, *Host tree resistance against the polyphagous wood-boring beetle Anoplophora glabripennis*, Entomologia Experimentalis et Applicata, 110 (2004), pp. 79–86.

[13] D. J. Nowak, J. E. Pasek, R. A. Sequeira, D. E. Crane, and V. C. Mastro, *Potential effect of Anoplophora glabripennis (Coleoptera: Cerambycidae) on urban trees in the United States*, Journal of Economic Entomology, 94 (2001), pp. 116–122.

[14] X. Qin, R. Gao, H. Yang, and G. Zhang, *Study on the application of entomopathogenic nematodes, Steinernema bibionis and S. feltiae, to control Anoplophora glabripennis and Holocercus insularis*, Forest Science and Research, 1 (1988), pp. 179–185.

[15] N. Shigesada and K. Kawasaki, *Biological Invasions: Theory and Practice*, Oxford University Press, Oxford, UK, 1997.

[16] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, American Mathematical Society, Providence, RI, 1995.

[17] M. T. Smith, P. C. Tobin, J. Bancroft, G. Li, and R. Gao, *Dispersal and spatiotemporal dynamics of asian longhorned beetle (Coleoptera: Cerambycidae) in China,* Environ. Entom., 33 (2004), pp. 435–442.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1688–1716

THE FREDERIKS EFFECT AND RELATED PHENOMENA IN FERRONEMATIC MATERIALS*

V. I. ZADOROZHNII†, T. J. SLUCKIN‡, V. YU. RESHETNYAK†, AND K. S. THOMAS§

Abstract. Using continuum and statistical mechanical theories, we study the switching properties of a ferronematic in a nematic liquid crystal cell subject to homeotropic boundary conditions at the cell and particle walls. An external magnetic field normal to the cell plane is also imposed. At low fields we find thresholdless switching of the nematic director, consistent with experimental data. At higher fields, there are three regimes, depending on the strength of the anchoring interaction between the director and the ferroparticle orientation. For low anchoring strengths, there is an inverse Frederiks effect, and the nematic reorientation reduces and then disappears continuously at a critical magnetic field. At intermediate fields, the degree of reorientation reduces at high fields but remains finite. For high fields, however, the director switching saturates. The dimensionless temperature scale in the problem involves the temperature, the mean nematic elastic constant, the colloidal density, and the cell dimension. If this quantity is sufficiently low, then high magnetic fields can cause magnetic segregation. The segregation order parameter is coupled to the director distortion, and this can change the inverse Frederiks transition into a first order transition, leading to bistability in an intermediate field regime. These features are perturbed but not changed structurally by the effect of a small bias magnetic field (< 10 Oe) normal to the unperturbed director. Subject to suitable choice of parameters, the theory is also quantitatively consistent with the results of the classic experiment of Chen and Amer in 1983.

Key words. liquid crystals, ferronematics, colloids

AMS subject classifications. 82B26, 82D30, 82D45

DOI. 10.1137/070703831

1. Introduction.

In 1970 Brochard and de Gennes [1] suggested on theoretical grounds that it might be possible to construct magnetic colloids based on a liquid crystal matrix. As a result of the anchoring at the surface of the colloidal particles, magnetic and nematic order in these materials would be coupled. The weak interaction between magnetic fields and nematic order, already exploited by the early liquid crystal pioneers such as Charles Mauguin [2], would then be dramatically increased.

The giant magnetic-nematic coupling might then be fruitfully used in devices controlled by easily accessible magnetic fields (< 10 Oe). These systems, including as they do elements of both ferromagnetism and nematic liquid crystalline behavior, have come to be known as *ferronematics* (FNs). The key relevant properties of a nematic crystal doped with single-domain ferromagnetic particles are (a) the intrinsic high magnetic susceptibility and (b) the uniform molecular reorientation of the entire liquid crystal (LC) matrix, or macroscopic collective behavior [1], in a varying magnetic field.

It was not until 1983 that Chen and Amer [3] were first able to construct a model experimental system. One feature complicating the interpretation of experiments is

*Received by the editors September 26, 2007; accepted for publication (in revised form) March 28, 2008; published electronically July 2, 2008. INTAS (grant 99-00312) and NATO (grant CBP.NUKR.CLG.981968) have supported visits to Southampton by VYR and VIZ and visits to Kiev by TJS. We thank both these agencies for their continuing support.

http://www.siam.org/journals/siap/68-6/70383.html

†Physics Faculty, Kyiv National Taras Shevchenko University, Prosp. Glushkova, 2, bldg 1, Kyiv, 03680, Ukraine (viza@mail.univ.kiev.ua, reshet@iop.kiev.ua).

‡School of Mathematics, University of Southampton, Southampton, SO17 1BJ, United Kingdom (t.j.sluckin@soton.ac.uk).

§School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom (K.S.Thomas1@soton.ac.uk).

the fact that the ferroparticles can coagulate irreversibly. Indeed, subsequent work has continually been hindered for a long time by difficulties in the manufacture of stable reproducible magnetic liquid crystalline colloids.

In the last two decades interest in these materials has grown. Not only do these systems pose interesting physical problems, but also they promise to provide an optical device technology based on magnetic switching [4, 5, 6, 7, 8]. In particular, Buluy et al. [9] have recently synthesized an FN which is stable against colloidal aggregation. This system consists of magnetite particles (Fe_3O_4) coated with oxyethyl-propylene glycol and suspended in 5CB.

The initial continuum theory [1] has been generalized by Burylov and Raikher [10, 11] to the case of a finite anchoring energy of the nematic at the ferroparticle surface. This provided an explanation for the absence of coalignment of the nematic director $\hat{\mathbf{n}}$ and the averaged local magnetization $\mathbf{M}$ in the FN, as had been found experimentally [12]. Experiments, among them one on 8CB-based FN with magnetite particles of nearly spherical shape [6], have subsequently confirmed the generalized theory. The loss of coalignment has been labeled *detachment* in the literature. In this paper we determine some more precise conditions for detachment.

One other important feature which the theory must include is the possibility of *magnetic segregation* [1, 13]. This involves the magnetic colloidal particles migrating toward regions in which the ferronematic coupling energy is minimized. Typically this occurs in the center of the sample, where the nematic director is most free to rotate toward the magnetically favored direction.

However, this tendency to segregate is opposed by entropic forces favoring a uniform colloidal concentration. The balance is subtle and is controlled by the dimensionless temperature, which is typically of order unity. Previous calculations [14] have suggested that this effect can lead to hysteretic effects as a function of magnetic field. The hysteretic effects have been interpreted as the signature of a decoupling of the nematic director and the ferroparticle orientation. In this paper we give a full account of the physics of the ordering process which occurs in a confined ferronematic system within the usual Frederiks geometry.

The plan of the paper is as follows. In section 2 we describe the theoretical model. In section 3 we give an analysis of the basic aspects of the model, specifically confining our interest to the case in which magnetic segregation is absent. Most of this work is analytical, but we conclude this section by presenting the results of some computational solutions of the equations. We also compare the numerical solutions and the analytical approximations; the analysis, even where only approximate, yields surprisingly accurate precisions. In section 4 we extend the model to include the so-called bias field. This is the extra in-plane magnetic field required to stabilize experimentally the systems we are discussing. In section 5 we add a discussion of magnetic segregation. Both bias field and magnetic segregation are important experimentally but might be regarded as complications to the basic underlying mathematical model. In section 6 we give a brief discussion linking our work to some of the available experiments. Finally, in section 7 we draw some conclusions and place our work in a larger context. A brief preliminary report of our work has been published elsewhere [15].

2. Model.

The basic geometry of the problem is shown in Figure 1. The ferroparticles are needle-like monodomain ferrite grains of length L and diameter $d \sim L/7$–$L/3$; they are significantly larger than the nematic molecule size. We consider a cell of thickness D and suppose strong homeotropic nematic anchoring at the cell walls at $z = 0$ and $z = D$.

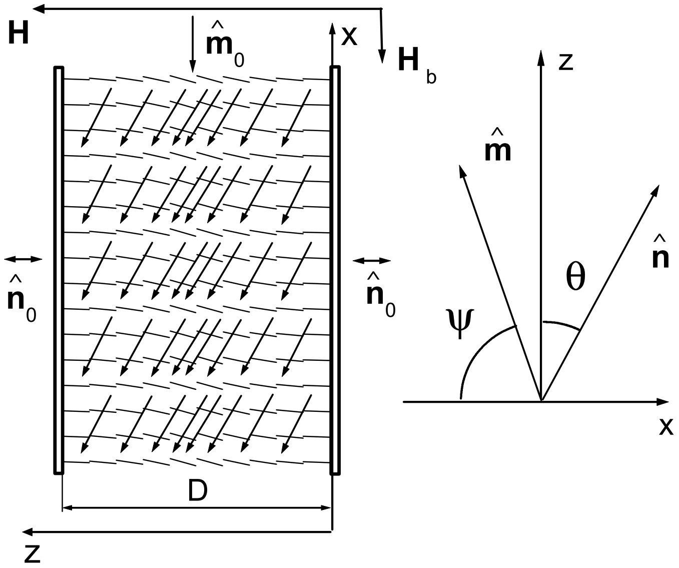


FIG. 1. *Schematic of the FN cell. The total magnetic field is given by* $\mathbf{H}_s = [H_b, 0, H]$, *with* H_b *the bias field and* H *the imposed field. In the left-hand diagram, the magnetic director is marked with an arrow, while the nematic director is marked by a line. The meaning of the angles* ψ, *signifying the deviation of the magnetic director from its original orientation, and* θ, *signifying the nematic deviation, are shown in the right-hand diagram.*

Burylov and Raikher [10, 11] have derived an effective soft homeotropic anchoring energy W_p per colloidal particle due to the nematic-ferroparticle surface interaction. We do not discuss the microscopic derivation of the magnitude of this interaction here but simply note that in our discussion it is a measurable parameter which enters the theory. The existence of this anchoring encourages, but does not oblige, the ferroparticles to orient perpendicular to the (averaged) nematic director. This term changes the physics of the classical nematic liquid crystal.

In addition, the cell may be initially subject to a small *bias* magnetic field $\mathbf{H}_b$ parallel to the cell substrates. In this initial state ferroparticles are homogeneously distributed over the cell volume and their magnetic moments are perpendicular to the unperturbed director $\hat{\mathbf{n}}_0$. $\hat{\mathbf{m}}_0$ is the unit vector in the direction of the sample magnetization at $\mathbf{H} = 0$. The sample is uniform in the x-y direction.

The physical states are defined by a nematic director $\hat{\mathbf{n}}(\mathbf{r})$, a local normalized magnetization $\hat{\mathbf{m}}(\mathbf{r})$, and a local colloidal particle density $\rho(\mathbf{r}) = \frac{f(\mathbf{r})}{\nu}$, with f the local colloidal packing fraction and ν the colloidal particle volume.

The FN equilibrium state at given magnetic field $\mathbf{H}_s = \mathbf{H} + \mathbf{H}_b$ can be determined by minimizing the full free energy functional [10]

$$\begin{aligned}\mathcal{F} = \int_V \Bigg\{ &\frac{1}{2}\left[K_1(\nabla\cdot\hat{\mathbf{n}})^2 + K_2(\hat{\mathbf{n}}\cdot\nabla\times\hat{\mathbf{n}})^2\right.\\ &\left.+ K_3(\hat{\mathbf{n}}\times\nabla\times\hat{\mathbf{n}})^2\right] - \frac{1}{2}\chi_a(\hat{\mathbf{n}}\cdot\mathbf{H}_s)^2 + \frac{fk_BT}{\nu}\ln f\\ &- \bar{M}f(\hat{\mathbf{m}}\cdot\mathbf{H}_s) + fW_p(\hat{\mathbf{n}}\cdot\hat{\mathbf{m}})^2\Bigg\}\,dV \end{aligned} \tag{1a}$$

subject to ferroparticle number conservation

$$\int f\,dV = \bar{f}V, \tag{1b}$$

where K_1, K_2, K_3 are elastic (Frank) constants, χ_a is the anisotropic part of the nematic diamagnetic susceptibility, f and $\bar{f}$ are the local and the mean volume particle fractions, and $\hat{\mathbf{m}}$ is the unit vector in the direction of the sample magnetization $\mathbf{M} = \bar{M}f\hat{\mathbf{m}}$, where normally $\bar{M} = M_s$, the saturation magnetization per unit volume within an individual colloidal particle. In the limit of low applied fields, we shall find, however, that sometimes $\bar{M}$ is reduced from this saturation value by thermal fluctuation effects.

The term in square brackets in (1a) is the Frank–Oseen–Zocher curvature energy. The remaining terms are, respectively, the magnetic energy density, the contribution of the mixing entropy of an ideal ferroparticle solution, the magnetic energy of the colloidal particles, and the anchoring-induced ferronematic interaction [10]. We ignore the ferroparticle magnetic dipole-dipole interaction energy, which disappears at low ferroparticle concentrations. It will be useful to define the quantity $\eta(\mathbf{r}) = \frac{f(\mathbf{r})}{\bar{f}}$. This is the local enhancement (or reduction) of the colloidal density induced by segregation effects. We also suppose the magnetization of the ferroparticles to have reached saturation [10].

In this geometry, the nematic and magnetic distortions are given, respectively, by $\hat{\mathbf{n}} = (\sin\theta, 0, \cos\theta)$ and $\hat{\mathbf{m}} = (-\cos\psi, 0, \sin\psi)$, where $\theta = \theta(z)$ and $\psi = \psi(z)$ are shown in Figure 1. In the simple case when the elastic constants K_1 and K_3 are taken to be equal ($K_1 = K_3 = K$), the free energy functional reduces to

$$\begin{aligned}\mathcal{F} = \int_0^D \Bigg[&\frac{1}{2}K\left(\frac{d\theta}{dz}\right)^2 - \frac{1}{2}\chi_a(H\cos\theta - H_b\sin\theta)^2 \\ &+ \eta(z)\frac{\bar{f}k_BT}{\nu}\ln\eta - \bar{M}\eta(z)\bar{f}(H\sin\psi + H_b\cos\psi) \\ &+ \eta(z)\bar{f}W_p\sin^2(\theta-\psi)\Bigg]\,dz,\end{aligned} \tag{2}$$

subject to the boundary condition $\theta(0) = \theta(D) = 0$ and the constraint $\frac{1}{D}\int_0^D \eta(z)\,dz = 1$.

We may sensibly enquire what conditions are required for (1a) and (2) to hold. This point has been addressed briefly by Burylov and Raikher [10, 11]. However, a detailed discussion of this point goes beyond the scope of this paper. The key point is that it should be possible to define the local magnetic director $\hat{\mathbf{m}}$ as a good macroscopic variable in a continuum theory. This in turn requires that there be effective coupling of the local magnetic directors over length scales larger than the typical interparticle distance. We return to this point briefly in section 6, when we discuss the application of our work to experimental interpretation.

The effective coupling between the magnetic and nematic director will necessarily involve some orientational distortions induced by the colloidal particles over a length scale which may be considerably larger than the individual particle. A sufficient condition for the existence of the director $\hat{\mathbf{m}}$ will be that this distortion region be larger than the typical interparticle distance, so that the distortions induced by neighboring colloidal particles overlap. There will then be an indirect interaction between the magnetic particles, and a locally macroscopic structure. This condition has been termed the requirement for collective behavior [1, 10, 16].

To simplify we initially suppose the following:

(a) There is no bias field. The *raison d'être* of the bias field is to hold $\hat{\mathbf{m}}$ and hence $\hat{\mathbf{n}}$ in the x-z plane. In the absence of the bias field at very low applied fields $\bar{M}/M_s < 1$, but here we shall suppose that this is not the case. We shall discuss the low bias field case (and criteria for determining what is meant by this limit) at the end of the paper.

(b) The direct magnetic-nematic interaction as a result of the anisotropic nematic molecular susceptibility can be ignored as compared to the indirect (but giant) colloidally mediated coupling. Normally, the anisotropic part of the nematic diamagnetic susceptibility is extremely low, for 5CB $\chi_a = 1.7 \times 10^{-7}$ [17]. Thus neglecting the bare magnetic-nematic interaction is a reasonable approximation for the magnetic fields which we shall consider ($H < 200$ Oe), as we will see below.

We now nondimensionalize the problem. Length scales are now measured in units of the cell width D (i.e., z in scaled units is equal to z/D in unscaled units). The scaled free energy is now given by the following formula:

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 + \eta t \ln \eta - \eta h \sin\psi \right. \\ \left. + \eta w \sin^2(\theta - \psi) \right], \tag{3}$$

subject to the constraint $\int_0^1 \eta(z)\, dz = 1$, and with

(a) $h = \bar{f}\bar{M}HD^2/K$ the dimensionless magnetic field;

(b) $w = \bar{f}W_pD^2/K$ the dimensionless coupling (W_p now per unit volume) between the nematic and magnetic orientations; we refer to this quantity as the *ferronematic coupling parameter*; and

(c) $t = k_BT\bar{f}D^2/(\nu K)$ the dimensionless temperature; this is roughly the ratio of the thermal energy of the colloidal particles to the nematic elastic free energy density.

We note that this nondimensionalization scheme differs from that in a number of previous papers [10, 11, 13, 14]. In the previous normalization scheme, it turns out not to be possible to remove the magnetic segregation in a regular way. By contrast, the scheme introduced in our previous letter [15] and further developed here permits a simple limit in which there is no magnetic segregation. This corresponds to an infinite temperature limit with respect to the energy parameter driving the segregation. Segregation can then be considered as a finite temperature perturbation. The natural perturbation parameter is then a nondimensionalized inverse temperature.

Suitable surrogates for the global behaviors of the parameters will be the quantities $\theta_0(h) = \theta\left(h, z = \frac{1}{2}\right)$ and the analogous quantity $\psi_0(h) = \psi\left(h, z = \frac{1}{2}\right)$. We shall also define the degree of segregation in a number of ways. The segregation order parameter $s(h)$ is an integral quantity and is defined by

$$s = -\int_0^1 dz\, \eta(z) \cos 2\pi z. \tag{4}$$

The quantity

$$\eta_0(h) = \eta\left(h, z = \frac{1}{2}\right) \tag{5}$$

is in some respects analogous to $\theta_0(h)$ and $\psi_0(h)$ and measures the ratio of the concentration in the center of the cell to its average value.

3. Unsegregated limit. To begin with we put $t \to \infty$ [18]. The colloidal concentration responds in a Boltzmann-like fashion to its local potential energy. In the infinite temperature limit, therefore, there is no response; the concentration $\eta(z)$ remains constant, and there is no colloidal segregation. Equation (3) and the resulting field theory are both now much simplified.

The theory now reduces to minimizing

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - h\sin\psi + w\sin^2(\theta - \psi)\right]. \tag{6}$$

3.1. Infinitely strong ferronematic coupling. Now $w \to \infty$, which enforces $\theta = \psi$; the nematic director follows the magnetic particle distortion and is always perpendicular to the magnetic particles. The resulting problem is now similar to the classical Frederiks problem but lacks the symmetry-breaking characteristic of this problem. The theory now reduces to

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - h\sin\theta\right]. \tag{7}$$

The Euler–Lagrange equation is

$$\frac{d^2\theta}{dz^2} + h\cos\theta = 0. \tag{8}$$

The *weak field* solution can be calculated by putting $\cos\theta = 1$ in this equation. Substituting $\theta(0) = \theta(1) = 0$ then yields

$$\theta(z) = \frac{h}{2}\, z\,(1 - z). \tag{9}$$

One suitable figure of merit for the total degree of reorientation is $\Theta = \int_0^1 \theta(z)\, dz$. At low fields $\Theta = h/12$; the effect is proportional to the imposed field and is thresholdless. Alternatively, we may choose $\theta_0(h) = \theta\left(h, z = \frac{1}{2}\right)$. From (9), we find $\theta_0(h) = \frac{h}{8}$. Later in this paper we shall find it useful to expand $\theta(h, z) \approx \theta_0(h)\sin\pi z$ in a harmonic approximation. If we minimize the linearized version of (7) within this approximation, we obtain $\theta_0(h) \approx (4/\pi^3)h \approx 0.129h$, rather close to the $\theta_0(h) = 0.125h$, given by (9).

A *strong field* solution can be found using a matched asymptotic expansion method [19]. A high field is here defined by $h \geq 1$ in the absence of any other scale on which to compare it. We omit the details, as only the conclusion is important for the subsequent discussion. There is a boundary region close to the wall, in which the distortion is small but rapidly increases:

$$\theta(z) \approx 1.46h^{1/2}z - 0.5\left(h^{1/2}z\right)^2. \tag{10}$$

This region has thickness $z_0 = 0.46h^{-1/2}$. In the bulk of the cell, the solution for θ is approximately given by

$$\theta(z) = \frac{\pi}{2} - 3.16\exp\left[-\frac{h^{1/2}}{2}\right]\cosh\left[h^{1/2}\left(z - \frac{1}{2}\right)\right]. \tag{11}$$

The key result is that the director distortion saturates. There are boundary regions close to the walls in which the distortion reduces to zero of thickness $z_0 \sim h^{-1/2}$, which become increasingly thin with increasing field. The value of the distortion in the center of the cell becomes increasingly close to saturation, with

$$\theta_0(h) = \frac{\pi}{2} - 3.16 \exp\left[-\frac{h^{1/2}}{2}\right] \tag{12}$$

and

$$\frac{\pi}{2} - \Theta(h) \sim h^{-1/2}. \tag{13}$$

We note that all figures of merit for the total degree of reorientation, including the optical phase lag measured through the cell, *monotonically increase* with magnetic field up to saturation, although this increase does of course slow at high fields.

These results do not, however, carry over exactly into the finite coupling case. This is because the degree to which the saturated alignment of the magnetic director is limited by the degree of coupling between the nematic and the magnetic directors, as we shall see below.

3.2. Finite ferronematic coupling. This corresponds to the more realistic case of finite anchoring at the colloidal particle surface. The appropriate free energy functional is given by (6):

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - h \sin\psi + w \sin^2(\theta - \psi)\right]. \tag{6}$$

The Euler–Lagrange equations corresponding to the free energy (6) are

$$\frac{d^2\theta}{dz^2} - w \sin(2(\theta - \psi)) = 0, \tag{14a}$$

$$h \cos\psi + w \sin(2(\theta - \psi)) = 0. \tag{14b}$$

Equation (14b) is the *bonding equation* of Burylov and Raikher [10, 11]. It is apparent that so long as the ratio $\frac{h}{w} \ll 1$, we can expect that $\theta \approx \psi$. The magnetic and nematic directors will be coaligned (or bonded) in this circumstance. If this condition does not hold, then $\theta \approx \psi$ may still hold, but only if $\cos\psi$ is sufficiently small (i.e., ψ is sufficiently close to $\frac{\pi}{2}$). Much of this paper is concerned with determining the details of when and in what way this relation holds. It ceases to hold with increasing field when h and w are of the same order of magnitude. The magnetic director becomes more aligned with the field, but the nematic director is no longer aligned with it. This phenomenon has been described as *decoupling* by Burylov et al. [13].

3.2.1. Boundary conditions. From a mathematical point of view these equations are slightly peculiar. The strong surface anchoring condition on θ yields $\theta(0) = \theta(1) = \theta_s = 0$. However, the boundary conditions $\psi(0) = \psi(1) = \psi_s$ are not defined explicitly. However, they are implicit in (14b) from the surface condition $\theta = 0$. This yields

$$h \cos\psi_s = w \sin(2\psi_s), \tag{15}$$

which sustains two solutions:

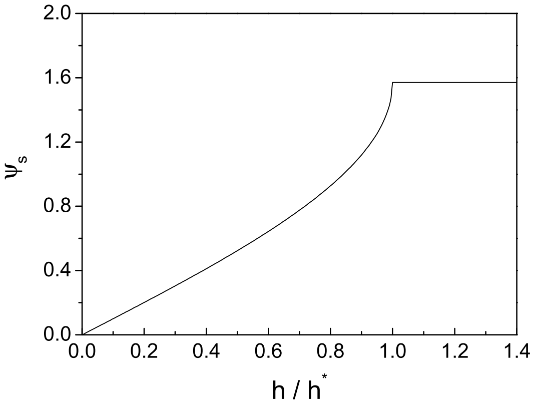


FIG. 2. *Behavior of the displacement of the magnetic director at the surface ψ_s, as a function of scaled magnetic field h. Note the linear behavior at low h, and the sharp singularity at $h = h^*$, at which the ψ_s abruptly saturates.*

(a) $\psi_s = \arcsin(h/2w)$, which is the low magnetic field case. For $h/2w > 1$, however, $\psi_s = \arcsin(h/2w)$ is no longer well defined.

(b) $\psi_s = \pi/2$, which is the high magnetic field case. The surface value of the magnetic angle ψ_s *saturates*.

Thus $\psi_s(h)$ is a monotonic function of h, increasing smoothly from zero at $h = 0$. For low h, $\psi_s \sim h$; the behavior is linear. But there is a change of regime at $h = h^*(w) = 2w$, at which point $\psi_s = \psi(0) = \psi(1)$ saturates. The behavior of ψ_s is shown in Figure 2. The quantity ψ_s saturates abruptly at $h = h^*$. There is strongly discontinuous behavior in the gradient $\frac{d\psi_s}{dh}$ at $h = h^*$. Just below $h = h^*$, we have

$$\psi_s \sim \frac{\pi}{2} - \sqrt{2}\left(1 - \frac{h}{h^*}\right)^{1/2}. \tag{16}$$

Given the coupling between ψ and θ, one might also expect some nonanalytic properties in $\psi(z)$ (and hence in $\theta(z)$) for all z as a function of h. Numerical evidence suggests, however, that the singularity is rounded away from the boundary. We shall return to this problem elsewhere.

3.2.2. Low field properties. We can now expand (6) in powers of small quantities h, θ_0, and γ_0, where

$$\theta(z) \approx \theta_0 \sin \pi z; \quad \psi(z) = \psi_s + \gamma_0 \sin \pi z; \quad \psi_s = \arcsin\left(\frac{h}{2w}\right). \tag{17}$$

The free energy is then given by

$$\begin{aligned} F = \int_0^1 \Bigg\{ & \frac{\pi^2}{2}\,\theta_0^2 \cos^2 \pi z - h \sin\left[\arcsin\frac{h}{2w} + \gamma_0 \sin \pi z\right] \\ & + w \sin^2\left[\arcsin\frac{h}{2w} + (\gamma_0 - \theta_0)\sin \pi z\right]\Bigg\}\, dz. \end{aligned} \tag{18}$$

We now expand the integrand of (18) in a multivariate Taylor series up to second order in θ_0 and γ_0, noting that $\sin\left(\arccos\frac{h}{2w}\right) = \cos\left(\arcsin\frac{h}{2w}\right) = \sqrt{1 - \frac{h^2}{4w^2}}$. Performing the integrals, including those over $\sin \pi z$ and $\sin^2 \pi z$, we obtain the following expansion in θ_0 and γ_0:

(19)

$$F = -\frac{h^2}{4w} - \frac{2h}{\pi}\sqrt{1-\frac{h^2}{4w^2}}\,\theta_0 + \frac{\pi^2}{4}\theta_0^2 + \frac{h^2}{8w}\gamma_0^2 + \frac{w}{2}\left(1-\frac{h^2}{2w^2}\right)(\theta_0-\gamma_0)^2 + O(\theta_0^3,\gamma_0^3).$$

This leads to Lagrange equations

$$\text{(20a)}\qquad \frac{\pi^2}{2}\theta_0 + w\left(1-\frac{h^2}{2w^2}\right)(\theta_0-\gamma_0) - \frac{2h}{\pi}\sqrt{1-\frac{h^2}{4w^2}} = 0,$$

$$\text{(20b)}\qquad \frac{h^2}{4w}\gamma_0 - w\left(1-\frac{h^2}{2w^2}\right)(\theta_0-\gamma_0) = 0.$$

Solving these we find

(21)

$$\theta_0 = \frac{16hw^3}{\pi(4w^3\pi^2+2w^2h^2-\pi^2h^2w-h^4)}\left(1-\frac{h^2}{4w^2}\right)^{3/2} = \frac{4h}{\pi^3} - \frac{\pi^2+4w}{2\pi^5w^2}h^3 + O(h^5)$$

and

$$\text{(22)}\qquad \theta_0-\gamma_0 = \frac{h^3}{\pi^3w^2} + O(h^5).$$

In the limit $\frac{h}{w}\to 0$, as expected, this solution tends to the behavior of the infinitely strong-coupling regime (9). In the very low field regime, the response is as though the torque is acting directly on the nematic director. But this response is modified by a third order term in h. This reduces the response as compared to the infinite coupling case, as one might expect, given that the restoring force on the nematic director acts only through the intermediary effect of the magnetic director.

3.2.3. Infinite field limiting properties. For sufficiently high fields, the minimizer of the free energy (6) with respect to ψ requires that $\psi(z) = \frac{\pi}{2}$ everywhere. Now (6) reduces to

$$\text{(23)}\qquad F \sim \int_0^1 dz\left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 + w\sin^2\left(\theta-\frac{\pi}{2}\right)\right] \sim \int_0^1 dz\left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - w\sin^2\theta\right].$$

This is just the classic Frederiks transition problem, whose solution is well known [20].

The nematic distortion now no longer saturates at high fields. Indeed, for $w \le w_c = \frac{\pi^2}{2}$, $\psi(z) = \frac{\pi}{2}$ everywhere; i.e., the magnetic distortion is *maximal*. But apparently paradoxically the cost of *any* nematic distortion is positive, and hence $\theta(z)\equiv 0$; the nematic distortion is *minimal*.

However, as w increases beyond w_c, the high field nematic distortion increases. In the region $w\sim w_c$, we can again make the approximation $\theta(z) = \theta_0\sin\pi z$, in which case (23) reduces to

$$\text{(24)}\qquad F(\theta_0) \sim \frac{1}{2}\left(\frac{\pi^2}{2}-w\right)\theta_0^2 + \frac{w}{8}\theta_0^4 + \dots .$$

Minimizing this with respect to θ_0 yields

$$\text{(25)}\qquad \theta_0 \approx \sqrt{2}\left(1-\frac{w_c}{w}\right)^{\frac{1}{2}}.$$

Finally for $w \gg w_c$, $\theta(z)\approx\frac{\pi}{2}$ almost everywhere. There is a small healing region close to the boundaries, of dimension $w^{-\frac{1}{2}}$, over which $\theta(z)$ goes from zero to $\frac{\pi}{2}$.

3.2.4. General classification. On the basis of the evidence adduced above, we are now in a position to make a general classification of FN behavior into three regimes. The regimes are as follows.

(a) *Weak ferronematic coupling.* The weak coupling regime is defined by $w \leq w_c = \frac{\pi^2}{2}$. In this regime, as we have shown in section 3.2.2, the initial response $\theta_0(h)$ is proportional to h. We have shown in section 3.2.3 that $\lim_{h\to\infty} \theta_0 \equiv 0$, but in fact the condition is stronger. It will turn out that there is a critical field $h_c(w)$ at which the nematic distortion abruptly *disappears*, so that for $h \geq h_c(w)$, $\theta_0(h) \equiv 0$. In any event, there is a turning point at intermediate h_M, and for $h > h_M$, $\theta_0(h)$ decreases.

(b) *Intermediate ferronematic coupling.* We know from section 3.2.2 that at low fields the nematic response increases, with $\theta_0 \sim h$. If $w > w_c$, but $\left|\left(\frac{w}{w_c}\right) - 1\right| \ll 1$, we also know from section 3.2.3 that $\theta_0(h \to \infty)$ is small. Thus there will be a regime $w_c < w < w_{c2}$ for which necessarily $\theta_0(h)$ reaches a maximum as a function of h before decreasing at high fields. The precise value of w_{c2} remains to be determined.

(c) *Strong ferronematic coupling.* We know from section 3.1 that in the infinite coupling regime $\theta_0(h)$ increases for all h. We thus expect a regime defined by $w > w_{c2}$ for which this behavior is retained.

3.2.5. High field limit. We can now extend the considerations of the infinite field limit to fields which are merely high. In this case, given the considerations of section 3.2.1, this means $h \geq 2w$. The analysis involves constructing a free energy expansion in both nematic and magnetic orientation variables. Specifically, the expansion extends the considerations of (24) involving θ to include departures in ψ from complete alignment.

We recall from section 3.2.1 that for large h, $\psi_s = \frac{\pi}{2}$. Hence $\psi(z) = \frac{\pi}{2} - \gamma(z)$, where the angle $\gamma(z)$ is small and zero at the boundary. Likewise, at least in the weak and intermediate ferronematic coupling regimes discussed in section 3.2.4, the angle $\theta(z)$ may be regarded as small.

The appropriate expansion variables are then γ_0, θ_0, where

$$\psi(z) \approx \frac{\pi}{2} - \gamma(z), \quad \gamma(z) = \gamma_0 \sin \pi z, \quad \theta = \theta_0 \sin \pi z. \tag{26}$$

We shall construct a Landau expansion of F in the high field regime. From (6), we obtain

$$F = F_0 + \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - h\cos\gamma - w\sin^2(\theta+\gamma)\right], \tag{27}$$

where F_0 is a reference free energy defined for $\theta(z) \equiv 0$ and $\psi \equiv \frac{\pi}{2}$. Expanding (27) to fourth order in the angular variables θ_0, γ_0, using the relations (26), yields

$$F - F_0 = \left[\frac{\pi^2}{4}\theta_0^2 + \frac{h}{4}\gamma_0^2 - \frac{1}{2}w\left(\theta_0+\gamma_0\right)^2\right] - \frac{h}{64}\gamma_0^4 + \frac{1}{8}w\left(\theta_0+\gamma_0\right)^4 + \dots. \tag{28}$$

Here we have isolated the crucial terms quadratic in the variables θ_0, γ_0.

Stability is defined by this quadratic term, which can be written as

$$F_Q = \frac{1}{2}\left[w_c\,\theta_0^2 + \frac{h}{2}\gamma_0^2 - w\left(\theta_0+\gamma_0\right)^2\right] = \frac{1}{2}\left[\left(w_c - w\right)\theta_0^2 + \left(\frac{h}{2} - w\right)\gamma_0^2 - 2w\,\theta_0\gamma_0\right], \tag{29}$$

where we have substituted $w_c = \frac{\pi^2}{2}$. We have seen in section 3.2.3 that this is the critical ferronematic coupling beyond which the infinite field nematic distortion no longer vanishes, i.e., $\lim_{h\to\infty} \theta_0 \neq 0$. And indeed, when $\gamma_0 \equiv 0$, then by inspection it is clear that the expression given in (29) is positive definite (and hence $\theta_0 \equiv 0$) if and only if $w < w_c$.

For finite h, $\theta_0 = \gamma_0 = 0$ are minimizers of the free energy expression (28) if and only if the quadratic term (29) is positive definite. This will be the case if

$$w_c - w > 0 \tag{30a}$$

and also if the discriminant of expression (29) is positive:

$$\left(\frac{h}{2} - w\right)(w_c - w) > w^2. \tag{30b}$$

Rearranging (30b), we obtain a condition

$$\frac{h}{2}(w_c - w) - ww_c > 0 \tag{31a}$$

or

$$h^{-1} < h_c^{-1}(w) = \frac{1}{2}(w^{-1} - w_c^{-1}). \tag{31b}$$

The result of this calculation is that in the weak FN coupling limit, i.e., if $w - w_c < 0$, then the magnetic director is completely saturated, the nematic distortion will be zero not only in the limit of infinitely high field but also *for all* fields $h > h_c(w)$, where

$$h_c(w) = \frac{2}{(w^{-1} - w_c^{-1})} = \frac{2ww_c}{(w_c - w)}. \tag{32}$$

Equivalently, the quadratic form (29) is positive definite if both $w_c - w > 0$ and $h - h_c > 0$. Using (29), the free energy (28) can then be recast in a diagonal form in which the change of character at $h = h_c$, $w = w_c$ becomes explicit:

$$\begin{aligned} F - F_0 = \frac{1}{2}\Bigg[&\frac{2w_c^2}{2w_c + h}\left(\theta_0 - \frac{h}{2w_c}\gamma_0\right)^2 + \frac{(h - h_c(w))(w_c - w)}{2w_c + h}(\theta_0 + \gamma_0)^2 \\ &+ \frac{w}{4}(\theta_0 + \gamma_0)^4 - \frac{h}{32}\gamma_0^4\Bigg]. \end{aligned} \tag{33}$$

In all cases of interest there exists a minimizer of $F(\theta_0, \gamma_0)$ such that θ_0, γ_0 are either zero or small. The γ_0^4 term with a negative coefficient is swamped by the $(\gamma_0 + \theta_0)^4$ term.

3.2.6. The weak coupling regime. This is the regime $w < w_c = \frac{\pi^2}{2}$. We have seen that the nematic response increases rapidly at low fields and reaches a maximum. θ_0 then *decreases*, reaching zero at $h_c(w)$, where from (31b)

$$\frac{2}{h_c(w)} = \frac{1}{w} - \frac{1}{w_c}.$$

We describe this transition as an inverse Frederiks transition, because at high fields the nematic director remains undistorted, whereas for lower fields, deviation from the zero field equilibrium occurs. We note also that

$$\lim_{w\to w_c} h_c^{-1}(w) = 0; \quad \frac{h_c(w)}{2w} = \frac{1}{(1-w/w_c)}. \tag{34}$$

The critical field $h_c(w)$ *diverges* as $w \to w_c$, so that for $w > w_c$, the undistorted state no longer exists.

For $h > h_c(w)$ the magnetic director $\hat{\mathbf{m}}$ is completely aligned with the magnetic field, corresponding to $\psi(z) = \pi/2$ everywhere. This magnetic alignment is coupled to the nematic director through the colloidal particles, giving an effective perpendicular field on the nematic of magnitude w. As we have seen above, there is a strong analogy with the conventional Frederiks effect. If w is too low, even at high fields the effective aligning force on the nematic particles cannot overcome the elastic energy of the nematic director. The director thus remains unmoved by the field. The critical value w_c is just that field which corresponds to the Frederiks transition.

We also note that the alignment of the magnetic director ψ_s at the *surface* saturates at $h^*(w) = 2w$. However, *in the bulk* the magnetic saturation occurs only at the higher field $h_c(w) = 2w(1-w/w_c)^{-1} = h^*(w)(1-w/w_c)^{-1}$. The effect of the surface singularities in the response of the bulk system close to $h^*(w)$ is not clear; preliminary evidence suggests that they may be smoothed out. Cells with *weaker* nematic-magnetic coupling saturate at *lower* fields (i.e., more easily) because the saturation is discouraged by the nematic elastic term.

At $h_c(w)$ there is a transition to nonzero values of θ because the magnetic field is no longer sufficiently strong to hold the magnetic director perpendicular to the walls. The magnetic director then orients at some angle to the walls, and this in turn breaks the left-right symmetry to which the nematic director is subject. The nematic director thus follows the distorting magnetic director.

In order to analyze the behavior of ψ and θ just below h_c, it is necessary to analyze (33). The quadratic form in this equation is diagonal. The coefficient of the term

$$\frac{2w_c^2}{2w_c+h}\left(\theta_0 - \frac{h}{2w_c}\gamma_0\right)^2$$

is always positive. Thus, apart from corrections, $\theta(z) = \frac{h}{2w_c}\gamma(z)$, or

$$\gamma_0 = \frac{2w_c}{h}\theta_0. \tag{35}$$

We can now rewrite (33) as an expansion in θ_0 alone. In the spirit of Landau theory, close to $h = h_c(w)$, we replace all values of h by $h_c(w)$ except where the relevant term is $(h - h_c(w))$. After some algebra, we obtain

$$F - F_0 = \frac{w_c^2}{2w}\left[\left(1 - \frac{h_c(w)}{h}\right)\left(1 - \frac{w}{w_c}\right)\theta_0^2 + \frac{1}{4}\left(\frac{w_c}{w}\right)^2\left(1 - \frac{1}{4}\left(1 - \frac{w}{w_c}\right)^3\right)\theta_0^4\right]. \tag{36}$$

We note that the coefficient of the quartic term is always positive, despite the apparent negative term in γ_0^4 in (28). Minimizing (36) with respect to θ_0 yields

$$\theta_0 = \sqrt{2}\frac{w}{w_c}\left(\frac{(1-w/w_c)}{\left[1 - \frac{1}{4}(1-w/w_c)^3\right]}\right)^{1/2}\left[\frac{h_c(w)}{h} - 1\right]^{1/2}, \tag{37}$$

exhibiting the expected square root singularity as function of the field. The transition thus follows the normal paradigm.

As the magnetic field is further reduced, the nematic response increases. However, at very low fields, the magnetic director responds little to the magnetic field. Then finally the nematic response also decreases, in sympathy. The qualitative explanation of this phenomenon is as follows. The degree of distortion is associated with the couple exerted on the nematic director. This is proportional to $\sin 2(\theta - \psi)$. For the undistorted nematic ($\theta = 0$) this has a maximum at $\psi = \pi/4$. We thus expect a maximum in $\theta_0 = \theta(z = 1/2)$ to occur when $\psi \approx \pi/4$.

3.2.7. The intermediate coupling regime. We now discuss the regime for which $\theta_0(h)$ is monotonically decreasing at high h, but $\lim_{h\to\infty} \theta_0 \neq 0$. Now the ferronematic coupling is somewhat larger than in the previous case: $w_c < w < w_{c2}$, where we shall determine w_{c2}.

The Landau expansion again uses (35) connecting γ_0 and θ_0. Re-expanding in terms of the one relevant variable θ_0 yields

$$F \sim w_c \left(1 - \frac{w}{w_c} - \frac{2w}{h}\right)\left(1 + \frac{2w_c}{h}\right)\theta_0^2 + \frac{w}{4}\left(1 + \frac{2w_c}{h}\right)^4 \theta_0^4, \tag{38}$$

where now the γ_0^4 term in (33) is negligibly small at $w \gg \pi^2/6$. Minimizing this function, we obtain the following expression for $\theta_0(h)$:

$$\theta_0^2(h) = 2\frac{\left(1 - \dfrac{w_c}{w} + \dfrac{2w_c}{h}\right)}{\left(1 + \dfrac{2w_c}{h}\right)^3}. \tag{39}$$

We observe that in the limit $h \to \infty$, this expression is consistent with (37). Indeed, we can expand $\theta_0(h)$ in terms of $\theta_0(h = \infty)$, yielding

$$\theta_0^2(h) = \theta_0^2(h = \infty)\frac{\left(1 + \dfrac{2w_c}{h\left(1 - \dfrac{w_c}{w}\right)}\right)}{\left(1 + \dfrac{2w_c}{h}\right)^3}. \tag{40}$$

Expanding this, and taking square roots, we obtain, now to leading order in h^{-1},

$$\theta_0(h) = \theta_0(\infty)\left[1 + \left(\frac{w_c}{h}\right)\left(\frac{1}{1 - \dfrac{w_c}{w}} - 3\right)\right]. \tag{41}$$

This is the key result of this section. It shows that for w such that $\frac{1}{1-w_c/w} > 3$, $\theta_0(h)$ for large fields is an *increasing* function of *inverse* field and hence a *decreasing* function of h. This condition can be rewritten as

$$w_c < w < w_{c2}, \quad w_{c2} = \frac{3}{2}w_c. \tag{42}$$

Equation (42) improves the estimate of w_{c2} given in our previous paper [15].

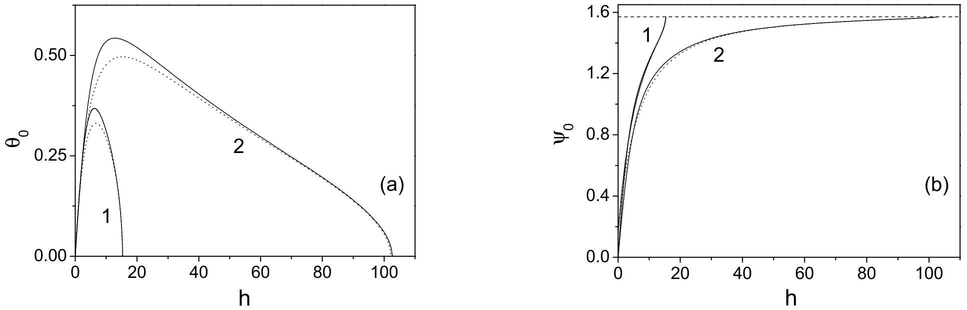


FIG. 3. *Weak coupling ferronematics: Nematic* (a) *and magnetic* (b) *directors as a function of dimensionless external field, in the absence of a bias field. Quantities θ_0 and ψ_0 in the center of the cell are used as surrogates for global behavior. Curve* 1 *corresponds to $w = 3$ and curve* 2 *to $w = 4.5$; $w < w_c \approx 4.93$. The dotted lines have been added to show comparison with the results of the asymptotic calculation* (36) *in section* 3.2.6.

3.2.8. The strong coupling regime. This is the regime $w > w_{c2} = \frac{3}{2}w_c$. We can see from (41) that in this regime, the angle θ_0 *increases* at high field. It is further not clear from the asymptotic analysis whether there is a region of w for which θ_0 is not a monotonic function of h, or whether the maximum in $\theta_0(h)$ disappears exactly at w_{c2}. The numerical results are consistent with the latter hypothesis.

3.3. Numerical results. We have also carried out a numerical minimization of the relevant free energy (3). The method does not involve a direct brute force quadrature-based solution of the resulting Euler–Lagrange equations. Rather we use the existence of a set of first integrals, which allows us to parameterize the solutions in terms of the values of the parameters ψ_0, θ_0, η_0, where these quantities are the values of the relevant parameters in the middle of the cell. These quantities satisfy algebraic self-consistency conditions. The method has been used in previous publications [13, 14]. We shall present a detailed discussion of the merits of this approach elsewhere.

Figures 3–6 illustrate the behavior of weak, intermediate, and strong-coupling FNs in the unsegregated limit. Solid curves are numerical solutions. A comparison of numerical calculations and asymptotic results is given by dotted lines.

Figure 3 shows the *inverse Frederiks transition.* As h increases, the nematic response first increases and then decreases, disappearing at an inverse Frederiks transition at $h = h_c(w)$. The magnetic response is a monotonically increasing function of h but saturates at $h = h_c(w)$ and $\psi_0 = \pi/2$.

The profiles of the nematic and magnetic directors for weak-coupling FNs at low and high fields are shown in Figure 4. These profiles demonstrate that the harmonic approximations for $\theta(z)$ and $\psi(z)$ given in (17) and (26) are extremely good. There is a change of the magnetic director profile shape when h is close to $2w$. Our investigation shows that this change in the profile structure is due to the ambiguity of the dependence of θ (14b) on h or ψ when $\psi_s > \pi/4$ or $h > \sqrt{2}w$, respectively.

Figure 5 also illustrates peculiar changes in the magnetic director behavior with h. For low fields ($h < \sqrt{2}w$) the magnetic profile is concave-down and $\psi_0 - \psi_s > 0$. For high fields the profile is concave-up. In the process of changing the profile from concave-up to concave-down, there is the intermediate concave-convex profile in the narrow interval of h close to $h = 2w$. At a certain value of h in this interval $\psi_0 = \psi_s$ and $\Delta\psi = 0$.

Figure 6 shows the nematic and magnetic response in the intermediate and strong-

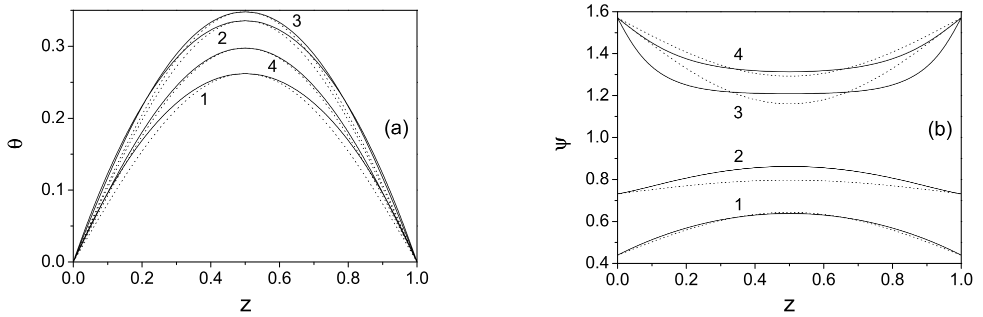


FIG. 4. *Weak coupling ferronematics: Nematic distortion* $\theta(z)$ (a) *and magnetic distortion* $\psi(z)$ (b) *as a function of position for different dimensionless external fields, in the absence of a bias field, with* $w = 3$*,* $w < w_c$*. Curve* 1*,* $h = 2.55$*; curve* 2*,* $h = 4.0$*; curve* 3*,* $h = 8.38$*; curve* 4*,* $h = 10.56$*. Solid curves are numerical solutions. Dotted curves are asymptotic solutions; in curves* 1 *and* 2 *we used the low* h *expansion, and in curves* 3 *and* 4 *we used the high* h *expansion. Note how* $\theta(z)$ *always increases in the center of the cell, but in the weak coupling case* $\psi(z)$ *increases for low* h *but decreases for higher* h*.*

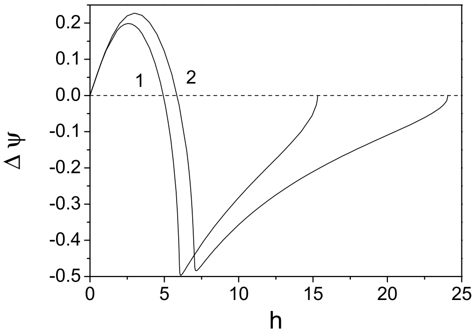


FIG. 5. *Differences between surface and bulk behavior as a function of field* h *for weak ferronematic couplings* $w = 3$ *(curve* 1*) and* $w = 3.5$ *(curve* 2*). The quantity* $\Delta\psi = \psi_0 - \psi_s$ *is the difference between the magnetic distortion at the center of the cell and its value at the surface. Note the discussion in the text.*

coupling regimes. As h increases, the nematic response of the intermediate coupling FN first increases and then decreases, but $\lim_{h\to\infty} \theta_0 = \sqrt{2(1 - w_c/w)} \approx 0.6 \neq 0$. In the strong-coupling regime the nematic response is a monotonically increasing function of h and saturates at high h. The magnetic response is a monotonically increasing function of the field in both regimes.

4. Bias field. Experimentally, the bias field is imposed in order to maintain particles and hence the director in the x-z plane, whereas we simply *assume*, even in the absence of a bias field, that the orientation is maintained in this plane. The zero-bias-field case then becomes the distinguished limit which we have discussed in the last section. *Physically* the important point about the bias field is that the resultant magnetic field is never entirely perpendicular to the cell plane. The consequence of this is that the angle ψ can reach its saturation value of $\pi/2$ only at infinitely high fields (i.e., $h^{-1} \equiv 0$). This contrasts with the zero-bias-field case, for which, as we have seen, at least for $w < w_c$, $\psi \equiv 0$ for $h > h_c(w)$.

For analytical purposes it is convenient to discuss small perturbations from the

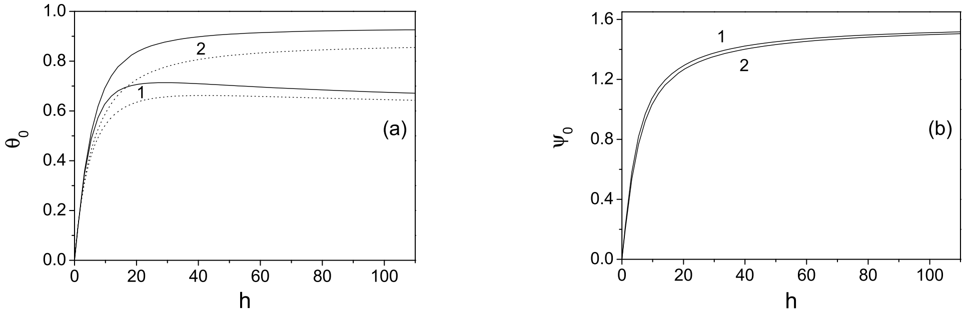


FIG. 6. *Intermediate and strong coupling ferronematics: Nematic* (a) *and magnetic* (b) *directors as a function of dimensionless external field, in the absence of a bias field. (Curve* 1*)* $w = 6$, $w_c < w < 3w_c/2$*, corresponding to intermediate ferronematic coupling; (curve* 2*)* $w = 8$, $w > 3w_c/2$*, corresponding to strong ferronematic coupling. Dotted lines show comparison with the results of the asymptotic calculation which minimizes the free energy* (28)*. This figure is the intermediate and strong FN coupling version of Figure* 3.

zero-bias-field case already considered. It is sensible to scale the bias field, by analogy with other nondimensionalization in the problem, leading to a dimensionless bias field $h_b = \bar{f}\bar{M}H_bD^2/K$. The constant η, high t free energy is now

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 - h\sin\psi - h_b\cos\psi \right.$$
$$\left. + w\sin^2(\theta - \psi)\right]. \tag{43}$$

This is the bias field analogue of (3). The Euler–Lagrange equations now become

$$\frac{d^2\theta}{dz^2} - w\sin(2(\theta - \psi)) = 0, \tag{44a}$$

$$h\cos(\psi) - h_b\sin\psi + w\sin(2(\theta - \psi)) = 0, \tag{44b}$$

where (14a) and (44a) are identical, and (44b) differs from its zero-bias-field analogue (14b) by an extra factor $-h_b\sin\psi$.

The most dramatic effects of the bias field occur in the low w regime. In this regime, when there is no bias field, the magnetic director saturates. The saturation drives the high field absence of nematic director distortion and the inverse Frederiks transition. The bias field destroys the magnetic saturation and thus fundamentally affects the inverse Frederiks effect.

We first discuss the boundary values ψ_s. We recall from section 3.2.1 that in the absence of a bias field, $\psi_s(h)$ is a monotonically increasing function of h, saturating at $\psi_s = \frac{\pi}{2}$ at $h = h^* = 2w$. We shall investigate the behavior of $\gamma_s = \frac{\pi}{2} - \psi_s$, which in the $h_b = 0$ case vanishes identically for $h > 2w$. Combining (44b) and the condition $\theta_s = 0$ yields the bias field analogue of (15):

$$h\cos\psi_s - h_b\sin\psi_s = w\sin 2\psi_s. \tag{45}$$

For $h_b = 0$ there are two regimes separated by a singularity at $h = 2w$. In the $h > 2w$ regime $\psi_s = \pi/2$; $\gamma_s = 0$, whereas for $h < 2w$, $\gamma_s \neq 0$. However, a finite value

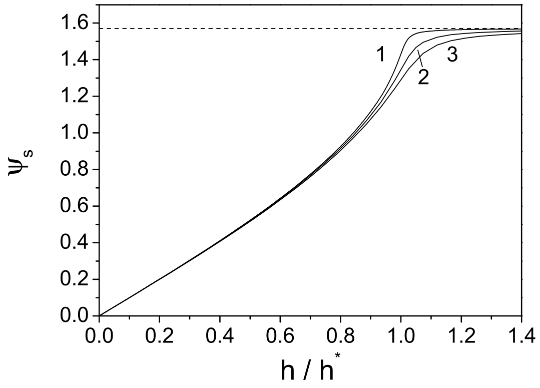


FIG. 7. *Behavior of the magnetic surface director deviation $\psi_s(h)$ as a function of bulk field h for various different values of bias field h_b, with $w = 0.9$, $w < w_c$. Curve* 1: $h_b = 0.01$. *Curve* 2: $h_b = 0.05$. *Curve* 3: $h_b = 0.1$. *Compare to the zero-bias-field case shown in Figure* 2. *Note the rounding of the sharp singularity at $h = h^* = 2w$.*

of h_b rounds this singularity, and the behavior of ψ_s and hence all other quantities can no longer be divided into two distinct regimes. Rewriting (45) in terms of γ_s, we obtain

$$h \sin\gamma_s - h_b \cos\gamma_s = w \sin 2\gamma_s. \tag{46}$$

Linearizing in the regime of small $\gamma_s \ll 1$ yields

$$h\gamma_s - h_b = 2w\gamma_s, \tag{47}$$

yielding

$$\gamma_s = \frac{h_b}{h - 2w}. \tag{48}$$

We plot exact numerical results for $\psi_s(h)$ for a number of different values of h_b in Figure 7. Equation (48) can be regarded as a response by the FN surface to a bias field probe. The unbiased system is unstable with respect to a perturbation of γ_s at $h = 2w$. One should thus expect that the susceptibility of the magnetic director to the small bias field would diverge at $h = 2w$, as indeed occurs (although for fields of this order, the small γ_s approximation no longer holds). The nonzero γ_s is a signature of a nonzero γ_0, and hence of a nonzero θ_0, as we now show.

Analogous results are found for the magnetic and nematic director deviations in the center of the sample. Numerical results are presented in Figure 8.

The results in Figure 8 can be understood semiquantitatively as follows. Using the expansion of (26), we can rewrite the functional (43) in a power law expansion in γ_0 and θ_0, yielding for the relevant terms

$$\begin{aligned} F - F_0 = \frac{1}{2}\Bigg\{ &(w_c - w)\left[\theta_0 - \frac{h_c(w)}{2w}\gamma_0\right]^2 + \frac{1}{2}(h - h_c(w))\gamma_0^2 \\ &- \frac{4}{\pi}h_b\gamma_0 + \frac{w}{4}(\theta_0 + \gamma_0)^4 - \frac{h}{32}\gamma_0^4 + \frac{4}{9\pi}h_b\gamma_0^3\Bigg\}. \end{aligned} \tag{49}$$

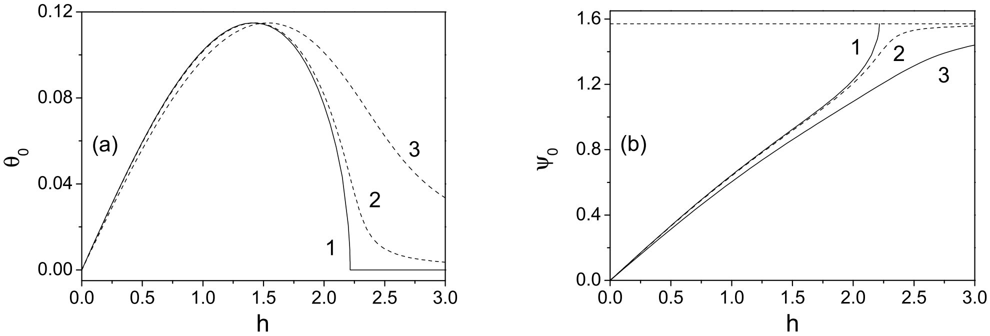


FIG. 8. *Effect of bias field on the bulk distortion in the weak ferronematic coupling regime. In all cases $w = 0.9$, $w < w_c$, and the nematic director* (a) *behavior is shown. Curve* 1*: No bias field* ($h_b = 0$)*. Curves* 2 *and* 3*: Weak bias field* ($h_b = 0.01, 0.1$)*. The re-entrant Frederiks transition is rounded in the presence of a bias field. The analogous curve for the magnetic director* (b) *is monotonic as a function of h, saturates at $\pi/2$ for $h_b = 0$, but merely tends to $\pi/2$ as $h \to \infty$ for $h_b \neq 0$.*

For high fields, the main balance defining γ_0 comes from the linear and quadratic terms; the fourth order terms can be neglected. With this assumption, we minimize (49). This yields $\gamma_0 = (2w_c/h_c(w))\,\theta_0$ for the relationship between γ_0 and θ_0, as well as an equation for $\gamma_0(h)$:

$$\gamma_0 = \frac{4}{\pi(h - h_c(w))} h_b. \tag{50}$$

There is thus no longer an unambiguous distinction between the high field $h > h_c(w)$ regime, for which $\gamma = \theta_0 = 0$, and the low field $h < h_c(w)$ regime, for which $\gamma_0 \neq \theta_0 \neq 0$. In the high field regime both θ and ψ differ from their zero-bias values 0 and $\pi/2$, respectively, by quantities which are proportional to the bias field, and appear for $h \gg h_c(w)$ to be heading for a divergence at $h = h_c^+(w)$. This divergence is, however, prevented by higher order terms in (49) which have been neglected here. On the other hand, there is no qualitative change in the low field response; the magnitude of the response is proportional to h_b.

Taking into account fourth order terms in (49), we find at $h = h_c(w)$

$$\gamma_0 \approx 4w \left[\frac{h_b}{\pi h_c(w)(h_c^3(w) - 2w^3)} \right]^{1/3}. \tag{51}$$

Here the magnitude of the response is proportional to $h_b^{1/3}$.

5. Ferroparticle segregation. In the theory as written so far, the ferroparticle density is kept constant. This corresponds to the limit $\eta = 1$ in (3), which in turn follows in the infinite temperature limit $t \to \infty$. However, inserting a finite value of t permits the ferroparticle density to respond so that more particles can migrate to regions where the ferronematic coupling energy is minimized. Interestingly (and apparently paradoxically) the constant η limit corresponds to inserting an apparently *infinite* rather than a zero term. We note that in the real problem t takes its physical value; it is useful nevertheless to treat it as a variable parameter in the theory.

We recall the figures of merit for the degree of segregation s defined in (4) and η_0 defined in (5) in section 2. Then $s = 0, 1$ are the limits of complete lack of segregation and segregation, respectively. The quantity s is a segregation order parameter, and it possesses a status comparable in the theory to the angular quantities θ_0 and ψ_0. In the low segregation limit we can expand the normalized colloidal density as follows:

$$\eta(z) = 1 - 2s\cos 2\pi z. \tag{52}$$

5.1. Zero bias field. The full free energy is given by (3):

$$F = \int_0^1 dz \left[\frac{1}{2}\left(\frac{d\theta}{dz}\right)^2 + \eta t \ln \eta - \eta h \sin\psi + \eta w \sin^2(\theta - \psi) \right],$$

subject to the constraint $\int_0^1 \eta(z)\,dz = 1$. The Euler–Lagrange equations are now

$$\frac{d^2\theta}{dz^2} - \eta w \sin(2(\theta-\psi)) = 0, \tag{53a}$$

$$h\cos\psi + w\sin(2(\theta-\psi)) = 0, \tag{53b}$$

$$t\ln\eta - \left(h\sin\psi - w\sin^2(\theta-\psi)\right) = \lambda, \tag{53c}$$

where λ is a Lagrange parameter which enforces density conservation. Equation (53a) modifies (14a) to the case in which segregation is allowed; (53b) is in fact identical to (14b) and is unchanged by the addition of segregation while (53c) is a new equation for the self-consistent degree of segregation.

The exact solution for $\eta(z)$ comes from inverting (53c) and enforcing the density conservation condition $\bar{\eta} = 1$. This solution is

$$\eta(z) = \frac{\exp\left\{\left[h\sin\psi - w\sin^2(\theta-\psi)\right]/t\right\}}{\int_0^1 dz \exp\left\{\left[h\sin\psi - w\sin^2(\theta-\psi)\right]/t\right\}}. \tag{54}$$

In general, this solution must be determined self-consistently with solutions for $\psi(z)$ and $\theta(z)$, and the detailed picture is complicated.

However, in the high temperature limit η is always small. We can then describe the degree of segregation using the order parameter s and perturb away from the infinite t solution using (4). In this limit

$$s = -\frac{1}{t}\int_0^1 dz \cos 2\pi z \left[h\sin\psi - w\sin^2(\theta-\psi)\right], \tag{55}$$

where values of θ and ψ are given by the infinite t limit. In general, we expect values of $s(h)$ to peak at intermediate h, for it is in these cases that θ and ψ change most across the cell.

A particularly interesting case occurs for high fields in the low ferronematic coupling $w < w_c$ regime. Here it is possible to include the segregation order parameter s in an extended Landau expansion closely analogous to (28). The parameter s couples to the variables θ_0, ψ_0 in this expansion. We recall (17): in this regime $\psi \approx \frac{\pi}{2} - \gamma_0 \sin\pi z$; $\theta(z) = \theta_0 \sin\pi z$.

We obtain the following free energy:

$$\begin{aligned} F - F_0 = \frac{1}{2}\Bigg\{ & \frac{2w_c^2}{2w_c + h}\left(\theta_0 - \frac{h}{2w_c}\gamma_0\right)^2 + \frac{(h - h_c(w))(w_c - w)}{2w_c + h}(\theta_0 + \gamma_0)^2 \\ & + \frac{w}{4}(\theta_0+\gamma_0)^4 - \frac{h}{32}\gamma_0^4 + 2s^2 t - s\left[w(\theta_0+\gamma_0)^2 - \frac{h}{2}\gamma_0^2\right]\Bigg\}. \end{aligned} \tag{56}$$

Equation (56) is (33) modified by some extra terms in s. The quadratic term in s comes from the $\eta \ln \eta$ term in the free energy (3). The linear term comes from the evident fact that the colloidal density is coupled linearly both to the nematic-magnetic coupling and to the coupling of the magnetic director with the field.

In the high field $h > h_c(w)$ limit there is no structure at all in the infinite t regime, for now $\psi(z) = \pi/2$ and $\theta(z) = 0$. As a result, (54) and (55) show that in this regime $s = 0$ and there is no segregation. However, in the low h limit there is structure in $\psi(z), \theta(z)$ and so $\eta(h) \neq 0$ for $h < h_c(w)$. As a result, the order parameters s and γ_0 couple in a Landau expansion of the free energy of the system close to $h_c(w)$. The coupling is constant, but the stabilizing term in s is proportional to t. It is this fact which gives rise to the general result $s \sim t^{-1}$.

However, the linear coupling between s and quadratic terms in the other order parameters does have profound consequences. This is a consequence of the Halperin–Lubensky–Ma theorem [24]. This theorem states that coupling a critical order parameter to a second noncritical order parameter can under some circumstances drive a continuous phase transition first-order.

To show what happens in this case, we minimize (56). We obtain (35) for the relationship between γ_0 and θ_0, and additional equations for s and γ_0:

$$s = \frac{hh_c}{16tw_c}\left[1 + \frac{4w_c^2 - hh_c}{h(2w_c + h_c)}\left(1 - \frac{h}{h_c}\right)\right]\gamma_0^2, \tag{57}$$

$$\gamma_0^2 = 8\left\{\left[\frac{(2w_c + h)^2}{2w_c(2w_c + h_c)} - \frac{h}{h_c}\right]s + \frac{2w_c + h}{2w_c + h_c}\left(1 - \frac{h}{h_c}\right)\right\}\left[\frac{(2w_c + h)^4}{2w_c^3(2w_c + h_c)} - \frac{h}{h_c}\right]^{-1}. \tag{58}$$

Substituting (35) and (57) into (56), we can evaluate the fourth order term in γ_0 at $h = h_c(w)$:

$$F - F_0 = \left[\frac{h_c^3(w)}{2}\left(\frac{1}{w^3} - \frac{1}{2\,t\,w_c^2}\right) - 1\right]\frac{h_c(w)}{64}\,\gamma_0^4. \tag{59}$$

The key point is that the coupling terms provide a negative definite contribution to the fourth order term. The negative magnitude increases as the coupling (which in this case is the temperature) decreases. Eventually, at a sufficiently low temperature the sign of the γ_0^4 term in the Landau expansion close to $h = h_c(w)$ changes. A negative γ_0^4 term signals that the continuous transition at $h = h_c(w)$ becomes first order.

From (59) we find that the tricritical point (i.e., the point at which the continuous phase transition becomes first order) occurs for

$$t_c(w) = \frac{w^3}{2w_c^2\left(1 - \dfrac{2w^3}{h_c^3(w)}\right)} \approx \frac{w_c}{2}\left(\frac{w}{w_c}\right)^3. \tag{60}$$

We can also look at the properties of $\gamma_0(h)$ and $s(h)$ *below* $h = h_c(w)$ but *above* $t = t_c(w)$. In this region the transition is still continuous but is approaching tricriticality. By substituting (58) into (57) and expanding s in powers of $(h_c(w) - h)$, we find the segregation order parameter

$$s = \frac{2w_c t_c(w)}{h_c^2(w)\,(t - t_c(w))}\,(h_c(w) - h) \tag{61}$$

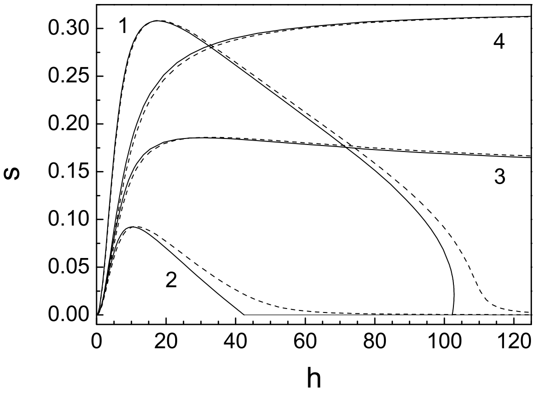


FIG. 9. *Behavior of the segregation order parameter as a function of the dimensionless external field in the three ferronematic coupling regimes. (Curve* 1*) Weak coupling regime;* $w = 4.5$*,* $t = 1.78 < t_c = 1.87$*,* $h_c = 102.1$*. (Curve* 2*) Weak coupling regime;* $w = 4$*,* $t > t_c = 1.32$*,* $h_c = 42.3$*. (Curve* 3*) Intermediate coupling regime;* $w = 5.5$*. (Curve* 4*) Strong coupling regime;* $w = 7$*.* $t = 3.85$ *in curves* 2–4*. Solid curves: No bias field. Dashed curves: Weak bias field;* $h_b = 0.1$ *in curve* 1 *and* $h_b = 0.4$ *in curves* 2–4*.*

in the lowest order approximation. We find that s goes *linearly* with $(h_c(w) - h)$ but that the linear coefficient diverges at the tricritical point.

Likewise, from (57) and (58) we find

$$\gamma_0^2 = \frac{32tw_c^2 t_c(w)}{h_c^4(w)(t - t_c(w))}(h_c(w) - h). \tag{62}$$

Here the characteristic square root behavior for γ_0 is maintained, but the coefficient of proportionality diverges with a square root divergence as the tricritical point is approached.

Beyond the tricritical point (i.e., $t < t_c(w)$), there are solutions for $h > h_c(w)$. These solutions belong to a van der Waals loop. Thus there are two qualitatively different behaviors, separated by a singularity at $t = t_c(w)$, with $w < w_c$. For $t > t_c(w)$ we have the inverse Frederiks transition described in detail in the sections above. For $t < t_c(w)$, on the other hand, the functional dependence of s, θ, and ψ develops a van der Waals loop. These cases are illustrated in Figure 9 (solid curves 2 and 1), where we show curves calculated numerically by minimizing the cell free energy (1a).

To show what happens in the intermediate and strong-coupling regimes, we rewrite (56) in the form

$$\begin{aligned} F - F_0 = \frac{1}{2}\Bigg\{ &\frac{2w_c w}{2w_c + h}\left[\frac{w_c}{w}\left(\theta_0 - \frac{h}{2w_c}\gamma_0\right)^2 - \left(1 + h\frac{w - w_c}{2w_c w}\right)(\theta_0 + \gamma_0)^2\right] \\ &+ \frac{w}{4}(\theta_0 + \gamma_0)^4 - \frac{h}{32}\gamma_0^4 + 2s^2 t + s\left[\frac{h}{2}\gamma_0^2 - w(\theta_0 + \gamma_0)^2\right]\Bigg\}, \end{aligned} \tag{63}$$

where $w > w_c$. We now minimize (63). For $t \gg 1$ this yields (35) for the relationship between γ_0 and θ_0. The equations for s and θ_0 are now

$$s = \frac{1}{4t}\left[w\left(1+\frac{2w_c}{h}\right)^2 - \frac{2w_c^2}{h}\right]\theta_0^2, \tag{64a}$$

$$\theta_0^2 = \frac{2\left(1-\frac{w_c}{w}+\frac{2w_c}{h}\right)\left(1+\frac{2w_c}{h}\right)}{\left(1+\frac{2w_c}{h}\right)^4 - \frac{2w_c^4}{wh^3} - \frac{w}{2t}\left[\left(1+\frac{2w_c}{h}\right)^2 - \frac{2w_c^2}{wh}\right]^2}. \tag{64b}$$

From (64a) and (64b) one can obtain to lowest order in h^{-1}

$$s = \frac{w-w_c}{2t-w}\left[1+\frac{2w_c^2}{wh}\left(\frac{2w}{w-2t} - \frac{w_c}{w_c-w}\right)\right], \tag{65a}$$

$$\theta_0^2 = \frac{4t}{2t-w}\left[\frac{w-w_c}{w} + \frac{2w_c}{wh}\left(5w_c - 2w + \frac{2w_c(2t-w_c)}{w-2t}\right)\right]. \tag{65b}$$

Equations (65a) and (65b) are valid in the high field limit if $t > w - w_c/2$ (by definition, $s \leq 1$). In the limit $t \to \infty$, (65b) reduces to its no-segregation limit (41). It follows from (65b) that the boundary w_{c2} of the intermediate coupling regime at finite temperature reduces to

$$w_{c2}(t) = t + \frac{5}{4}w_c - \frac{1}{4}\left(16t^2 - 8tw_c + 9w_c^2\right)^{1/2} \approx \frac{3}{2}\,w_c\left(1-\frac{w_c}{6t}\right), \tag{66}$$

and this regime takes place at $w_c < w < w_{c2}(t)$. The strong coupling regime should begin at $w_{c2}(t)$. It is seen from (64a) that $s(h)$ has a maximum if $w_c < w_{sc}(t)$, where

$$w_{sc}(t) = \frac{w_c}{4}\left(1+\sqrt{1+\frac{16t}{w_c}}\right). \tag{67}$$

The behavior of the segregation order parameter in the intermediate and strong-coupling regimes is shown in Figure 9 (curves 3 and 4).

5.2. Nonzero bias field. For the nonzero bias field, the free energy can be expressed as

$$\begin{aligned} F - F_0 = \frac{1}{2}\Bigg\{ & \frac{2w_c w}{2w_c+h}\left[\frac{w_c}{w}\left(\theta_0 - \frac{h}{2w_c}\gamma_0\right)^2 - \left(1+h\frac{w-w_c}{2w_c w}\right)(\theta_0+\gamma_0)^2\right] \\ & - \frac{4h_b}{\pi}\,\gamma_0 + \frac{4}{9\pi}h_b\gamma_0^3 - \frac{8h_b}{3\pi}\,\gamma_0 s + \frac{w}{4}(\theta_0+\gamma_0)^4 + 2ts^2 \\ & - w(\theta_0+\gamma_0)^2 s + \frac{h}{2}\,\gamma_0^2 s - \frac{h}{32}\,\gamma_0^4\Bigg\}. \end{aligned} \tag{68}$$

Here (68) is (56) modified by some extra terms in h_b. Minimizing this equation, we obtain (35) for the relationship between γ_0 and θ_0. For high fields we can find the relationship between s and γ_0:

$$s = \frac{1}{8t}\left\{\frac{16}{3\pi}\,h_b\gamma_0 - \left[h - 2w\left(1+\frac{h}{2w_c}\right)^2\right]\gamma_0^2\right\}. \tag{69}$$

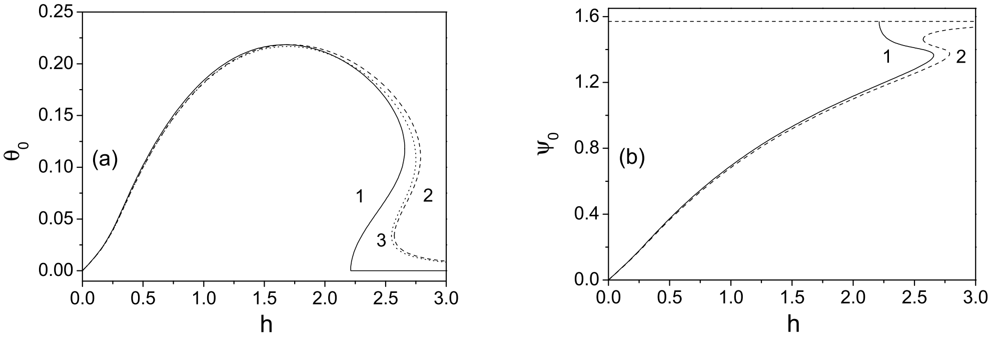


FIG. 10. (a) *Director distortion for weak ferronematic coupling at a finite temperature as a function of the external field, showing the development of a van der Waals loop at the inverse Frederiks transition. In both curves* $t = 0.005$*,* $w = 0.9$ *(*$t < t_c \approx 0.018$*). Curve* 1*: No bias field;* $h_b = 0$*. Curve* 2*: Weak bias field;* $h_b = 0.025$*. The analogous curve for the magnetic director* (b) *is nonmonotonic as a function of* h*, but saturates at* $\pi/2$ *for* $h_b = 0$*, but merely tends to* $\pi/2$ *as* $h \to \infty$ *for* $h_b \neq 0$*. Curve* 3*: The effect of small diamagnetic anisotropy (*$\kappa = \frac{1}{2}\chi_a K/(\bar{f}\bar{M}D)^2 = 0.015$*).*

A solution for θ_0 can be obtained from a cubic equation in which linear and cubic terms play an important role. However, this equation is too complicated and is not presented here.

The dashed curve 2 in Figure 9 shows that the bias field rounds the re-entrant Frederiks transition. The van der Waals loop is retained at the small bias field (not shown in the figure), but no longer occurs at a sufficiently high bias field (dashed curve 1).

We now make remarks concerning the importance of the bias field. The role of the bias field in the x direction is to restrict the nematic and magnetic directors to the x-z plane. In the absence of the bias field but in high applied fields, the system will *choose* a (broken symmetry) plane in which to orient. A detailed analysis requires a full treatment of the local statistical mechanics of the ferronematic ordering and will be discussed elsewhere.

We have analyzed in detail the behavior of a ferronematic system at high dimensionless temperatures ($t \geq 1$). However, our analysis is restricted to a weak bias field and runs into difficulties if the magnetic particles are not well aligned. Our segregation parameter s is a useful tool only for high t.

In experiments, however, it is more common to encounter a low-coupling regime at low t, and it is this regime which is of prime interest for ferronematic applications. However, to investigate the segregation effect we now have to use the quantity η, which can take values much larger than unity. Furthermore, the asymptotic analysis gives little insight into the system properties at large deviations from the initial alignment. We thus resort to numerical studies. Figures 10–13 show aspects of the system behavior in the low-coupling regime at low temperature t for system parameter values in the experimental region.

In Figures 10 and 11 we illustrate the orientational behavior of the nematic and magnetic directors and the quantity η in the varying magnetic field in the middle of the cell. The figures show the development of the van der Waals loop.

We remark that the van der Waals loop, and thus the first order transition between a highly segregated low field phase and a slightly segregated high field case, is retained even when a bias field is introduced. In this case, however, if there is no first order

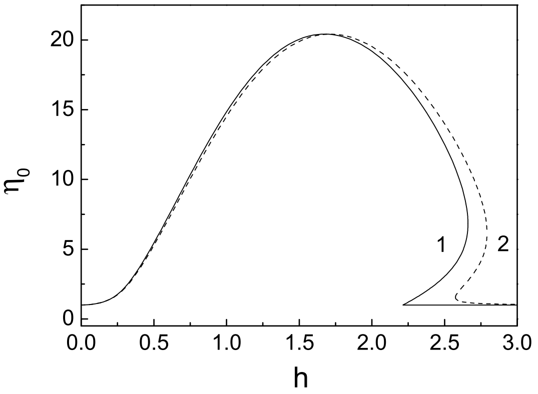


FIG. 11. *Ratio of the local to the mean volume particle fraction for weak ferronematic coupling at a finite temperature as a function of the external field, showing the development of a van der Waals loop at the inverse Frederiks transition. In both curves* $t = 0.005$, $w = 0.9$*;* $t < t_c \approx 0.018$*. Curve* 1*: No bias field;* $h_b = 0$*. Curve* 2*: Weak bias field;* $h_b = 0.025$*.*

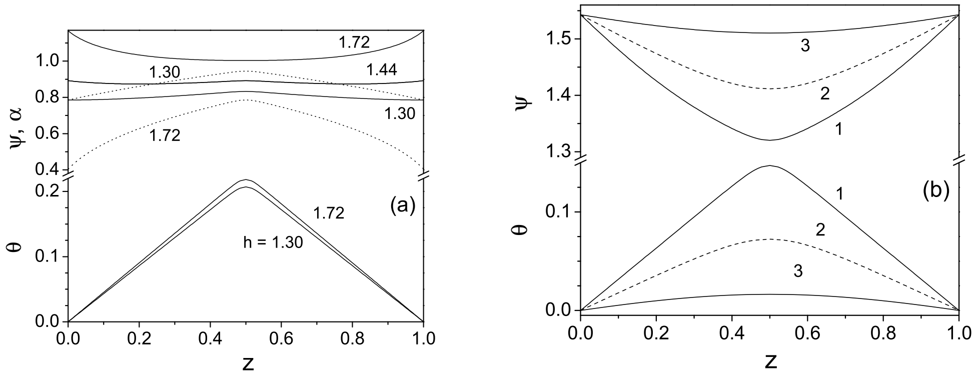


FIG. 12. (a), (b) *Orientational profiles at* $t = 0.005$, $w = 0.9$, *and* $h_b = 0.025$*.* α *(dotted lines) is the angle between magnetic and nematic directors. Curves* 1–3 *in* (b) *correspond, respectively, to the upper, middle (unstable), and lower parts of the van der Waals loop at* $h = 2.7$*.*

transition, then the continuous transition found in the zero-bias-field case disappears and is replaced by smooth high field behavior (curve 2 in Figures 10 and 11). In fact, we have found numerical evidence for van der Waals loops in previous calculations [14, 15]. This analysis presented here provides for the first time a consistent explanation.

We also show, in Figures 12 and 13, some examples of the profiles of the nematic and magnetic directors and of magnetic colloid concentration profiles through the cell. In these examples the system is in the weak coupling regime; as the field is increased the system exhibits a van der Waals loop and a consequent first order phase. We recall that in this regime, for low fields, the degree of nematic distortion, represented by $\theta(1/2) = \theta_{\max}$, goes through a maximum before tumbling in a discontinuous way and subsequently decreasing to zero in the limit of the high field.

The profile of the magnetic director in Figure 12(a) (given by the angle ψ) shows a change of regime from low fields to high fields. For low fields the profile is concave-down (i.e., $\psi(1/2) > \psi_s$), whereas for high fields it is concave-up ($\psi(1/2) < \psi_s$). We

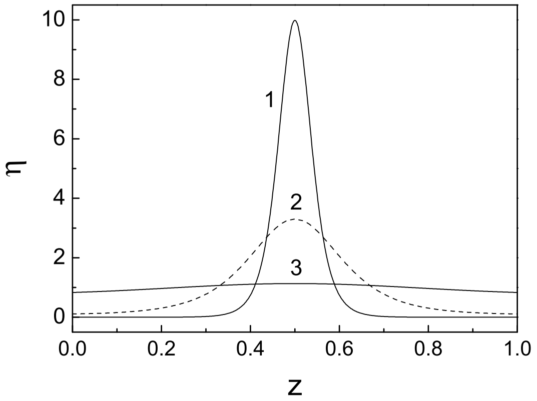


FIG. 13. *Concentration profiles at* $t = 0.005$*,* $w = 0.9$*, and* $h = 2.7$*,* $h_b = 0.025$*. Curves* 1*–*3 *correspond to the upper, middle (unstable), and lower parts of the van der Waals loop, respectively.*

have seen this behavior above in the non-bias-field case in Figure 4(b). The change from concave-down to concave-up profiles occurs near the peaks in $\theta_{\max}$ (Figure 10(a)) and $\eta_{\max}$ (Figure 11) ($h_{\max} = 1.68$). From this figure we note a well-defined intermediate concave-convex profile. The orientational profile of the nematic director θ is concave-down everywhere as a result of the boundary conditions on θ. However, we note that the behavior of $\theta(z)$ is unlike that in conventional nematics. In conventional nematics the nematic director can saturate over most of the cell apart from a region very close to the boundary. Here, however, we see a constant (absolute) *gradient* in θ, apart from very close to the center of the cell, where θ abruptly changes its gradient.

We give here a brief semiquantitative discussion of the changing properties of $\psi(z)$. As h is increased, the magnetic director profile begins to change when $\psi(0) = \psi(1) = \psi_s$ reaches the value $\frac{\pi}{4}$. Using (44b) it can be shown that this occurs at $h^{**} = h_b + \sqrt{2}w$. For $h < h^{**}$, $\frac{d\psi(0)}{dz} > 0$ and ψ increases away from the boundary. At h^{**}, $\frac{d\psi(0)}{dz} = 0$, and for $h > h^{**}$, $\frac{d\psi(0)}{dz} < 0$; ψ now *decreases* away from the boundary. There is then a narrow interval in h over which the turning points in $\psi(z)$ move from the edge of the cell toward its center. In this interval for $0 < z < z_c(h)$, $\frac{d\psi}{dz} < 0$, whereas for $z_c < z < \frac{1}{2}$, $\frac{d\psi}{dz} > 0$, with $\psi(z) = \psi(1-z)$ everywhere. Eventually, for $h > h^{\dagger}$ (which depends in specific cases on h_b), z_c reaches $z_c = \frac{1}{2}$, and for $h > h^{\dagger}$, $\psi(1/2) < \psi_s$, and the behavior of $\psi(z)$ is monotonic in the interval $0 \leq \frac{1}{2}$.[1]

The inverse Frederiks effect can also be interpreted in terms of a force law. The torque of the magnetic particles on the nematic matrix is not a monotonic function of the angle $\alpha = \pi/2 - (\psi - \theta)$ between the directors and reaches its maximum value at $\alpha \approx \pi/4$.

Finally, in this section, we note that it seems likely that the transition to a ferronematically distorted state actually occurs first through a first order transition at lower fields. Physical values of t are lower than unity; the result is that our analysis remains true in the high field limit but is modified in low field limit. Similarly, we

[1] Further computation shows that the change in the profiles $\psi(z)$ is connected with an ambiguity of the dependence $\theta(h,\psi)$ when $\psi_s > \pi/4$ or $h > h_b + \sqrt{2}w$. There are two branches, but only one of these satisfies the boundary condition $\theta_s = 0$. On this branch the angle θ decreases with h and hence with ψ, leading to the inverse Frederiks effect.

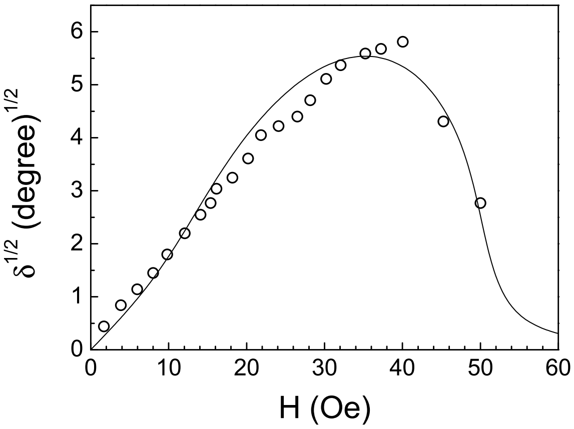


Fig. 14. *Experimental data by Chen and Amer* [3] *of the square root of the phase difference versus the magnetic field strength (open circles) and the theoretical curve obtained by the numerical minimization of the cell free energy equations* (1a) *and* (1b).

find that in dimensionless units the bias field required for the analysis presented here to hold true is $h_b < t$. Unsurprisingly, at low temperatures, a small bias field suffices to saturate the local magnetic order.

Figures 12(b) and 13 show the profiles in the magnetic field range where the van der Waals loop is developed. All profiles are symmetrical due to equal nematic anchoring strengths at the walls of the cell.

6. Experimental considerations. To make contact with the experiment [9, 17, 18, 21], we consider a cell of thickness $D = 125$ μm and suppose magnetic particles of $L = 0.2$ μm, $d = L/3$. Plausible mean packing fractions are in the range $\bar{f} \approx 2 \times 10^{-7} - 6 \times 10^{-5}$, with magnetite particles for which $\bar{M} = 485$ G. The quantity $W_p d$ is a quantity with the dimensions of an anchoring energy W. Naive microscopic theory suggests that this might be expected to be of the same order of magnitude as the anchoring energy at the surface of the colloidal particle. In this spirit, we expect $W_p\, d \approx 10^{-3} - 10^{-1}$ erg/cm^2. Finally, for 5CB at $T = 25^o$C, $K \sim 5.3 \times 10^{-7}$ dyn. The resulting dimensionless quantities are in the range $w \approx 9 \times 10^{-3} - 2.7 \times 10^{2}$ and $t \approx 3.5 \times 10^{-3} - 1$.

To obtain higher temperatures t, in order that real experiments more closely match our ideal systems, it would be necessary to use thicker cells and higher particle concentrations. For example, with a cell of thickness $D = 460$ μm, magnetic particles with $L = 0.15$ μm, $d = L/3$, and a mean volume fraction of $f = 6.9 \times 10^{-6}$ ($c = 2.3 \times 10^{10}$ cm^{-3}), we have $t = 3.85$.

The size and the aspect ratio we consider for the colloidal ferromagnetic particles are consistent with these particles being in a single-domain magnetic state [22]. We suppose that the magnetic grains are coated by suitable surfactant [3, 9, 23] to prevent coagulation. This theoretical idealization has in practice presented problems for experimentalists, but we do not address these here. Finally, we impose homeotropic boundary conditions at the particle surfaces.

We show in Figure 14 a fit of the experimental data by Chen and Amer [3] to our theory. The field-induced molecular reorientation of the FN was found in [3] by measuring the corresponding induced change in phase difference

$$\delta = \frac{2\pi}{\lambda}\int_0^D \left[\frac{n_o n_e}{(n_e^2\cos^2\theta + n_o^2\sin^2\theta)^{1/2}} - n_o\right] dz, \tag{70}$$

where λ is the wavelength of the incident laser beam and n_o, n_e are, respectively, the ordinary and extraordinary refractive indices of the sample. The continuous curve is calculated for $\lambda = 632.8$ nm, the width of an FN cell $D = 337$ μm, $H_b = 0.6$ Oe, $M_s = 340$ G (γ-Fe$_2$O$_3$ magnetic particles), $K_3 = 7.63 \times 10^{-7}$dyn, $K_1 = 6 \times 10^{-7}$dyn, $n_o = 1.5443$, $n_e = 1.7582$, and $\chi_a = 0.97 \times 10^{-7}$ (MBBA, $T = 25^o$C [17]). The parameters of the best fit are $L = 0.47$ μm, $L/d = 7.7$ (close to that in [3]), $W_p\, d = 4.9 \times 10^{-2}$ erg/cm^2, and $\bar{f} = 1.83 \times 10^{-8}$. For the above parameter values, $w \approx 0.22$ (low coupling regime). A clear decrease in δ for $H > 40$ Oe is in agreement with our prediction of an inverse Frederiks effect.

We note [10] that the ferronematic is expected to exhibit collective behavior only for colloidal particle concentrations exceeding a critical value w_{cr}, with $w_{cr} \sim 1/2$. However, our estimate for w is slightly below 1/2. There is thus an apparent paradox that results can be fitted to a model, but only with parameters for which the microscopic foundation of the model is weak. The solution to the paradox is not entirely clear, but it is possible to speculate. Among possible weaknesses in the model are (a) a failure to take account of the polydispersity in size of magnetic particles and (b) the use of a simple phenomenological Rapini–Papoular expression for the anchoring energy [10].

The best fit of the birefringence gives the value of mean particle concentration $\bar{c}$ as about 7% of the total concentration reported in [3]. We speculate that this may be related to a partial coagulation of the magnetic particles during the preparation of the ferronematic, i.e., a formation of large multiparticle aggregates whose total magnetic moment is close to zero [11]. These aggregates would be insensitive to a weak external magnetic field, thus reducing the birefringence effect.

7. Discussion. In this paper we have carried out an exhaustive analysis of the ordering processes which take place in a Frederiks-like cell when the liquid crystal is doped by magnetic colloidal particles. These systems are otherwise known as ferronematics. The original motivation for introducing the magnetic particles is to amplify the otherwise low magnetic response. The magnetic Frederiks transition would then be observable at experimentally accessible fields, and the effect could be utilized in magnetically switched liquid crystal devices.

Our calculations show that the simple amplification picture outlined in the last paragraph is at best a great simplification. The mathematical structure of these ferronematic systems seems extremely simple to formulate. There is also an interesting homogenization problem concerned with determining the magnitude of the effective magnetic-nematic director interaction which we have not addressed here, but which is under study elsewhere [25]. The simple formulation nevertheless exhibits a complex and rich set of behaviors as a function of magnetic field, colloidal particle structure, and colloidal concentration.

Although these behaviors can be analyzed easily using computational solutions, we have sought here where possible to examine the structural predictions of the model using mathematical tools. Without this perspective, the model predictions may look counterintuitive. A particular advantage of the method presented here, as opposed to earlier attempts at the same problem, is a change of scaling. The result of this apparently trivial scaling change is that temperature effects on the colloidal density can be added as a perturbation rather than present themselves as an essential element in the theory.

We find that low magnetic fields do indeed produce a switching analogous to the switching in Frederiks cells in an electric field. However, if the colloidal particles are insufficiently anisotropic, or if their volume fraction is too low, the alignment at low fields is followed by a disorientation process at high fields. In this circumstance, even though the magnetic particles are ordered by the field, the nematic order is dominated by the boundaries at high and low fields, although not at intermediate fields. We have further shown that a first-order disorientation transition is expected as a result of coupling between the orientation and segregation of the magnetic particles into regions where the magnetic force is highest.

The key difference between the ferronematic-induced effects and the direct magnetic effects is the fact that at high fields the magnitude of the ferronematic effects saturates. For very high fields, the director is more strongly tied to the magnetic field than it is to the nematic director. Our treatment completely ignores the direct interaction between the nematic and the magnetic fields. This is normally down by several orders of magnitude, but in the very high field limit, this will no longer be the case. In this very high field limit, our high field asymptotics would need to be modified. In extreme cases, where the direct and induced fields compete, there is further potentially interesting very high field physics. But we anticipate that this will occur at experimentally inaccessible magnetic fields. In any event, we postpone this to future work.

Finally, we note again the paucity of experimental data on these systems. This is a result of the difficulties of aligning magnetic particles themselves at higher temperatures, and also of preventing van der Waals forces from forcing irreversible colloidal aggregation. Attempts in the physics and engineering communities to make progress in this area continue. If these attempts bear fruit, there are further interesting mathematical problems to attack in this area. The most obvious of these is the dynamics of the ordering process itself, which, because of the nature of the couple exerted by the magnetic field on the local dipole moment, could in principle lead to a slow and oscillatory response. We postpone this problem to a future paper.

Acknowledgments. We acknowledge useful discussions over the course of this work with M. P. Allen, S. V. Burylov, M. Carme Calderer, D. Golovaty, I. P. Pinkevich, Y. Reznikov, C. Rosenblatt, and A. N. Vasilev.

REFERENCES

[1] F. Brochard and P. G. de Gennes, *Theory of magnetic suspensions in liquid crystals*, J. Physique (France), 31 (1970), pp. 691–708.

[2] C. Mauguin, *Orientation of liquid crystals by a magnetic field*, C. R. Acad. Sci., 152 (1911), pp. 1680–1683 (in French); translation appears in Crystals that Flow: Classic Papers from the History of Liquid Crystals, T. J. Sluckin, D. A. Dunmur, and H. Stegemeyer, eds., Taylor and Francis, London, 2004, pp. 122–127.

[3] S.-H. Chen and N. M. Amer, *Observation of macroscopic collective behavior and new texture in magnetically doped liquid crystals*, Phys. Rev. Lett., 51 (1983), pp. 2298–2301.

[4] S. K. Srivatsa and G. S. Ranganath, *Nematic kink states in a laser field*, Phys. Rev. E (3), 60 (1999), pp. 5639–5646.

[5] C. Y. Matuo and A. M. Figueiredo Neto, *Time dependence of the magnetic grain concentration and secondary grain aggregation in ferronematic lyotropic liquid crystals subjected to magnetic field gradients*, Phys. Rev. E (3), 60 (1999), pp. 1815–1820.

[6] P. Kopčanský, I. Potočová, M. Timko, M. Koneracká, A. M. G. Jansen, J. Jadzyn, and G. Czechowski, *The structural transitions in ferronematics in combined electric and magnetic fields*, J. Magn. Magn. Mater., 272–276 (2004), pp. 2355–2356.

[7] Yu. L. Raikher and V. I. Stepanov, *Transient field-induced birefringence in a ferronematic,* J. Magn. Magn. Mater., 201 (1999), pp. 182–185.

[8] V. Berejnov, J.-C. Bacri, V. Cabuil, R. Perzynski, and Yu. Raikher, *Lyotropic ferronematics: Magnetic orientational transition in the discotic phase*, Europhys. Lett., 41 (1998), pp. 507–512.

[9] O. Buluy, E. Ouskova, Yu. Reznikov, and P. Litvin, *Preparation and properties of a ferromagnetic nematic suspension*, Ukr. J. Phys., 49 (12A) (2004), pp. A48–A50.

[10] S. V. Burylov and Yu. L. Raikher, *Macroscopic properties of ferronematics caused by orientational interactions on the particle surfaces* I: *Extended continuum model*, Mol. Cryst. Liq. Cryst., 258 (1995), pp. 107–122.

[11] S. V. Burylov and Yu. L. Raikher, *Macroscopic properties of ferronematics caused by orientational interactions on the particle surfaces* II: *Behavior of real ferronematics in external fields*, Mol. Cryst. Liq. Cryst., 258 (1995), pp. 123–141.

[12] B. J. Liang and S.-H. Chen, *Electric-field-induced molecular reorientation of a magnetically biased ferronematic liquid-crystal film*, Phys. Rev. A (3), 39 (1989), pp. 1441–1446.

[13] S. V. Burylov, V. I. Zadorozhnii, I. P. Pinkevich, V. Yu. Reshetnyak, and T. J. Sluckin, *Weak anchoring effects in ferronematic systems*, J. Magn. Magn. Mater., 252 (2002), pp. 153–155.

[14] V. I. Zadorozhnii, I. P. Pinkevich, V. Yu. Reshetnyak, S. V. Burylov, and T. J. Sluckin, *Adsorption phenomena and macroscopic properties of ferronematics caused by orientational interactions*, Mol. Cryst. Liq. Cryst., 409 (2004), pp. 285–292.

[15] V. I. Zadorozhnii, A. N. Vasilev, V. Yu. Reshetnyak, K. S. Thomas, and T. J. Sluckin, *Nematic director response in ferronematic cells*, Europhys. Lett., 73 (2006), pp. 408–414.

[16] B. I. Lev, S. B. Chernyshuk, P. M. Tomchuk, and H. Yokoyama, *Symmetry breaking and interaction of colloidal particles in nematic liquid crystals*, Phys. Rev. E (3), 65 (2002), 021709.

[17] L. M. Blinov and V. G. Chigrinov, *Electrooptic Effects in Liquid Crystal Materials*, Springer-Verlag, New York, 1994.

[18] A. N. Zakhlevnykh, *Threshold magnetic fields and Fréedericksz transition in a ferronematic*, J. Magn. Magn. Mater., 269 (2004), pp. 238–244.

[19] R. H. Self, C. P. Please, and T. J. Sluckin, *Deformation of nematic liquid crystals in an electric field*, European J. Appl. Math., 13 (2002), pp. 1–23.

[20] E. G. Virga, *Variational Theories for Liquid Crystals*, Chapman and Hall, London, 1994.

[21] Z. Wang and C. Holm, *Structure and magnetic properties of polydisperse ferrofluids: A molecular dynamics study*, Phys. Rev. E (3), 68 (2003), 041401.

[22] L. L. Afremov and A. V. Panov, *Magnetic states and hysteresis properties of small magnetite particles*, The Physics of Metals and Metallography, 86 (1998), pp. 269–275.

[23] P. Poulin, V. Cabuil, and D. A. Weitz, *Direct measurement of colloidal forces in an anisotropic solvent*, Phys. Rev. Lett., 79 (1997), pp. 4862–4865.

[24] B. I. Halperin, T. C. Lubensky, and S. K. Ma, *First-order phase transitions in superconductors and smectic-A liquid crystals*, Phys. Rev. Lett., 32 (1974), pp. 292–295.

[25] M. Carme Calderer and D. Golovaty, *Private communication*, University of Minnesota, 2007.

SIAM J. APPL. MATH.
Vol. 68, No. 6, pp. 1717–1737

MODELS OF VIRULENT PHAGE GROWTH WITH APPLICATION TO PHAGE THERAPY*

HAL L. SMITH†

Abstract. We modify existing models of bacteriophage growth on an exponentially growing bacterial population by including (1) density dependent phage attack rates and (2) loss to phage due to adsorption to both infected and uninfected bacteria. The effects of these modifications on key pharmacokinetic parameters associated with phage therapy are examined. More general phage growth models are explored which account for infection-age of bacteria, bacteria-phage complex formation, and decoupling phage progeny release from host cell lysis.

Key words. phage therapy, infection-age structure, multiple adsorptions, bacteria-phage complex, passive therapy, active therapy, proliferation threshold

AMS subject classifications. 92D25, 34K60

DOI. 10.1137/070704514

1. Introduction.

As pathogenic bacteria have increasingly become resistant to our arsenal of antibiotics, there has been renewed interest in the use of bacteriophage to control bacterial infections [13, 11, 12, 14, 15, 23]. Bacteriophage, phage for short, are viruses which prey on bacteria. Almost as soon as they were discovered there was interest in using them to control infections and bacterial contamination. The history of early attempts to use them for such purposes during the last century is fascinating [13, 12, 6]. It is not hard to see the potential in phage therapy, for, unlike chemotherapy, which simply results in the death of a susceptible bacteria, phage therapy results in the death of the host cell and the release of hundreds more lethal phage. The author found the review articles [13, 12] on phage therapy useful.

Mathematical modeling has long played a significant role in the study of phage-bacterial interactions for ecological reasons [2, 22, 17, 9] as well as for medical ones [3, 11, 12, 10, 14, 15, 23]. See also the additional references in these papers. The reasons for this are obvious—among them being the difficulty of carrying out controlled experiments in vivo and the novelty of a self-replicating therapeutic agent.

It will be useful to briefly review the life cycle of a virulent phage and some of the associated terminology for later use. Typically, phage specialize to attack only one or a few strains of a bacterial host whose cell surface contains an appropriate binding site. Phage attach to a preferred binding site and then inject their genetic material, DNA or RNA depending on the phage, and perhaps some enzymes into the cell, which thereafter is called an infected cell. In the case of virulent (also called lytic) phage, the infected host cell machinery is then immediately co-opted to make new phage particles which are subsequently released in a burst when phage enzymes cause the host cell to lyse and the cycle repeats. The latent period is the time between phage-host binding and subsequent release of the phage at cell lysis, usually on the order of 20 minutes to an hour depending on the host-phage system. The burst size, ranging between several to thousands, is a measure of the average number of phage

[*] Received by the editors October 4, 2007; accepted for publication (in revised form) April 4, 2008; published electronically July 2, 2008. This work was supported by NSF grant DMS 0414270.
http://www.siam.org/journals/siap/68-6/70451.html

[†] Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (halsmith@asu.edu).

progeny resulting from a single infected host cell and also depends on the host-phage system.

This paper addresses some issues that seem not to have been explored in the mathematical modeling of bacteriophage growth that may be important in phage therapy. First, one finds that mass action kinetics, e.g., bxv, where x denotes the concentration of uninfected bacteria and v is the concentration of phage, is invariably used to model both the phage attack rate on uninfected bacteria and the rate of loss of free phage due to attachment [9, 10, 14, 15, 23]. Weld, Butts, and Heinemann [23] experimentally measured the "adsorption rate" b in the context of the rate of loss of phage (number of adsorbed phage per free phage per bacterium per minute), found a wide variation of values, and noted that it decreased as the sum of phage and bacterial densities $(x + y + v)$ increased, where y is the density of infected bacteria.

In this paper, we will argue that the phage attack rate and the rate of phage loss due to attachment are distinct. As the former involves attachment and injection of the one primary (first to inject) phage while the latter takes account of all secondary phage that attach to a cell, this should not be unexpected. We propose that the phage attack rate deviates from bxv when phage densities are large due to the higher likelihood of multiple phage binding to a cell between the time of initial binding and lysing and therefore to a lower impact per phage particle. Mathematically, this will be achieved here not by making b dependent on the densities but by adding an extra multiplicative term to the phage attack rate which depends on the phage density. Specifically, our analysis leads to the reduced phage attack rate

$$\frac{bxv}{F_N(cv)}, \quad c = b/\rho, \tag{1.1}$$

where $1/\rho$ denotes the injection time, the time between binding of a phage to a host bacteria and subsequent injection of genetic material into the host, N denotes the number of binding sites for phage per host, and

$$F_N(u) = 1 + \frac{u}{1+u} + \frac{u^2}{(1+u)(2+u)} + \cdots + \frac{u^N}{(1+u)(2+u)\cdots(N-1+u)N}.$$

Observe that $F_N(u) > 1$, so the attack rate is strictly less than bxv. Despite appearances, $F_N(u)$ depends rather weakly on N; $F_3(u)$ is a good approximation of $F_{100}(u)$ on $0 < u < 5$. To lowest order, $F_N(u) \approx 1 + u$, so the effect of the term cv in (1.1) is nonnegligible precisely when $bv \times \frac{1}{\rho}$ is nonnegligible compared to one. bv/ρ gives the number of potential irreversible phage attachments that could be formed with a typical host cell during the injection time. A rough estimate of $c = O(10^{-8})$ for a strain of *E. coli* and phage implies that the term cv in (1.1) is significant when $v \geq O(10^7)$, well within the range used in experimental and theoretical studies.

Second, it seems to us that the rate of loss of phage is underestimated in existing models and moreover that the effect of this underestimation may be significant for bacteriophage therapy. For example, if we assume that a phage cannot detect the state (uninfected or infected) of the host cell to which it binds, then one should not ignore the loss of the phage due to "wasted attacks" on already infected hosts. We take into account that a host cell has a multiplicity of potential phage binding sites on its surface, more than one of which may be simultaneously bound by phage. This leads, with good approximation, to the expression

$$-bv(x + y) \tag{1.2}$$

for the rate of phage loss due to attachment. As a result our model differs from others in the literature in that the phage loss rate due to attachment differs from the phage attack rate on the host.

Finally, most existing models either assume an exponentially distributed latent period [14, 15] leading to ordinary differential equations or assume a fixed-length latent period which results in delay differential equations [9, 10, 23]. Here, in the appendix, we explore a more general model, where infected cells are structured by age-since-injection and where the release rate of phage progeny may be either a continuous "budding off" or an abrupt burst at host lysis typical of virulent phage. We also allow for variable phage progeny size by decoupling cell death from the release of progeny. This structured model may lead to ordinary differential equations, to delay equations, or to more general integro-differential equations depending on whether the infected cell mortality rate is independent of age-since-injection, sharply dependent on it, or a more smooth nonconstant function of it. However, much of our effort is devoted to a delay differential equation model resulting from the assumption of a fixed-length latent period followed by a discrete burst of phage. This model is similar to models considered by Lenski and Levin [10] and by Beretta and Kuang [2] except for the modifications already noted. Our treatment of the initial conditions and their effect during the initial latent period is more natural than in [10, 2], and it facilitates the consideration of a proliferation threshold for phage therapy.

We show that the density dependent attack rate (1.1) and the modified rate of phage loss due to attachment (1.2) result in potentially significant modifications in key pharmacokinetic quantities associated with active phage therapy first identified by Payne and Jansen [14, 15]. Unlike passive therapy which relies on a massive dose of phage to kill bacteria in only one phage generation, active therapy does not need such a large dose since it relies on second and third generation phage for its success. Payne and Jansen argued for the existence of a threshold number of uninfected host cells required to support the amplification in phage numbers from one phage generation to the next (the phage reproductive number exceeds one). This idea has generated some controversy in the field [7, 8, 13, 16] due to a misunderstanding of its meaning. However, it is certainly valid for the mathematical models of phage growth treated in [14, 23]. We derive a threshold condition for phage proliferation which reduces to one comparable to Payne and Jansen's when total bacterial density is not too large but changes character when densities become large.

2. Phage growth on an exponentially growing host population: Fixed-length latent period. In the appendix we derive a general model of phage growth which includes the one described below as a special case. Here, we make the following assumptions:

(a) Bacteria first injected by phage at time $t-\tau$ lyse (die) at time t.

(b) Uninfected host cells grow at rate a; infected cells do not grow.

(c) Free phage, those unattached to host cells, decay or wash out at rate m.

(d) Bacteria are removed by washout or death unrelated to phage at rate p.

(e) Phage do not distinguish between infected and uninfected cells with regard to attachment and injection.

(f) The rate of release of phage progeny from an infected host of infection-age (time since injection) $s \in [0, \tau]$ is $\eta(s)$. In particular, it is independent of the number of phage injections.

In other words, we assume the latent period has duration precisely τ units of time as in [2, 23, 10]; this is relaxed in the more general model in the appendix. As we are primarily motivated by applications to phage therapy where one is interested in treating the initial phase of a bacterial infection which, in the worst case, is characterized by an exponentially growing pathogen, the model equations do not include density effects on host growth such as in [10, 2]. Payne and Jansen [14, 15] assume that infected hosts grow at the same rate as uninfected hosts, but we follow Weld, Butts, and Heinemann [23] and Abedon, Herschler, and Stopar [1], who suggest that infected host cells do not grow. According to (f), the integral

$$L = \int_0^\tau \eta(s)ds$$

gives the phage progeny from an infected host assuming that it survives the latent period. The general release rate $\eta(s)$ easily accommodates both what is sometimes called "budding" of phage progeny from a living host cell as well as a "burst" of phage progeny released at cell lysis. Assumption (e) is used by Schrag and Mittler in [18] in an ecological setting. We are unaware of any evidence supporting or refuting it. Our assumption in (f) that the number of progeny is independent of the number of phage injections is consistent with observations of Stent [20, p. 74], who remarks that "latent period and burst size do not, however, depend in any very striking way on the number of phage particles with which each bacterial cell has been infected." However, he later mentions lysis inhibition that occurs in the infected host of certain T-even phage where superinfection late in the latent stage may substantially prolong the latent stage and enhance the burst size. See also [1]. Hypothesis (d) is rather standard [14, 23].

Let x denote the density of uninfected host bacteria, y the density of phage-infected bacteria, and v phage density. The expression (1.1) is used for the phage attack rate rather than the usual mass action rate bxv for the reasons described in the introduction. The corresponding loss rate for phage due to attachment to host cells is modeled by (1.2) to account for wasted attachments following (e) above. We stress that both these rates are derived in the appendix.

Initial conditions at time $t = 0$ must take account of the infection-age of the infected cells since these cells may have been infected at different times in the past. Moreover, it is reasonable to assume that the initial set of infected cells was obtained in a manner independent of the initial set of uninfected cells and phage. Therefore, we prescribe the initial uninfected host density $x(0)$, the initial phage density $v(0)$, and the initial distribution of infected cells:

$$U_0(s), \ 0 \le s \le \tau,$$

where s denotes the age since infection (injection). Thus, for $0 < c < d < \tau$, $\int_c^d U_0(s)ds$ denotes the number of infected cells with infection-age between c and d. These cells, infected at times between $t = -d$ and $t = -c$, will, if not washed out, lyse at times between $\tau - d$ and $\tau - c$.

In our view, the manner in which initial data are formulated here is more natural than that in [2], where the past history of phage and uninfected host cells are prescribed over a latent period, assuming that the initial population of infected cells at $t = 0$ arises through infection of the initial host population by the initial phage.

For the initial latent period, the equations of the model are given by

$$
\begin{aligned}
&\frac{dx}{dt} = ax - \frac{bxv}{F_N(cv)} - px,\\
&y(t) = \int_0^t \frac{bx(t-s)v(t-s)}{F_N(cv(t-s))} e^{-ps} ds + e^{-pt} \int_0^{\tau - t} U_0(s) ds, \quad 0 \le t \le \tau,\\
&\frac{dv}{dt} = -b(x+y)v - mv + \int_0^t \frac{bx(t-s)v(t-s)}{F_N(cv(t-s))} e^{-ps} \eta(s) ds\\
&\qquad + e^{-pt} \int_0^{\tau - t} U_0(s) \eta(s+t) ds.
\end{aligned} \tag{2.1}
$$

Observe that infected cells at time $t < \tau$ arise either from cells infected after $t = 0$ or from surviving cells from the founding population U_0. The first integral in the equation for y gives the number of cells infected after $t = 0$ which survive washout to be alive at time t. The second gives the survivors from the founding population still alive at time t; clearly these must have had age less than $\tau - t$ in order to be alive at time t. A similar analysis explains the two integrals in the equation for phage: the first integral gives phage progeny issuing from cells infected after $t = 0$, and the second integral gives phage progeny issuing from survivors from the founding population of infected cells.

In all our simulations and in most experimental setups, the founding population of infected cells is taken to vanish, $U_0 = 0$, simplifying the equations during the initial latent period. Indeed, the natural initial conditions for phage therapy correspond to administering a dose of phage to an exponentially growing bacterial population that has not been exposed to phage, and therefore there are initially no infected hosts of any age of infection. This does not mean that our care in formulating the initial conditions is wasted since we will have reason to consider nonzero U_0 in our later discussion related to phage therapy.

After the latent period, the founding population of infected cells have all lysed and the equations are simpler:

$$
\begin{aligned}
&\frac{dx}{dt} = ax - \frac{bxv}{F_N(cv)} - px,\\
&y(t) = \int_{t-\tau}^t \frac{bx(s)v(s)}{F_N(cv(s))} e^{-p(t-s)} ds, \quad t > \tau,\\
&\frac{dv}{dt} = \int_0^\tau e^{-ps} \frac{bx(t-s)v(t-s)}{F_N(cv(t-s))} \eta(s) ds - b(x+y)v - mv.
\end{aligned} \tag{2.2}
$$

Key properties of the function $F_N(u)$ are given in the lemma below, which is proved in the appendix. As Figure 2.1 shows, these functions depend rather weakly on N for small u.

LEMMA 2.1. *The functions* $F_N : [0, \infty) \to [1, \infty)$ *satisfy the following:*

1. $\frac{d}{du} \frac{u}{F_N(u)} > 0$ *and* $\frac{d}{du} F_N(u) > 0$,
2. $\frac{u}{F_N(u)} \to N$ *as* $u \to \infty$, *and*
3. $F_\infty(u) \le F_{N+1} \le F_N(u) \le F_1(u) = 1 + u$,

where

$$
F_\infty(u) = 1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{u}{i+u}.
$$

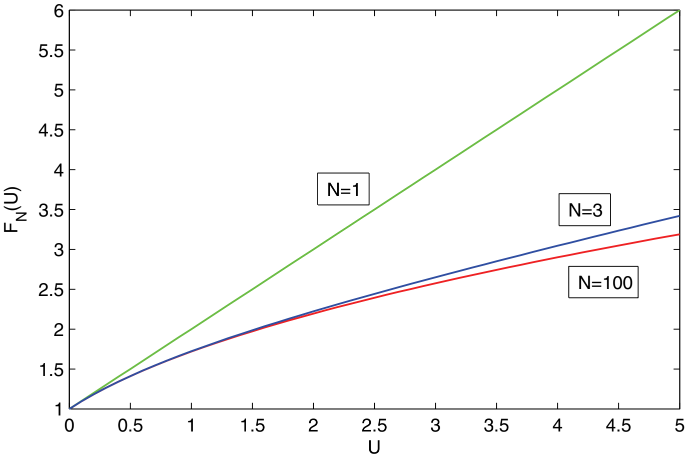


FIG. 2.1. *Plot of $F_N(u)$ versus u for $N = 1, 3, 100$.*

The biological implications of these assertions are not surprising. The fact that $F_N(u) > 1$ for $u > 0$ means that the attack rate (1.1) is smaller than the corresponding mass action rate bxv. According to assertion 3, it increases with N, the number of host binding sites; more sites means the potential for more attached phage, which decreases the time to injection. The first assertions ensure that the phage attack rate increases with increasing phage density v. The second indicates that the maximum effect of phage on the specific growth rate of bacteria (x'/x) is bN/c. Observe that $F_3(2) \approx F_{100}(2) > 2$, so the attack rate is less than half the corresponding mass action rate bxv when $cv > 2$.

For $t > \tau$, the y equation can be differentiated to yield

$$y' = \frac{bxv}{F_N(cv)} - e^{-p\tau}\frac{bx_\tau v_\tau}{F_N(cv_\tau)} - py, \tag{2.3}$$

where $x_\tau = x(t-\tau)$, $v_\tau = v(t-\tau)$. As noted above, the natural initial conditions for phage therapy are to administer a dose of phage to an exponentially growing bacterial population which has not previously been exposed to phage. Thus, $U_0 \equiv 0$ and the differentiated form of the y equation during the initial latent period differs from (2.3) in that the second term is removed.

Well-posedness issues related to our system (2.1)–(2.2) can be treated using results in [5]. Corollary 2.2 of Chapter 12 can be used to prove the existence and uniqueness of a maximally defined continuous solution corresponding to nonnegative initial data if, for example, U_0 is integrable and η is essentially bounded. Easy arguments give that the solution is nonnegative, and prior bounds show that it is globally defined.

If the rate of phage progeny release is highly peaked about the age at lysis τ, it is reasonable to assume that cells produce L-phage exactly on reaching age τ:

$$\eta(s) = L\delta(s-\tau), \tag{2.4}$$

where $\delta(r)$ is the Dirac impulse function with unit mass concentrated at $r = 0$. In that case, the equation for phage simplifies for the latent period to

$$\frac{dv}{dt} = -b(x+y)v - mv + Le^{-pt}U_0(\tau - t), \quad 0 < t \le \tau, \tag{2.5}$$

and thereafter to

$$\frac{dv}{dt} = Le^{-p\tau}\frac{bx_\tau v_\tau}{F_N(cv_\tau)} - b(x+y)v - mv, \quad t > \tau. \tag{2.6}$$

2.1. Remarks on long term dynamics. In contrast to ecologically motivated theoretical studies of phage growth where attention has been paid to long term dynamics [2, 17], there has been very little consideration of long term dynamics of phage growth models aimed at understanding phage therapy [14, 23]. This is probably due to a tacit assumption that the equations are valid only until such time as an immune response is mounted. Indeed, the hypothesis of exponential bacterial growth captures this focus on the short term dynamics.

Returning to the general system (2.1)–(2.2) with budding rate η, if $a > p$, which we assume hereafter, then the trivial equilibrium point $(x, y, v) = (0, 0, 0)$ is an unstable saddle point. For if there are no virus or infected cells, then the uninfected cells grow at the exponential rate $a - p > 0$. If, on the other hand, there are no uninfected cells, then any free virus and infected cells are removed at an exponential rate. Therefore, it is plausible that no solution starting with $x(0) > 0$ can satisfy $x(t) \to 0$, $t \to \infty$. We wish to stress that this fact is common to all models of virulent phage growth in the literature—not just the one treated here. It is possible to overlook this observation on viewing the simulations reported here and in [14, 23]. Below we give a proof which carries over to these other models with only minor changes.

Proposition 2.2. *If $a > p$, then no solution of* (2.2) *with* $x(0) > 0$ *can satisfy* $x(t) \to 0$, $t \to \infty$.

Proof. The assertion is obvious if $bN/c \le a - p$ since then $x'/x \ge 0$. Hereafter, assume that $bN/c > a - p$. If $x(t) \to 0$ as $t \to \infty$ for some nonzero solution of (2.2), then necessarily $y(t) \to 0$ as well. It can then be seen that v must remain bounded. By standard arguments, this and the convergence of $x(t)$ imply that $x'(t) \to 0$ and, therefore, since $x(t) \ne 0$, $v(t) \to V_I > 0$, where V_I is the unique positive root of $\frac{bv}{F_N(cv)} = a - p$ guaranteed by Lemma 2.1. The convergence of $v(t)$ implies $v'(t) \to 0$, but this immediately leads to the contradiction that $v \to 0$. □

Proposition 2.2 does not preclude that bacteria levels fall below a singleton, $x(t) < O(1)$, which means extinction and hence successful treatment. This will be evident in our numerical simulations. Obviously, our deterministic model breaks down at low densities of bacteria, but aside from this therapy can be successful if, even with large initial bacterial populations, there are feasible phage doses that will result in bacteria levels decreasing to a small fraction of pretreatment levels.

Because we assume that bacteria grow exponentially, rather than, say, logistically as in [2] or controlled by nutrient limitation as in [11, 10], there is no "phage-free" equilibrium with host cells at some positive level and no phage or infected cells.

If $\frac{bN}{c} > a - p$ and

$$(a-p)\int_0^\tau e^{-ps}\eta(s)ds > bV_I\left(1 + (a-p)\tau\frac{(1-e^{-p\tau})}{p\tau}\right), \tag{2.7}$$

where the quotient $\frac{(1-e^{-p\tau})}{p\tau} = 1$ when $p = 0$, then there exists a unique positive steady state (X, Y, V). The analytical determination of the stability properties of the positive steady state via linearization is very challenging. See Beretta and Kuang [2] for some partial results. Simulations of our system extending for hundreds of hours, not shown here, are oscillatory in nature and characterized by long periods with cell

Table 2.1
System parameters.

Parameter	Value
a	0.3 /hr
b	$9 \times 10^{-7} ml/hr$
c	$3 \times 10^{-8} ml$
L	150
m	1.8 /hr
p	0
τ	.43 hr
N	100
ρ	30/hr

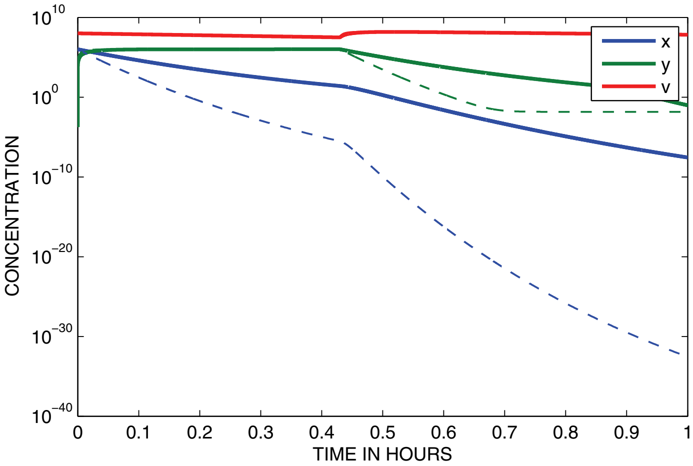


Fig. 2.2. $x(0) = 10^6$, $y(0) = 0$, $v(0) = 10^8$. *Dashed lines correspond to* $c = 0$.

densities well below one cell per milliliter. We conclude that the asymptotic behavior of the model system has limited relevance for phage growth.

2.2. Parameter values for simulations. Measurements of Weld, Butts, and Heinemann [23] found that the adsorption rate for phage T4 growing on an *E. coli* strain averaged 1.5×10^{-8} ml/min, which translates to the per hour rate in Table 2.1. Similar values can be found in [14, 20].

Grayson et al. [4] observe that the waiting time following phage binding to the host for initiation of injection is random, ranging from seconds to minutes. Their data reasonably fit an increasing but saturating exponential function with time constant t_0 which ranges from 79 seconds to 166 seconds. As they find that the injection process takes roughly 10 seconds, it is reasonable to take our injection time $1/\rho$ to be 2 minutes.

This leads to a value $c = b/\rho = 3 \times 10^{-8}$ ml, which means that the term cv appearing in (1.1) is significant; i.e., $cv = O(0.1)$ when $v > 0.3 \times 10^7$. In the numerical simulation shown in Figure 2.2, we take $v(0) = 10^8$ (compare with Payne and Jansen [14] who use even 10^9 initial phage in simulations), so $cv(0) = 3$.

In his monograph, Stent [20] describes experiments of Schlesinger, who measured the "adsorption capacity" of *E. coli* for WLL phage by adding phage to a suspension of bacteria and noting when additional phage could no longer become attached. He found that this threshold occurred at about 300 phage per bacterium. Although the adsorption capacity may differ from the number of binding sites for a variety of reasons, the experiment suggests that the number of binding sites satisfies $N = O(10^2)$. The lack of sensitivity of the attack rate (1.1) to N suggests that $N = 100$ should give sufficient accuracy.

Parameter values used in our simulations are displayed in Table 2.1. Those not described above were taken from [14, 23]. Our simulations are restricted to the special case that a burst of L phage is produced from an infected cell at lysis. Therefore, (2.5) and (2.6) were used. Initial data were taken following Payne and Jansen.

Assuming that $U_0(s) \equiv 0$, the simulated system is

$$\begin{aligned} \frac{dx}{dt} &= ax - \frac{bxv}{F_N(cv)} - px, \\ \frac{dy}{dt} &= \frac{bxv}{F_N(cv)} - H(t-\tau)e^{-p\tau}\frac{bx_\tau v_\tau}{F_N(cv_\tau)} - py, \\ \frac{dv}{dt} &= H(t-\tau)Le^{-p\tau}\frac{bx_\tau v_\tau}{F_N(v_\tau)} - b(x+y)v - mv, \end{aligned} \tag{2.8}$$

where $H(t-\tau)$ denotes the Heaviside function and $y(0) = 0$ (reflecting that $U_0 \equiv 0$) unless mentioned otherwise. Numerical solutions were computed using the dde23 delay differential equation solver on MATLAB.

Thick lines in Figure 2.2 display the time series simulating phage therapy over an hour time frame. A notable feature of the time series is the sharp discontinuity in the derivative of solution components at the time of the first latent period, roughly 26 minutes or 0.43 hours, caused by the burst of fresh phage. Cell concentrations below one cell per milliliter should be viewed as the absence of cells; this occurs for uninfected cells at approximately 0.6 hours. For comparison purposes, the dashed lines show the result of replacing both the attack rate (1.1) and rate of phage loss (1.2) by the traditional mass-action term bxv, keeping initial data the same. Note that the traditional mass-action rate results in a substantially quicker reduction in uninfected hosts over the first two latent periods and a reduction in infected cells over the second latent period. Uninfected cells essentially vanish at 0.2 hours and infected cells at 0.6 hours. Virus levels appear to be affected to a lesser degree. We conclude that the effect of our modifications in phage attack rate and phage loss rate is to significantly lengthen host cell survival, at least for initial data used here. As viral densities appear to be less affected by our modifications, it is probably the case that the modified attack rate is most significant.

3. Implications for phage therapy.

Payne and Jansen [14, 15] discovered some key pharmacokinetic parameters related to in vivo phage therapy against bacterial infection using a very simple model of phage growth. Of course, effects of an immune response to phage and to bacteria are ignored for these calculations, assuming that they kick in at a later time.

As noted above, in the context of our deterministic model, successful phage therapy can at best mean that bacterial levels are driven to a sufficiently low level that the immune system easily finishes them off. Successful phage therapy requires at minimum that uninfected bacterial density decreases. Our equations imply that

$$x' < 0 \Leftrightarrow v > V_I, \tag{3.1}$$

where V_I is the unique positive root of $\frac{bv}{F_N(cv)} = a - p$ (see Lemma 2.1). Payne and Jansen [14] refer to their V_I, obtained by setting $c = 0$ in ours, as the "inundation threshold" value. For the parameter values in Table 2.1, $V_I = 3.3 \times 10^5$.

Note that there is no guarantee, if one starts out with (3.1) holding, that it will continue to hold. In fact, Proposition 2.2 implies that $x' < 0$ cannot hold indefinitely. Therefore, merely arranging for initial phage densities to exceed the inundation threshold does not ensure "successful treatment."

Payne and Jansen identify two phage therapy strategies: passive therapy and active therapy. Passive therapy is the attempt to substantially knock down the bacterial population with the initial dose of phage, ignoring contributing effects of subsequent phage generations. Active therapy, presumably requiring a much smaller initial dose of phage, relies on the proliferation potential of phage reproduction to build up phage densities to levels sufficient to eventually drive down bacterial levels.

3.1. Passive therapy. Payne and Jansen's derivation of an explicit minimal phage dose for successful passive therapy, i.e., reaching $x(t) = O(1)$, in [14] relies on the simplicity of their model. Indeed, there are excellent reasons for simple models, and the ability to perform explicit calculations is one of them. Our model, which includes additional features such as the unproductive loss of phage due to phage attacking uninfected bacteria that are already bound to phage and to wasted phage attacks on infected cells, does not permit easy calculations. These additional features are most likely to be nonnegligible at the required large phage doses used in passive therapy.

It is not clear that the notion of a minimal dose for passive therapy is well defined for our model. Here, we take passive therapy to mean that bacterial density is reduced to a suitably small fraction of its initial size *within the initial latent period*. We take this obviously restrictive view of passive therapy only for analytical convenience; it is not, however, at great variance from the rather subjective definition of passive therapy—not to rely on subsequent generations of phage for success.

It seems obvious on biological grounds that if we fix the initial uninfected bacterial density $x(0)$ and assume as usual that $y(0) = 0$ but vary the initial phage dose $v(0)$, then larger doses will lead to smaller uninfected bacterial levels. However, this is far from obvious from a mathematical viewpoint. The next result establishes this point and provides a theoretical basis for the concept of a minimal dose for passive therapy.

PROPOSITION 3.1. *Let $(x(t), y(t), v(t))$ and $(\bar{x}(t), \bar{y}(t), \bar{v}(t))$ be two solutions of (2.8) with $x(0) = \bar{x}(0)$ and $y(0) = \bar{y}(0) = 0$. If $v(0) < \bar{v}(0)$, then $\bar{x}(t) < x(t)$ and $\bar{x}(t) + \bar{y}(t) < x(t) + y(t)$ on $0 < t \leq \tau$.*

For every $\theta \in (0, 1)$ and every $U > 0$ there exists $V = V(U) > 0$ such that if $(x(t), y(t), v(t))$ is a solution of (2.8) with $x(0) + y(0) \leq U$ and $v(0) \geq V$, then

$$\frac{x(\tau)}{x(0)} \leq e^{(a-p-(bN/c)\theta)\tau}. \tag{3.2}$$

In words, bigger phage doses result in smaller bacteria levels over the first latent period. In addition, *for any* initial bacterial density, there is a phage dose which will reduce bacterial density at the end of the first latent period by a factor that can be taken as close to $e^{(a-p-(bN/c))\tau}$ as desired. Recall that we are assuming $a - p - (bN/c) < 0$. Indeed, for the parameters of Table 2.1 it is approximately -1290. Therefore, $e^{(a-p-(b/c)\theta)\tau} \approx \exp(-1290)$ if $\theta \approx 1$. We do not claim that the required doses are medically feasible or even that there are the required number of phage in the universe. The proof of Proposition 3.1 is provided in the appendix.

3.2. Active therapy. In their consideration of active treatment, Payne and Jansen give an intuitive argument for the existence of a threshold bacterial density such that phage numbers are amplified over each successive phage generation only if bacterial density exceeds threshold. We can apply this intuitive reasoning to our model as well, although we arrive at a somewhat different threshold for phage amplification. According to our delay model, a cohort of y_0 newly infected cells, i.e., $U_0(s) = y_0\delta(s)$, where δ is the Dirac impulse concentrated at zero so all cells have infection-age zero, gives rise to $y_0 Le^{-p\tau}$ phage τ units of time later, when the surviving members of this cohort of the infected host have lysed. These phage survive $\frac{1}{b(x+y)+m}$ hours during which they infect the host at rate $\frac{bx}{F_N(cv)}$ per phage. We conclude that the $y_0 Le^{-p\tau}$ phage produce $Le^{-p\tau}\frac{bx}{(b(x+y)+m)F_N(cv)}y_0$ second generation infected cells. The amplification factor (of y_0) is therefore

$$R_0 = Le^{-p\tau}\frac{bx}{(b(x+y)+m)F_N(cv)}.$$

The condition for amplification of infected cells, and hence the condition for proliferation of phage, is that $R_0 > 1$:

$$Le^{-p\tau}\frac{bx}{(b(x+y)+m)F_N(cv)} > 1. \tag{3.3}$$

Unlike the proliferation threshold derived in [14], ours does not lead to a threshold condition for uninfected bacteria alone. However, as active therapy should not require such large phage doses that cv is significant, it may be reasonable in some cases to assume $cv \ll 1$, in which case $F_N(cv) \approx 1$ and we may ignore this factor. For parameter values in Table 2.1, if $v < 10^7$, this is a good approximation. Immediately below, we assume this approximation is valid.

Our proliferation condition involves both infected and uninfected hosts. If $b(x+y)$ is small relative to m, then we arrive at a threshold bacterial density for phage proliferation that is comparable to the one obtained in [14] and corrects the one given in [23]:

$$x > X_p \approx \frac{m}{bLe^{-p\tau}}. \tag{3.4}$$

However, if $b(x+y)$ is large relative to m, then (3.3) yields a threshold on the fraction of uninfected cells

$$Le^{-p\tau}\frac{x}{x+y} > 1.$$

In summary, the proliferation threshold computed from our model is more complex than that of Payne and Jansen since it involves all three quantities x, y, v. If $cv \ll 1$, it reduces to one similar to theirs, a lower bound on uninfected bacteria, when the total bacterial density is not too large but gives a lower bound for the fraction of uninfected bacteria when the total bacterial density is large. For parameter values in Table 2.1, $\frac{m}{bLe^{-p\tau}} \approx 13,500$; $b(x+y) \ll m$ if $x + y < 10^4$.

We provide two simulations where active therapy can be clearly identified. Initial data in Figure 3.1 yield a value $R_0 \approx 7$ well above unity. Bacteria initially grow before the phage gain control after the initial latent period. Figure 3.2 can be compared to Payne and Jansen's Figure 1(c) in [14], which shows passive therapy. We chose parameter values to correspond to those in that figure: $a = 0.3$, $b = 10^{-6}$, $\tau = 0.83$,

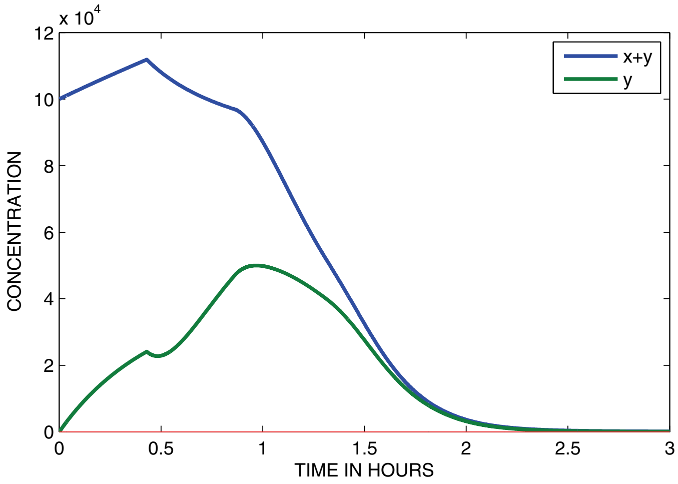


Fig. 3.1. *Active therapy with* $x(0) = 10^5$*,* $v(0) = 10^6$*.*

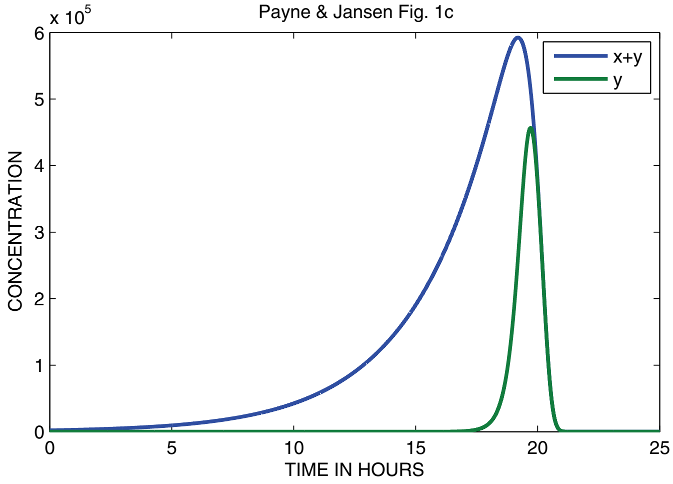


Fig. 3.2. *Compare to* [14, *Figure* 1(c)]*:* $x(0) = 2117$*,* $v(0) = 100$*.*

$L = 100$, $m = 1.8$, $p = 0$ with our value of N, and $c = b/\rho = 0.33 \times 10^{-7}$. Initial data were chosen as in their figure as well except that our starting time corresponds to their t_ϕ. It must be kept in mind that our model differs from theirs in the attack rate as well as our assumption that infected hosts do not grow. Our simulation agrees qualitatively with theirs, although our peak bacterial density exceeds theirs by a factor of three.

3.3. Summary of conclusions for phage therapy. We have explored a mathematical model of virulent phage growth on an exponentially growing bacterial population, system (2.8), which differs from previous models in two ways: (i) the density-dependent phage attack rate (1.1) is used in place of the mass action rate, and (ii) the loss rate of phage due to attachment (1.2) includes all bacterial cells and not just uninfected ones. Our simulations, particularly Figure 2.2, appear to be most strongly influenced by the modification (i), leading to significantly lengthened host cell survival. However, (ii) played a role in the calculation of the proliferation threshold (3.3) for active therapy which has a more complicated character than the simpler one deduced in [14]. Finally, in Proposition 3.1 we established that passive therapy works: for any initial bacterial population there is a phage dose that can reduce the host population to an insignificant fraction of its initial level within the first latent period. Although the dose may be impractically large and the restriction to a single latent period unduly restrictive, our result establishes the principle. It remains to be seen whether any of these effects are important for phage therapy.

In the following appendix, we show how (i) and (ii) arise from a careful modeling of bacteria-phage complexes consisting of a single host and a number of attached phage. The modeling framework we introduce there may be more important in the long run than the conclusions described above since it leads to much more flexibility in modeling phage release and host cell survival.

Appendix. We explore several modeling issues in this appendix which may be of interest for general bacteriophage-host interactions. First, we construct a model where infected hosts are structured by injection-age and where various complexes consisting of a host cell and one or more attached phage are explicitly included. We assume that each cell has a fixed number N of binding sites where host cells can have multiple attached phage. Finally, the age-structured infected cell population is reduced to integro-differential equations. Two cases lead to relatively simple equations: (1) the case where infected cell death rate by lysis is zero until time τ after which it is infinite, equivalent to fixed-length latent period, and (2) the case where the lysis rate is constant meaning an exponentially distributed latent period. The first case leads to the model considered in previous sections, while the second case leads to simpler ordinary differential equations similar to those in [14].

A.1. Phage-host complex formation. We call a host cell infected when a phage has injected genetic material into it; until then it is called uninfected. Furthermore, a phage ceases to exist once it has injected its genetic material.

The time between injection and lysis is on the order of 20 minutes or so, depending on the host-phage system. We will keep track of this "infection age" of infected cells. Denote by $Y(t,s)$ the distribution of infected cells of age s at time t. The integral

$$\int_{a_1}^{a_2} Y(t,s)ds$$

then gives the number of infected (postinjection) cells that were injected between $t-a_1$ and $t-a_2$, and

$$Y = \int_0^\infty Y(t,s)ds$$

gives the size of the infected class of cells.

Assume that each host cell has N potential phage binding sites. Then we partition uninfected host X, infected host Y, and phage V as follows:

$$X(t) = \sum_i C_x^i(t),$$
$$Y(t,s) = \sum_i C_y^i(t,s),$$

$$V(t) = v(t) + \sum_i i \left(C_x^i(t) + \int_0^\infty C_y^i(t,s)ds \right).$$

C_x^i (C_y^i) denotes the concentration of uninfected (infected) cells having i attached phage for $0 \leq i \leq N$, and $v(t)$ denotes the concentration of unadsorbed phage. In the case of infected cells, we stress that the age variable s denotes time since injection and not time in a particular compartment. All sums are over the range $0 \leq i \leq N$.

Recall that a host cell is infected once an attached phage injects, and a phage ceases to exist once it injects.

We assume that a host complex with $i < N$ attached phage may adsorb an additional phage at the rate bvC_z^i, where $z = x, y$. The rate constant b, the adsorption rate, is assumed independent of i and whether the host complex is infected ($z = y$) or uninfected ($z = x$). Let $\nu(s)$ denote the death (lysis) rate of infected cells of age s, and let $\eta(s)$ denote the rate of release of phage from an infected cell of age s. Here, we explore the case that both are independent of the number of attached phage. Recall that ρ is the injection rate; equivalently, $1/\rho$ is the average time between phage binding and subsequent injection of genetic material. The model equations are as follows:

$$\begin{aligned}
X' &= aX - \rho \sum_{i=1}^N iC_x^i - pX,\\
(C_x^i)' &= aC_x^i + bvC_x^{i-1} - (i\rho + bv)C_x^i - pC_x^i,\\
\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) C_y^i(t,s) &= bC_y^{i-1}v + (i+1)\rho C_y^{i+1} - (i\rho + bv)C_y^i - (p + \nu(s))C_y^i,\\
C_y^i(t,0) &= \rho(i+1)C_x^{i+1}(t),\\
\text{(A.1)}\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) Y(t,s) &= -(\nu(s) + p)Y(t,s),\\
Y(t,0) &= \rho \sum_{i=1}^N iC_x^i(t),\\
v' &= \int_0^\infty Y(t,s)\eta(s)ds - mv\\
&\quad -bv \sum_{i=0}^{N-1} \left(C_x^i + \int_0^\infty C_y^i(t,s)ds \right),
\end{aligned}$$

where $C_x^{N+1} = C_y^{N+1} = 0$ and where for $i = N$ the loss term $-bvC_z^N$, $z = x, y$, in the equations for C_z^N is dropped since all binding sites are filled so no more attachments are allowed.

Some comments are in order as these equations may not at first appear transparent. Uninfected hosts are lost due to injection of a complex $C_x^i(t)$ by one of its attached phage, which then becomes a newly infected host $C_y^{i-1}(t,0)$ with one less attached phage. The rate of injection is $\rho i C_x^i$ since any one of the i attached phage may inject. This accounts for the loss term $-i\rho C_x^i(t)$ in the equation for X and the boundary condition for $C_y^i(t,s)$ and $Y(t,s)$ at $s = 0$. The injection of an infected host simply reduces the number of attached phage by one. Figure A.1 shows the flow between the compartments C_x^i and C_y^i. Free phage are lost due to attachment to host cells.

Initial conditions for (A.1) consist of specifying nonnegative values for $X(0)$, $C_x^i(0)$, $v(0)$ and nonnegative functions $C_y^i(0,s)$ and $Y(0,s) = U_0(s)$ defined for $s \geq 0$.

$$\begin{tikzcd}
C_x^{i-1} \arrow[r, "bv"'] & C_x^{i} \arrow[r, "bv"'] \arrow[dl, "i\rho"'] & C_x^{i+1} \arrow[dl, "(i+1)\rho"'] \\
C_y^{i-1} \arrow[r, shift right, "bv"'] & C_y^{i} \arrow[l, shift right, "i\rho"'] \arrow[r, shift right, "bv"'] & C_y^{i+1} \arrow[l, shift right, "(i+1)\rho"']
\end{tikzcd}$$

FIG. A.1. *Transfer diagram between complexes for* (A.1)*: ρ denotes injection rate, and bv denotes adsorption rate.*

We intend to employ a quasi–steady state analysis in order to remove the equations for the host/phage complexes. Before doing so it is convenient to integrate the equation for C_y^i with respect to infection-age, yielding

$$(C_y^i)' = -\int_0^\infty (\nu(s) + p)C_y^i(t,s)ds - (i\rho + bv)C_y^i + bvC_y^{i-1} + (i+1)\rho(C_y^{i+1} + C_x^{i+1}).$$

Hereafter, unless specifically mentioned, C_y^i denotes

$$C_y^i(t) = \int_0^\infty C_y^i(t,s)ds.$$

We assume that irreversible binding and injection are fast compared to such processes as growth and washout of host and the latent period. Therefore, we ignore these slower processes in the equations for complexes C_x^i and C_y^i in (A.1). Setting the time derivatives to zero in the C_x^i and C_y^i equations yields

$$\begin{aligned}
&0 = bvC_x^{i-1} - (i\rho + bv)C_x^i,\ 1 \le i \le N,\\
&0 = bvC_y^{i-1} + (i+1)\rho(C_x^{i+1} + C_y^{i+1}) - (i\rho + bv)C_y^i,
\end{aligned}$$

where we employ the conventions used in (A.1).

Adding the first N equations for the C_x^i and then adding all $2N$ equations yields

$$\sum_{i=1}^N iC_x^i = bvC_x^0/\rho, \quad C_x^1 + C_y^1 = bv(C_x^0 + C_y^0)/\rho.$$

Then straightforward calculations give the distribution of phage occupancy of host cell binding sites:

$$C_x^i + C_y^i = \frac{(bv/\rho)^i}{i!}(C_x^0 + C_y^0), \tag{A.2}$$

meaning phage are distributed according to a Poisson distribution with parameter (mean and variance) bv/ρ. We also have

$$C_x^i = C_x^0 \prod_{j=1}^i \frac{u}{j+u},\ 1 \le i \le N-1,\ C_x^N = \frac{u}{N}C_x^{N-1},\ u = bv/\rho. \tag{A.3}$$

We must express all quantities in terms of the variables X, Y, v used in (A.1). We summarize the results of doing so below.

Proposition A.1. *The phage attack rate is*

$$-\rho \sum_{i=1}^{N} i C_x^i = -\frac{bvX}{F_N(cv)}, \quad c = b/\rho,$$

where

$$F_N(u) = 1 + \frac{u}{1+u} + \frac{u^2}{(1+u)(2+u)} + \cdots + \frac{u^N}{(1+u)(2+u)\cdots(N-1+u)N}.$$

The rate of loss of free phage due to attachment satisfies

$$-bv \sum_{i=0}^{N-1} (C_x^i + C_y^i) = -bv(X+Y)\left(1 - \frac{\frac{u^N}{N!}}{\sum_{i=0}^{N} \frac{u^i}{i!}}\right),$$

where $Y = Y(t) = \int_0^\infty Y(t,s)ds$.

Proof. In order to relate C_x^0 to X, we use (A.3):

$$\begin{aligned} X &= \sum_i C_x^i \\ &= C_x^0 \left(1 + \frac{u}{1+u} + \frac{u^2}{(1+u)(2+u)} \right. \\ &\qquad \left. + \cdots + \frac{u^N}{(1+u)(2+u)\cdots(N-1+u)N}\right) \\ &= C_x^0 F_N(u). \end{aligned}$$

Similarly, using (A.2),

$$X + Y = \sum_{i=0}^{N} (C_x^i + C_y^i) = (C_x^0 + C_y^0) \sum_{i=0}^{N} \frac{u^i}{i!}.$$

Finally,

$$\begin{aligned} \sum_{i=0}^{N-1} (C_x^i + C_y^i) &= X + Y - (C_x^N + C_y^N) \\ &= (X+Y) - \frac{u^N}{N!}(C_x^0 + C_y^0) \\ &= (X+Y)\left(1 - \frac{\frac{u^N}{N!}}{\sum_{i=0}^{N} \frac{u^i}{i!}}\right) \\ &\approx X + Y. \qquad \square \end{aligned}$$

In view of the above result and the quasi–steady state analysis, our system (A.1) reduces to

$$\begin{aligned} X' &= aX - \frac{bvX}{F_N(cv)} - pX, \\ \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right) Y(t,s) &= -(\nu(s)+p)Y(t,s), \\ Y(t,0) &= \frac{bvX}{F_N(cv)}, \\ v' &= \int_0^\infty Y(t,s)\eta(s)ds - mv \\ &\quad - bv\left(X(t) + \int_0^\infty Y(t,s)ds\right). \end{aligned} \tag{A.4}$$

Our analysis above shows that the parameter c in (1.1) is given by

$$c = \frac{b}{\rho} = b \times \text{adsorption time.}$$

We may solve for $Y(t,s)$ by integrating along characteristics:

$$Y(t,s) = \left\{ \begin{array}{ll} R(t-s)e^{-ps}e^{-\int_0^s \nu(u)du}, & t > s \\ U_0(s-t)e^{-pt}e^{-\int_{s-t}^s \nu(u)du}, & t < s \end{array} \right\},$$

where $R(t) = \frac{bvX}{F_N(cv)}$. The total infected cell population, $Y(t) = \int_0^\infty Y(t,s)ds$, is easily computed:

$$Y(t) = \int_0^t R(t-s)e^{-ps}e^{-\int_0^s \nu(u)du}ds + \int_0^\infty U_0(s)e^{-pt}e^{-\int_s^{s+t}\nu(u)du}ds.$$

Similarly, for the equation for phage,

$$\begin{aligned} v' = -bv(X+Y) - mv + \int_0^t R(t-s)e^{-ps}\eta(s)e^{-\int_0^s \nu(u)du}ds \\ + e^{-pt}\int_0^\infty U_0(s)\eta(s+t)e^{-\int_s^{t+s}\nu(u)du}ds. \end{aligned}$$

System (A.4) can now be replaced by the equation for X together with the equations above for Y and v. Below we consider two special cases for the lysis rate ν.

A.2. Fixed-length latent period. The special case

$$\nu(s) = \left\{ \begin{array}{ll} 0, & s < \tau \\ \infty, & s > \tau \end{array} \right\}$$

leads to

$$e^{-\int_0^s \nu(u)du} = \chi_{[0,\tau]}(s),\ e^{-\int_s^{s+t}\nu(u)du} = \left\{ \begin{array}{ll} 1, & s+t < \tau \\ 0, & s+t > \tau \end{array} \right\},$$

where, for a set B, $\chi_B(z) = 1$ if $z \in B$; otherwise $\chi_B(z) = 0$.

Hence

$$Y(t) = \int_0^{\min\{t,\tau\}} R(t-s)e^{-ps}ds + \int_0^{\infty} U_0(s)e^{-pt}\chi_{\{s+t<\tau\}}(t,s)ds$$
$$= \int_0^{\min\{t,\tau\}} R(t-s)e^{-ps}ds + \chi_{[0,\tau]}(t)\int_0^{\tau-t} U_0(s)e^{-pt}ds.$$

Equivalently,

$$Y(t) = \left\{ \begin{array}{cc} \int_0^t R(t-s)e^{-ps}ds + \int_0^{\tau-t} U_0(s)e^{-pt}ds, & t<\tau \\ \int_0^\tau R(t-s)e^{-ps}ds, & t>\tau \end{array} \right\}.$$

The equation for phage becomes

$$\begin{array}{c} v' = -bv(X+Y) - mv + \int_0^{\min\{\tau,t\}} R(t-s)e^{-ps}\eta(s)ds \\ + e^{-pt}\chi_{[0,\tau]}(t)\int_0^{\tau-t} U_0(s)\eta(s+t)ds. \end{array}$$

Equivalently,

$$v' = \left\{ \begin{array}{cc} -bv(X+Y) - mv + \int_0^t R(t-s)e^{-ps}\eta(s)ds & \\ + e^{-pt}\int_0^{\tau-t} U_0(s)\eta(s+t)ds, & t<\tau \\ -bv(X+Y) - mv + \int_0^\tau R(t-s)e^{-ps}\eta(s)ds, & t>\tau \end{array} \right\}.$$

If

$$\eta(s) = L\delta(s-\tau),$$

then

$$v' = \left\{ \begin{array}{cc} -bv(X+Y) - mv + Le^{-pt}U_0(\tau-t), & t<\tau \\ -bv(X+Y) - mv + LR(t-\tau)e^{-p\tau}, & t>\tau \end{array} \right\}.$$

These, together with the equation for X from (A.4), are the equations considered in section 2.

A.3. Exponentially distributed latent period. In case of an exponentially distributed latent period, $\nu(s) \equiv \nu$, easy computations yield expressions for the infected:

$$Y(t) = \int_0^t R(s)e^{-(p+\nu)(t-s)}ds + e^{-(p+\nu)t}\int_0^\infty U_0(s)ds$$

or, on differentiation,

$$Y' = -(p+\nu)Y + R(t),\ Y(0) = \int_0^\infty U_0(s)ds.$$

The phage equation becomes

$$v' = -bv(X+Y) - mv + \int_0^t R(s)\eta(t-s)e^{-(p+\nu)(t-s)}ds + e^{-(p+\nu)t}\int_0^\infty U_0(s)\eta(s+t)ds.$$

If $\eta(s) \equiv L\nu$, then v satisfies

$$v' = -bv(X+Y) - mv + \nu L\left(\int_0^t R(s)e^{-(p+\nu)(t-s)}ds + e^{-(p+\nu)t}\int_0^\infty U_0(s)ds\right)$$
$$= -bv(X+Y) - mv + \nu Ly.$$

These, together with the equation for X from (A.4), can be compared to those of Payne and Jansen [14].

A.4. Proof of Lemma 2.1. Straightforward calculation yields that

$$F_{N+1}(u) = F_N(u) - \frac{u^{N+1}}{(1+u)\cdots(N+u)N(N+1)}$$

and that $F_N(u)/u \to 1/N$ as $u \to \infty$.

We show that $\frac{u}{F_N(u)}$ is monotonically increasing by showing it has a positive derivative. Since we may write

$$\frac{u}{F_N(u)} = \frac{u}{F_1(u)} \frac{F_1(u)}{F_2(u)} \cdots \frac{F_{N-1}(u)}{F_N(u)},$$

it suffices to show that $\frac{u}{F_1(u)}$ and $\frac{F_n(u)}{F_{n+1}(u)}$ have positive derivatives for $n \geq 1$. $\frac{u}{F_1(u)} = u/1+u$ is clearly increasing.

$$\begin{aligned} \frac{d}{du}\frac{F_n(u)}{F_{n+1}(u)} &= \frac{d}{du}\frac{F_{n+1}(u)+g(u)}{F_{n+1}(u)} \\ &= \frac{g(u)}{uF_{n+1}(u)}\left(\frac{ug'(u)}{g(u)} - \frac{uF'_{n+1}(u)}{F_{n+1}(u)}\right), \end{aligned}$$

where

$$g(u) = \frac{u^{n+1}}{(1+u)(2+u)\cdots(n+u)(n+1)n}.$$

Straightforward computation gives

$$\frac{ug'(u)}{g(u)} = 1 + \sum_{i=1}^{n} \frac{i}{i+u}$$

and

$$\begin{aligned} uF'_{n+1}(u) = \frac{u}{1+u}\left(\frac{1}{1+u}\right) &+ \frac{u^2}{(1+u)(2+u)}\left(\frac{1}{1+u} + \frac{2}{2+u}\right) + \cdots \\ &+ \frac{u^n}{(1+u)(2+u)\cdots(n+u)}\left(\frac{1}{1+u} + \frac{2}{2+u} + \cdots + \frac{n}{n+u}\right) \\ &+ \frac{u^{n+1}}{(1+u)(2+u)\cdots(n+u)(n+1)}\left(1 + \frac{1}{1+u} + \frac{2}{2+u} + \cdots + \frac{n}{n+u}\right). \end{aligned}$$

On dividing this expression by $F_{n+1}(u)$, we see that it can be viewed as a convex combination of the quantities in parentheses in the previous expression, and therefore it lies between the minimum and maximum of the quantities in parentheses. But the maximum is clearly inside the last parenthesis, which exactly agrees with $\frac{ug'(u)}{g(u)}$. We conclude that

$$\frac{ug'(u)}{g(u)} - \frac{uF'_{n+1}(u)}{F_{n+1}(u)} > 0.$$

A.5. Proof of Proposition 3.1. Setting $u = x + y$ in our system (2.8), we obtain the system

$$\begin{aligned} \frac{dx}{dt} &= ax - \frac{bxv}{F_N(cv)} - px, \\ \frac{du}{dt} &= ax - pu, \quad 0 \le t \le \tau, \\ \frac{dv}{dt} &= -buv - mv. \end{aligned} \tag{A.5}$$

This is a monotone system of differential equations whose forward flow preserves the order relation

$$(\bar{x}, \bar{u}, \bar{v}) \le_K (x, u, v) \Leftrightarrow \bar{x} \le x,\ \bar{u} \le u,\ v \le \bar{v}; \tag{A.6}$$

see [19, Chap. 3, sec. 5]. As the Jacobian matrix on the right-hand side is irreducible, it follows that for $0 < t \le \tau$

$$(\bar{x}(0), \bar{u}(0), \bar{v}(0)) <_K (x(0), u(0), v(0)) \Rightarrow (\bar{x}(t), \bar{u}(t), \bar{v}(t)) \ll_K (x(t), u(t), v(t)),$$

where $<_K$ means at least one strict inequality in (A.6) while $\ll_K$ means all inequalities are strict. The first assertion of Proposition 3.1 is an immediate consequence of the previous inequality.

As for the second assertion of Proposition 3.1, we need some elementary estimates. Inequality $u' \le (a-p)u$ leads to $u(t) \le u(0)e^{(a-p)t}$, so there exists $M \ge 1$ such that $u(t) \le Mu(0)$, $0 \le t \le \tau$. Similarly, $v' \ge -[bMu(0) + m]v$ leads to $v(t) \ge Kv(0)$, $0 \le t \le \tau$, where $K = e^{-(bMu(0)+m)\tau}$. If $0 < \theta < 1$, then

$$\frac{cv}{F_N(cv)} \ge \frac{cKv(0)}{F_N(cKv(0))} \ge \theta N$$

by taking $v(0)$ sufficiently large. Hence, $\frac{x'}{x} \le a - p - (bN/c)\theta$.

REFERENCES

[1] S. Abedon, T. Herschler, and D. Stopar, *Bacteriophage latent-period evolution as a response to resource availability*, Appl. & Environ. Microbiol., 67 (2001), pp. 4233–4241.

[2] E. Beretta and Y. Kuang, *Modeling and analysis of a marine bacteriophage infection with latency*, Nonlinear Anal. Real World Appl., 2 (2001), pp. 35–74.

[3] A. Campbell, *Conditions for the existence of bacteriophage*, Evolution, 15 (1961), pp. 153–165.

[4] P. Grayson, L. Han, T. Winther, and R. Phillips, *Real-time observations of single bacteriophage lambda DNA ejection in vitro*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 14652–14657.

[5] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Encyclopedia of Mathematics and Its Applications 34, Cambridge University Press, Cambridge, UK, 1990.

[6] T. Håusler, *Viruses vs. Superbugs, a Solution to the Antibiotic Crisis?*, MacMillan, New York, 2006.

[7] L. Kasman, A. Kasman, C. Westwater, J. Dolan, M. Schmidt, and J. Norris, *Overcoming the phage proliferation threshold, a mathematical model with implications for phage therapy*, J. Virol., 76 (2002), pp. 5557–5564.

[8] L. Kasman, A. Kasman, C. Westwater, J. Dolan, M. Schmidt, and J. Norris, *Letter to the editor, Author's reply*, J. Virol., 76 (2002), p. 13123.

[9] B. Levin, F. Stewart, and L. Chao, *Resource-limited growth, competition, and predation: A model and experimental studies with bacteria and bacteriophage*, Amer. Naturalist, 111 (1977), pp. 3–24.

[10] R. Lenski and B. Levin, *Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities*, Amer. Naturalist, 125 (1985), pp. 585–602.

[11] B. Levin and J. Bull, *Phage therapy revisited: The population biology of a bacterial infection and its treatment with bacteriophage and antibiotics*, Amer. Naturalist, 147 (1996), pp. 881–898.

[12] B. Levin and J. Bull, *Population and evolutionary dynamics of phage therapy*, Nature Reviews Microbiology, 2 (2004), pp. 166–173.

[13] S. Matsuzaki, M. Rashel, J. Uchiyama, S. Sakurai, T. Ujihara, M. Kuroda, M. Ikeuchi, T. Tani, M. Fujieda, H. Wakiguchi, and S. Imai, *Bacteriophage therapy: A revitalized therapy against bacterial infectious diseases*, J. Infect. Chemother., 11 (2005), pp. 211–219.

[14] R. Payne and V. Jansen, *Understanding bacteriophage therapy as a density-dependent kinetic process*, J. Theoret. Biol., 208 (2001), pp. 37–48.

[15] R. Payne and V. Jansen, *Pharmacokinetic principles of bacteriophage therapy*, Clin. Pharmacokinetics, 42 (2003), pp. 315–325.

[16] R. Payne and V. Jansen, *Evidence for a phage proliferation threshold, Letter to the editor*, J. Virol., 76 (2002), p. 13123.

[17] Z. Qui, *The analysis and regulation for the dynamics of a temperate bacteriophage model*, Math. Biosci., 209 (2007), pp. 417–450.

[18] S. Schrag and J. Mittler, *Host-parasite coexistence: The role of spatial refuges in stabilizing bacteria-phage interactions*, Amer. Naturalist, 148 (1996), pp. 348–377.

[19] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Mathematical Surveys and Monographs 41, AMS, Providence, RI, 1995.

[20] G. Stent, *Molecular Biology of Bacterial Viruses*, W. H. Freeman and Co., London, 1963.

[21] H. Thieme and J. Yang, *On the complex formation approach in modeling predator prey relations, mating, and sexual disease transmission*, in Electron. J. Differ. Equ. Conf. 5, Southwest Texas State Univ., San Marcos, TX, 2000, pp. 255–283.

[22] J. Weitz, H. Hartman, and S. Levin, *Coevolutionary arms race between bacteria and bacteriophage*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9535–9540.

[23] R. Weld, C. Butts, and J. Heinemann, *Models of phage growth and their applicability to phage therapy*, J. Theoret. Biol., 227 (2004), pp. 1–11.

SIAM J. Appl. Math.
Vol. 68, No. 6, pp. 1738–1762

TRANSPORT-BASED IMAGING IN RANDOM MEDIA[*]

GUILLAUME BAL[†] AND KUI REN[†]

Abstract. This paper generalizes well-established derivations of the radiative transfer equation from first principles to model the energy density of time-dependent and monochromatic high frequency waves propagating in a random medium composed of localized scatterers. The correlation length of the random scatterers is small compared to the overall distance of propagation so that ensemble averaging may take place. The correlation length may be either comparable to the typical wavelength in the system (the weak-coupling regime) or larger than the wavelength (the low-density regime). The paper also considers the detection and imaging of inclusions buried in highly scattering random media. In such multiple scattering environments, the coherent wave fields may be too weak to be used for imaging purposes. We thus propose to model the inclusions as parameters in the macroscopic radiative transfer equations and consider the imaging problem as an inverse transport problem. Numerical simulations address the domain of validity of the radiative transfer equation and of the imaging method. Wave propagation is solved by using a Foldy–Lax framework, and the forward and inverse transport problems are solved by using a Monte Carlo method. Since the inverse transport problem is ill-posed, the buried inclusions are parameterized by a small number of degrees of freedom, typically their position and a few geometric properties.

Key words. imaging in random media, high frequency waves in random media, inverse problems, radiative transfer equation, Foldy–Lax model

AMS subject classifications. 35R30, 35R60, 65Z05, 78A40, 78A48

DOI. 10.1137/070690122

1. Introduction.

The imaging of buried inclusions in random media from acoustic, electromagnetic, or elastic wave measurements has a long history. We refer the reader to, e.g., [2, 9, 15, 21, 30, 28, 36] and their references. Several imaging methods have been proposed based on the type of available data and on the regime of wave propagation.

The most favorable situation occurs when the specific realization of the random medium is explicitly known. We can then use the refocusing properties of time reversed waves to backpropagate measured wave fields numerically through the known random environment. The time reversed waves focus on the location of the buried inclusions and allow for fairly accurate imaging; see, e.g., [12, 14, 24]. Note, however, that even small errors in the assessment of the random medium, such as the mislocation of the random scatterers by an amount comparable to the typical wavelength of the propagating waves, have a very large effect on the refocusing of time reversed waves [13, 31].

In most applications, the random medium is not known, and this lack of knowledge inevitably degrades the quality of the reconstructions. Randomness may then roughly be treated in two different ways. One way is to assume that noise is sufficiently small so that it may be treated perturbatively. Imaging is then performed

[*] Received by the editors April 30, 2007; accepted for publication (in revised form) March 18, 2008; published electronically July 3, 2008. This work was supported in part by DARPA-ONR grant N00014-04-1-0224 and NSF grant DMS-0239097.

http://www.siam.org/journals/siap/68-6/69012.html

[†] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (gb2030@columbia.edu, http://www.columbia.edu/~gb2030, kr2002@columbia.edu, http://www.columbia.edu/~kr2002). The work of the first author was partially supported by an Alfred P. Sloan Fellowship.

by backpropagating appropriately mollified wave field measurements into a homogeneous medium or a medium with smooth variations, whose determination is often an important aspect of the reconstruction. One of the main pitfalls in such reconstructions is that many classical inversion techniques are not statistically stable, i.e., the reconstruction strongly depends on the realization in a given set of random media. We refer the reader to, e.g., [15] for optimal, statistically stable imaging methods in that context. Such methods no longer work adequately when the fluctuations in the random medium increase to a point where the coherent wave field, i.e., the part of the wave field that is not affected by random scattering, becomes too weak. In such regimes of strong scattering, other models are necessary.

The alternative to backpropagation in a homogeneous or smoothly varying medium is to find a model that describes wave scattering in the random medium. In regimes where the wavelength and the correlation length of the random medium are very small compared to the overall distance of propagation, such as in the propagation of light through the atmosphere or of near-infrared photons through human tissues, the kinetic description for wave propagation is extremely accurate [1, 18]. The wave energy density is then modeled quite accurately by a radiative transfer (transport) equation or a diffusion equation.

In this paper, we are interested in the validity of the radiative transfer model in a more intermediate regime, where the fluctuations in the random medium are too strong for imaging methods based on coherent information to work, and where the typical wavelength in the system is smaller, though not orders of magnitude smaller, than the overall distance of propagation. Typically, GHz microwaves with a wavelength of 30 cm propagate over distances of tens or hundreds of meters. The typical random medium we consider here is made of hundreds of scatterers, with a mean separation distance between scatterers comparable to or larger than compared to the wavelength but smaller compared to the overall distance of propagation. Our objective is then to detect and image a—sufficiently large—inclusion buried in the random medium.

When the density of scatterers is sufficiently large so that the coherent wave field is too weak to be useful in imaging, the incoherent part of the wave field, i.e., the part that has interacted with the unknown random medium, can no longer be neglected and needs to be modeled. Such a model has to depend on the regime of wave propagation. In the high frequency regime, when the wavelength is smaller than the propagation distance, the simplest extension of wave propagation in homogeneous domains is arguably the radiative transfer equation. Such an equation models the propagation of the phase-space energy density $a(t, \mathbf{x}, \mathbf{k})$ at time t, position $\mathbf{x}$, and wave number $\mathbf{k}$, and for acoustic waves takes the form

$$\frac{\partial a}{\partial t} + c\hat{\mathbf{k}} \cdot \nabla_{\mathbf{x}} a = \mathcal{Q}a, \qquad \mathcal{Q}a = \int_{\mathbb{R}^d} \sigma(\mathbf{x}, \mathbf{k}', \mathbf{k})(a(t, \mathbf{x}, \mathbf{k}') - a(t, \mathbf{x}, \mathbf{k}))\delta(c|\mathbf{k}'| - c|\mathbf{k}|)d\mathbf{k}', \tag{1}$$

where $\hat{\mathbf{k}} = \mathbf{k}/|\mathbf{k}|$, c is sound speed, and $\sigma(\mathbf{x}, \mathbf{k}, \mathbf{k}')$ is the scattering coefficient, which is inversely proportional to the mean free path, the mean distance between successive interactions of the wave energy with the underlying medium. Note that the radiative transfer equation may be seen as a perturbation of the propagation of high frequency waves in a homogeneous medium, which corresponds to the case $\sigma = 0$. The radiative transfer equation is thus characterized by the following two features: scattering is not sufficiently strong to modify the dispersion relation of high frequency waves; this is the left-hand side in (1). However, because of incoherent interactions with the underlying structure, the—partially incoherent—energy density, rather than the wave field, needs to be modeled, and while the energy density is still transported through the random medium, it does so by possibly changing direction in a way described by the scattering operator on the right-hand side of (1). In that sense, it may be seen as the simplest model for the energy density of high-frequency waves propagating in heterogeneous media.

Radiative transfer equations have also been extensively studied and derived either phenomenologically or from first principles (i.e., starting from a wave equation); see [3, 18, 23, 27, 35, 37, 38]. We derive the equation in the setting of localized (possibly) strong scatterers, adapting techniques in [3, 35] to model the energy density of time-dependent and mono-frequency (monochromatic) waves propagating in such a random medium. The localized scatterers are assumed here to have a Poisson distribution with a correlation length that is either comparable to the wavelength (weak-coupling regime) or much larger than the wavelength (low-density regime). We then propose to assess the range of validity of the radiative transfer model by comparing its predictions with wave field calculations. We use a Foldy–Lax model to compute the wave fields. We demonstrate, based on numerical simulations, that the radiative transfer equations are very accurate, provided that the wave energy measurements are sufficiently stable statistically.

Statistical stability is a cornerstone of the interferometric imaging techniques developed in [15]. Reconstructions based on coherent information are much enhanced when an inversion technique that is carefully calibrated to the random medium and statistically stable is employed. When macroscopic models for the incoherent energy density are used, statistical stability is an absolute prerequisite to any form of imaging. It is impossible, based on one measurement, to image an inclusion whose influence on the detectors is a random variable whose fluctuations cannot be averaged out one way or another (for instance, by collecting measurements on a larger detector). It turns out that in the high frequency limit, wave energy densities are indeed statistically stable for a wide class of random media, in the sense that they converge in probability to their deterministic limit as the wavelength tends to 0. This has been proved in simplified models of wave propagation [6, 8] and has been confirmed by numerical simulations [10, 11].

We present here numerical evidence of the statistical stability of the wave energy density in sufficiently mixing random media with localized scatterers. What we mean by sufficiently mixing is that the density of scatterers is sufficiently high. We demonstrate numerically the physically simple fact that the energy density is more stable statistically when the density of scatterers increases (their correlation length decreases) while their scattering strength decreases in such a way that the mean free path remains constant.

Once we are confident in the radiative transfer equation as a model for the wave energy density, we use the model to image buried inclusions in such random media. We assume here that the random medium is statistically homogeneous, i.e., that its statistics are invariant by spatial translation. We can consider two scenarios. In the first scenario, we measure energy densities in the presence of the inclusion. We thus have to estimate the mean free path of the random medium and image the inclusion at the same time. In such a configuration, the inclusion's influence on available measurements has to be larger than the statistical instability coming from our lack of knowledge of the underlying random medium. In the second scenario, we have access to energy measurements in the presence *and* in the absence of the inclusion. We may thus perform differential measurements. These differential measurements are then proportional to the inclusion. With a kinematic picture in mind, all the instability in the random paths that do not visit the object cancels out in differential measurements, thus allowing us to image much smaller objects. We consider reconstructions under these two scenarios based on forward wave field calculations and inverse transport problems.

The rest of the paper is structured as follows. Section 2 presents our model of random media with localized scatterers and derives the radiative transfer equation from a high frequency acoustic wave equation. Time-dependent and mono-frequency equations are considered. The Foldy–Lax approximation to wave propagation is also presented. The numerical validation of the radiative transfer equations based on Foldy–Lax wave simulations and transport Monte Carlo simulations is presented in section 3. Transport-based imaging of inclusions in random media is then considered in section 4.

2. Radiative transfer models.

In this section, we introduce the microscopic and macroscopic models for the propagation of high frequency waves in random media with localized scatterers.

We start with the acoustic wave equation with sound speed given by the superposition of a constant background sound speed and localized strong fluctuations. In section 2.1, the radiative transfer equation is obtained as the high frequency limit of the energy density of the acoustic waves following methods developed in, e.g., [3, 35]. The corresponding transport equation for mono-frequency (time-harmonic) waves is given in section 2.2. Because the scatterers are localized on a scale much smaller than the wavelength, the wave equation for time-harmonic wave fields is approximated by the Foldy–Lax model, which is recalled in section 2.3.

It remains to address the modeling of the buried inclusions. Extended objects are treated like any other pointlike object in the Foldy–Lax formalism. At the radiative transfer level, we assume that the inclusion is sufficiently large compared to the wavelength so that energy reflects specularly at the inclusion's boundary. Such models are explained in greater detail in section 2.4.

2.1. Derivation of the transport equation.

We consider the propagation of scalar waves in media with, for simplicity, constant density ρ_0 and spatially varied compressibility $\kappa(\mathbf{x})$,

$$\rho_0 \partial_t \mathbf{v} + \nabla p = 0, \quad \kappa(\mathbf{x})\partial_t p + \nabla \cdot \mathbf{v} = 0, \quad p(0,\mathbf{x}) = p_0(\mathbf{x}), \quad \mathbf{v}(0,\mathbf{x}) = \mathbf{v}_0(\mathbf{x}), \tag{2}$$

where $t > 0$ and $\mathbf{x} \in \mathbb{R}^d$ with $d \geq 2$ the spatial dimension. The theories that follow generalize to the context of electromagnetic and elastic waves; see, e.g., [3, 35]. For concreteness, we restrict ourselves to the case of acoustic waves.

In the high frequency regime of interest in this paper, the above equation rescales, after the change of variables $t \mapsto \frac{t}{\varepsilon}$ and $\mathbf{x} \mapsto \frac{\mathbf{x}}{\varepsilon}$, as

$$\rho_0 \varepsilon \partial_t \mathbf{v}_\varepsilon + \varepsilon \nabla p_\varepsilon = 0, \quad \kappa_\varepsilon(\mathbf{x}) \varepsilon \partial_t p_\varepsilon + \varepsilon \nabla \cdot \mathbf{v}_\varepsilon = 0, \tag{3}$$

where the initial conditions $p_\varepsilon(0,\mathbf{x}) = p_{0\varepsilon}(\mathbf{x})$ and $\mathbf{v}_\varepsilon(0,\mathbf{x}) = \mathbf{v}_{0\varepsilon}(\mathbf{x})$ oscillate at the frequency ε^{-1}. The parameter ε thus models the typical wavelength in the system. The wave speed is defined as

$$c_\varepsilon^2(\mathbf{x}) = \frac{1}{\rho_0 \kappa_\varepsilon(\mathbf{x})} = c_0^2(\mathbf{x}) - \sqrt{\varepsilon} V_\varepsilon \left(\frac{\mathbf{x}}{\varepsilon} \right). \tag{4}$$

We retain the scaling $\sqrt{\varepsilon}V_\varepsilon(\frac{\mathbf{x}}{\varepsilon})$ to use the results developed in, e.g., [3, 35]. Neither the amplitude nor the correlation length of the potential V_ε is necessarily of order $O(1)$, however. More precisely, the potential V_ε is chosen as

$$V_\varepsilon(\mathbf{x}) = \varepsilon^{-\frac{(\gamma+2\beta)d}{2}} \sum_j \tau_j V\Big(\frac{\mathbf{x}-\mathbf{x}_j^\varepsilon}{\varepsilon^\beta}\Big), \tag{5}$$

where $\beta > 0$, $\gamma < 1$, where $V(\mathbf{x})$ is a compactly supported nonnegative, uniformly bounded function, where the points $\mathbf{x}_j^\varepsilon(\omega)$ form a Poisson point process of density $\nu_\varepsilon = \varepsilon^{\gamma d} n_0$, and where the coefficients $\tau_j(\omega)$ are square-integrable, mean-zero, independent identically distributed random variables. Here ω is a point in a sufficiently large abstract probability space $(\Omega, \mathcal{F}, P)$.

The sound speed fluctuations are therefore of the form

$$\sqrt{\varepsilon}V_\varepsilon\left(\frac{\mathbf{x}}{\varepsilon}\right) = \varepsilon^{\frac{1-(\gamma+2\beta)d}{2}} \sum_j \tau_j V\Big(\frac{\mathbf{x}-\varepsilon\mathbf{x}_j^\varepsilon}{\varepsilon^{1+\beta}}\Big). \tag{6}$$

We thus conclude that the thickness of the scatterers is $t_\varepsilon = \varepsilon^{1+\beta} \ll \varepsilon$, the correlation length in the medium is $l_\varepsilon = \varepsilon^{1-\gamma}L$ for L a typical distance of propagation, which verifies $l_\varepsilon \gg \varepsilon$ when $0 < \gamma < 1$ and $L = O(1)$, and the density of scatterers is $n_\varepsilon = \varepsilon^{-d}\nu_\varepsilon = \varepsilon^{(\gamma-1)d} n_0 \gg O(1)$.

Note that the Poisson point process allows for the clustering of points $\mathbf{x}_j^\varepsilon$, although the number of points in a given bounded domain is bounded P-a.s. There is therefore a (P-)small subset Ω_ε of Ω, where the above fluctuation is larger than c_0^2, which would result in a negative c_ε^2. The process in (6) thus needs to be modified on Ω_ε, for instance, by setting the fluctuations to 0. We verify, although we shall not present this here, that such modifications of the process in (6) occur on a very small set and that the calculations of the power spectra presented below are not affected by the change.

We can now use the methodology developed in [35] to show that the wave energy density is such that

$$\mathcal{E}_\varepsilon(t,\mathbf{x}) - \int_{\mathbb{R}^d} a_\varepsilon(t,\mathbf{x},\mathbf{k})d\mathbf{k} \to 0 \tag{7}$$

in a weak sense (i.e., after integration against a test function in the spatial variables $\mathbf{x}$), where $\mathcal{E}_\varepsilon(t,\mathbf{x})$ is the energy density defined as

$$\mathcal{E}_\varepsilon(t,\mathbf{x}) = \rho_0|\mathbf{v}_\varepsilon|^2(t,\mathbf{x}) + \kappa_\varepsilon(\mathbf{x})p_\varepsilon^2(t,\mathbf{x}), \tag{8}$$

and where $a_\varepsilon(t,\mathbf{x},\mathbf{k})$ is a phase-space energy density, which solves the following radiative transfer equation:

$$\frac{\partial a_\varepsilon}{\partial t} + c_0\hat{\mathbf{k}}\cdot\nabla a_\varepsilon = \int_{\mathbb{R}^d} \sigma_\varepsilon(\mathbf{x},\mathbf{k},\mathbf{q})(a_\varepsilon(\mathbf{x},\mathbf{q}) - a_\varepsilon(\mathbf{x},\mathbf{k}))\delta(c_0|\mathbf{k}| - c_0|\mathbf{q}|)d\mathbf{q}, \tag{9}$$

with appropriate initial conditions, where

$$\sigma_\varepsilon(\mathbf{x},\mathbf{k},\mathbf{q}) = \frac{\pi c_0^2|\mathbf{k}|^2}{2(2\pi)^d}\hat{R}_\varepsilon(\mathbf{k}-\mathbf{q}). \tag{10}$$

Here, $\hat{R}_\varepsilon$ is the power spectrum of the fluctuations V_ε. It is the Fourier transform of the correlation function R_ε of the fluctuations V_ε. They are defined as follows:

$$\begin{aligned} c_0^4 R_\varepsilon(\mathbf{y}) &= \mathbb{E}\{V_\varepsilon(\mathbf{x})V_\varepsilon(\mathbf{x}+\mathbf{y})\}, \\ (2\pi)^d c_0^4 \hat{R}_\varepsilon(\mathbf{p})\delta(\mathbf{p}+\mathbf{q}) &= \mathbb{E}\{\hat{V}_\varepsilon(\mathbf{p})\hat{V}_\varepsilon(\mathbf{q})\}, \end{aligned} \tag{11}$$

where $\hat{V}_\varepsilon(\mathbf{p}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}\cdot\mathbf{p}} V_\varepsilon(\mathbf{x})d\mathbf{x}$ is the Fourier transform of $V_\varepsilon(\mathbf{x})$. We then have the following result on the asymptotic limit of the power spectrum.

LEMMA 2.1. *Let us assume that $V(\mathbf{x})$ is a nonnegative, integrable, compactly supported function such that $\hat{V}(\mathbf{0}) = c_0^2 L^d$ for some characteristic distance $0 < L = O(1)$, where $\hat{V}(\mathbf{k})$ is the Fourier transform of V. Then we find that the power spectrum $\hat{R}_\varepsilon(\mathbf{k})$ converges in the uniform norm uniformly on compact sets to the limit*

$$\hat{R}_0 = L^{2d}\mathbb{E}\{\tau^2\}n_0. \tag{12}$$

Proof. We calculate that

$$c_0^4 R_\varepsilon(\mathbf{y}) = \varepsilon^{-(\gamma+2\beta)d}\mathbb{E}\{\tau^2\}\mathbb{E}\left\{\sum_{j=1}^\infty V\Big(\frac{\mathbf{x}-\mathbf{x}_\varepsilon^j}{\varepsilon^\beta}\Big)V\Big(\frac{\mathbf{x}+\mathbf{y}-\mathbf{x}_\varepsilon^j}{\varepsilon^\beta}\Big)\right\}.$$

Since V is compactly supported, there is a domain D, at $\mathbf{x}$ and $\mathbf{y}$ fixed, such that the above product vanishes for $\mathbf{x}_j^\varepsilon$ outside of D. The Poisson point process verifies that the number of points on D satisfies a Poisson distribution and that, conditioned on the number of points, these points are uniformly and independently distributed on D. This yields that

$$\begin{aligned} &\mathbb{E}\left\{\sum_{j=1}^\infty V\Big(\frac{\mathbf{x}-\mathbf{x}_\varepsilon^j}{\varepsilon^\beta}\Big)V\Big(\frac{\mathbf{x}+\mathbf{y}-\mathbf{x}_\varepsilon^j}{\varepsilon^\beta}\Big)\right\} \\ &\quad = \sum_{m=0}^\infty e^{-|D|\nu_\varepsilon}\frac{(|D|\nu_\varepsilon)^m}{m!}\sum_{j=1}^m \int_D V\Big(\frac{\mathbf{x}-\mathbf{z}}{\varepsilon^\beta}\Big)V\Big(\frac{\mathbf{x}+\mathbf{y}-\mathbf{z}}{\varepsilon^\beta}\Big)\frac{d\mathbf{z}}{|D|} \\ &\quad = \nu_\varepsilon \int_{\mathbb{R}^d} V\Big(\frac{\mathbf{z}-\mathbf{x}}{\varepsilon^\beta}\Big)V\Big(\frac{\mathbf{z}-\mathbf{x}-\mathbf{y}}{\varepsilon^\beta}\Big)d\mathbf{z}, \end{aligned}$$

where $|D|$ is the Lebesgue measure of D. Now,

$$\begin{aligned} H_\varepsilon(\mathbf{y}) &= \int_D V\Big(\frac{\mathbf{z}-\mathbf{x}}{\varepsilon^\beta}\Big)V\Big(\frac{\mathbf{z}-\mathbf{x}-\mathbf{y}}{\varepsilon^\beta}\Big)d\mathbf{z} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\mathbf{x}\cdot\mathbf{k}}e^{i(\mathbf{x}+\mathbf{y})\cdot\mathbf{k}}\varepsilon^{2\beta d}|\hat{V}(\varepsilon^\beta\mathbf{k})|^2 d\mathbf{k} \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i\mathbf{y}\cdot\mathbf{k}}\varepsilon^{2\beta d}|\hat{V}(\varepsilon^\beta\mathbf{k})|^2 d\mathbf{k}. \end{aligned}$$

Its Fourier transform is thus given by

$$\hat{H}_\varepsilon(\mathbf{p}) = \varepsilon^{2\beta d}|\hat{V}(\varepsilon^\beta\mathbf{p})|^2 = \varepsilon^{2\beta d}(|\hat{V}(\mathbf{0})|^2 + O(\varepsilon^\beta)),$$

where $O(\varepsilon^\beta)$ means a term of order ε^β in the uniform norm, uniformly bounded on compact sets. This follows, e.g., from the analyticity of $\hat{V}(\mathbf{p})$. Since $\hat{R}_\varepsilon(\mathbf{k})$ is the Fourier transform of $R_\varepsilon(\mathbf{x})$, we find that

$$c_0^4\hat{R}_\varepsilon(\mathbf{k}) = (|\hat{V}(\mathbf{0})|^2 + O(\varepsilon^\beta))\mathbb{E}\{\tau^2\}n_0.$$

This proves the result. □

In the limit $\varepsilon \to 0$, we thus find that a_ε converges to the solution of the following radiative transfer equation:

$$\frac{\partial a}{\partial t} + c_0 \hat{\mathbf{k}} \cdot \nabla a + \Sigma(\mathbf{x}, |\mathbf{k}|) a = \int_{\mathbb{R}^d} \sigma(\mathbf{x}, \mathbf{k}, \mathbf{q}) a(t, \mathbf{x}, \mathbf{q}) \delta(c_0|\mathbf{k}| - c_0|\mathbf{q}|) d\mathbf{q}, \tag{13}$$

where

$$\begin{aligned} \sigma(\mathbf{x}, \mathbf{k}, \mathbf{q}) &= \frac{\pi |\mathbf{k}|^2 L^{2d} c_0^2 \mathbb{E}\{\tau^2\} n_0}{2(2\pi)^d}, \\ \Sigma(\mathbf{x}, |\mathbf{k}|) &= \frac{|S^{d-1}|\pi}{2(2\pi)^d} |\mathbf{k}|^{d+1} L^{2d} c_0 \mathbb{E}\{\tau^2\} n_0. \end{aligned} \tag{14}$$

Here, $|S^{d-1}|$ is the Lebesgue measure of the unit sphere in $\mathbb{R}^d$.

Note that the transport equations generalize to the case where the density of scatterers depends on space. For instance, we may assume that the sound speed fluctuations are of the form

$$\sqrt{\varepsilon} V_\varepsilon \left(\mathbf{x}, \frac{\mathbf{x}}{\varepsilon}\right) = \varepsilon^{\frac{1-(\gamma+2\beta)d}{2}} \sum_j \tau_j \varphi(\mathbf{x}_j^\varepsilon) V\left(\frac{\mathbf{x} - \varepsilon \mathbf{x}_j^\varepsilon}{\varepsilon^{1+\beta}}\right), \tag{15}$$

where $\varphi(\mathbf{x})$ is a deterministic nonnegative function on $\mathbb{R}^d$. In the numerical simulations considered below, $\varphi(\mathbf{x})$ is the indicatrix function of our computational domain. The limit in (12) is then modified as

$$\hat{R}_0(\mathbf{x}) = \varphi^2(\mathbf{x}) L^{2d} \mathbb{E}\{\tau^2\} n_0. \tag{16}$$

The scattering coefficients in (14) also need to be multiplied by $\varphi^2(\mathbf{x})$ and thus become spatially dependent.

2.2. Transport equation in the frequency domain.

The derivation of transport equations for mono-frequency waves cannot be deduced in a straightforward way from that of time-dependent equations. We refer to [17] for a derivation of kinetic models from the Helmholtz equation. We formally generalize these results by adding a scattering operator to the transport equation to model the interaction with the underlying random medium.

The Helmholtz equation for the mono-frequency wave field $u_\varepsilon(\mathbf{x})$ with high frequency $\frac{\omega}{\varepsilon}$ takes the form

$$\varepsilon^2 \Delta u_\varepsilon(\mathbf{x}) + \frac{\omega^2}{c_\varepsilon^2(\mathbf{x})} u_\varepsilon(\mathbf{x}) = \frac{1}{\varepsilon^{\frac{d-1}{2}}} \varphi\Big(\frac{\mathbf{x} - \mathbf{x}_0}{\varepsilon}\Big), \tag{17}$$

where $\varphi(\mathbf{x})$ is a smooth function which localizes in the vicinity of the point $\mathbf{x}_0$. We assume that the density of scatterers vanishes in the vicinity of the source location $\mathbf{x}_0$ so that the dispersion relation is $\omega = c_0|\mathbf{k}|$ locally.

In the absence of scatterers, the results obtained, e.g., in [17] show that there exists a positive bounded measure $a(\mathbf{x}, \mathbf{k})$ such that

$$\lim_{\varepsilon \to 0} |u_\varepsilon(\mathbf{x})|^2 := \nu(\mathbf{x}) = \int_{\mathbb{R}^d} a(\mathbf{x}, \mathbf{k}) d\mathbf{k}, \tag{18}$$

where $\nu(\mathbf{x})$ is a positive measure on $\mathbb{R}^d$. Moreover, the phase-space measure $a(\mathbf{x},\mathbf{k})$ solves the Liouville equation

$$c_0\hat{\mathbf{k}}\cdot\nabla a = Q(\mathbf{x},\mathbf{k}), \tag{19}$$

with the source $Q(\mathbf{x},\mathbf{k})$ given by

$$\begin{aligned} Q(\mathbf{x},\mathbf{k}) &= \frac{c_0^2}{2\omega(2\pi)^{d-1}}\delta(\mathbf{x}-\mathbf{x}_0)\delta\left(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\right)|\hat{\varphi}(\mathbf{k})|^2 \\ &= \frac{c_0^3}{4\omega^2(2\pi)^{d-1}}\delta(\mathbf{x}-\mathbf{x}_0)\delta\left(\frac{\omega}{c_0}-|\mathbf{k}|\right)|\hat{\varphi}(\mathbf{k})|^2. \end{aligned} \tag{20}$$

The above transport equation should be augmented with outgoing radiation conditions, namely, the incoming field $a(\mathbf{x}-t\mathbf{k},\mathbf{k})\to 0$ as $t\to+\infty$.

We briefly recall the derivation of the above equation and explain the scaling for the source term in (17). The regularized Helmholtz equation (17) with constant sound speed may be written as

$$\left(i\varepsilon\alpha+\varepsilon^2\Delta+\frac{\omega^2}{c_0^2}\right)u_\varepsilon = \frac{1}{\varepsilon^{\frac{d-1}{2}}}\varphi\Big(\frac{\mathbf{x}}{\varepsilon}\Big)$$

for some causality-preserving regularization parameter $0<\alpha\ll 1$ that will be sent to 0 eventually, and where we set $\mathbf{x}_0=0$ to simplify the presentation. In the Fourier domain at wave number $\boldsymbol{\xi}/\varepsilon$, this is

$$\hat{u}_\varepsilon\Big(\frac{\boldsymbol{\xi}}{\varepsilon}\Big) = \frac{\varepsilon^d\varepsilon^{-\frac{d-1}{2}}\hat{\varphi}(\boldsymbol{\xi})}{i\varepsilon\alpha+\frac{\omega^2}{c^2}-|\boldsymbol{\xi}|^2}.$$

Let us now introduce the following Wigner transform of u_ε:

$$W_\varepsilon(\mathbf{x},\mathbf{k}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} u_\varepsilon\left(\mathbf{x}-\frac{\varepsilon\mathbf{y}}{2}\right)u_\varepsilon^*\left(\mathbf{x}+\frac{\varepsilon\mathbf{y}}{2}\right)e^{i\mathbf{k}\cdot\mathbf{y}}d\mathbf{y},$$

which verifies that

$$|u_\varepsilon(\mathbf{x})|^2 = \int_{\mathbb{R}^d} W_\varepsilon(\mathbf{x},\mathbf{k})d\mathbf{k}.$$

We verify that the Fourier transform $\mathbf{x}\to\mathbf{q}$ of W_ε is given by

$$\begin{aligned} \hat{W}_\varepsilon(\mathbf{q},\mathbf{k}) &= \frac{1}{(2\pi\varepsilon)^d}\hat{u}_\varepsilon\Big(\frac{\mathbf{q}}{2}-\frac{\mathbf{k}}{\varepsilon}\Big)\widehat{u_\varepsilon^*}\Big(\frac{\mathbf{q}}{2}+\frac{\mathbf{k}}{\varepsilon}\Big) \\ &= \frac{\varepsilon}{(2\pi)^d}\frac{|\hat{\varphi}(\mathbf{k})|^2}{[i\varepsilon\alpha+\frac{\omega^2}{c_0^2}-|\mathbf{k}+\frac{\varepsilon\mathbf{q}}{2}|^2][-i\varepsilon\alpha+\frac{\omega^2}{c_0^2}-|\mathbf{k}-\frac{\varepsilon\mathbf{q}}{2}|^2]}+l.o.t. \\ &= \frac{\varepsilon}{(2\pi)^d}\frac{|\hat{\varphi}(\mathbf{k})|^2}{\big(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\big)^2-\big(i\varepsilon\alpha-\varepsilon\mathbf{k}\cdot\mathbf{q}\big)^2}+l.o.t. \\ &= \frac{\varepsilon}{(2\pi)^d}\frac{|\hat{\varphi}(\mathbf{k})|^2}{2i\varepsilon\alpha-2\varepsilon\mathbf{k}\cdot\mathbf{q}}\left[\frac{1}{\big(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\big)-i\varepsilon\alpha}-\frac{1}{\big(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\big)+i\varepsilon\alpha}\right]+l.o.t. \\ &= \frac{1}{i(2\pi)^d}\frac{|\hat{\varphi}(\mathbf{k})|^2}{2\alpha+i2\mathbf{k}\cdot\mathbf{q}}2\pi i\delta\left(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\right)+l.o.t. \\ &= \frac{1}{2\alpha+i2\mathbf{k}\cdot\mathbf{q}}\frac{|\hat{\varphi}(\mathbf{k})|^2}{(2\pi)^{d-1}}\delta\left(\frac{\omega^2}{c_0^2}-|\mathbf{k}|^2\right)+l.o.t. \end{aligned}$$

Here, *l.o.t.* refers to terms that tend to 0 in the sense of distributions as $\varepsilon \to 0$. We refer the reader to, e.g., [17] for a more rigorous derivation. This shows that in the limit $\varepsilon \to 0$, we have

$$(2\alpha + i2\mathbf{k}\cdot\mathbf{q})\hat{W}(\mathbf{q},\mathbf{k}) = \frac{|\hat{\varphi}(\mathbf{k})|^2}{(2\pi)^{d-1}}\delta\left(\frac{\omega^2}{c_0^2} - |\mathbf{k}|^2\right).$$

Sending the regularizing parameters $\alpha \to 0^+$ and denoting by $a(\mathbf{x},\mathbf{k})$ the inverse Fourier transform $\mathbf{q} \to \mathbf{x}$ of $\hat{W}$, we obtain the Liouville equation (19).

In the presence of scatterers whose density vanishes at $\mathbf{x} = \mathbf{x}_0$, the radiating source term $Q(\mathbf{x},\mathbf{k})$ is not modified, and propagation in a homogeneous medium is replaced formally by propagation in a scattering medium, as in the preceding section. The phase-space energy density $a(\mathbf{x},\mathbf{k})$ thus solves the following stationary transport equation:

$$(21)\qquad c_0\hat{\mathbf{k}}\cdot\nabla a + \Sigma(\mathbf{x},|\mathbf{k}|)a = \int_{\mathbb{R}^d}\sigma(\mathbf{x},\mathbf{k},\mathbf{q})a(\mathbf{x},\mathbf{q})\delta(c_0|\mathbf{k}| - c_0|\mathbf{q}|)d\mathbf{q} + Q(\mathbf{x},\mathbf{k}).$$

The equation should be augmented with zero-incoming radiation conditions, i.e., $a(\mathbf{x}-t\mathbf{k},\mathbf{k}) \to 0$ as $t \to +\infty$. In practice, we choose the source term $\varphi(\mathbf{x}) = \delta(\mathbf{x})$ so that $\hat{\varphi}(\mathbf{k}) = 1$.

2.3. Foldy–Lax model for point scatterers. The Helmholtz equation (17) is very demanding to solve numerically for the choice of sound speed fluctuations given in (5). Since $\beta > 0$ so that $\varepsilon^\beta \to 0$ as $\varepsilon \to 0$, the localized scatterers are very small compared to the wavelength of the propagating waves. As a consequence, we can replace the localized scatterers by point scatterers. We thus have to solve a Helmholtz equation with randomly distributed point scatterers.

Assuming that the number of scatterers on a given computational domain is N, we obtain that the solution to the Helmholtz equation is given by

$$(22)\qquad u(\mathbf{x}) = u^i(\mathbf{x}) + \sum_{j=1}^{N}\tau_j G_0(\mathbf{x},\mathbf{x}_j)u(\mathbf{x}_j),$$

where $u^i(\mathbf{x})$ is the wave field generated by the source and the Green's function $G_0(\mathbf{x},\mathbf{x}')$ is given by

$$(23)\qquad G_0(\mathbf{x},\mathbf{x}') = \begin{cases} \dfrac{i}{4}H_0^1(k|\mathbf{x}-\mathbf{x}'|), & d = 2,\\ \dfrac{e^{ik|\mathbf{x}-\mathbf{x}'|}}{4\pi|\mathbf{x}-\mathbf{x}'|}, & d = 3,\end{cases}$$

with H_0^1 the 0th order Hankel function of the first kind. When $\mathbf{x} = \mathbf{x}_j$, the above solution needs modification. The Foldy–Lax model [25, 29, 40] removes the singularities in a self-consistent fashion by imposing that

$$(24)\qquad u(\mathbf{x}_j) = u^i(\mathbf{x}_j) + \sum_{\substack{j'=1\\ j'\neq j}}^{N}\tau_{j'}G_0(\mathbf{x}_j;\mathbf{x}_{j'})u(\mathbf{x}_{j'})$$

for $j = 1, \ldots, N$. We thus need to solve the system (24) first, and then we can evaluate the field $u(\mathbf{x})$ at each point $\mathbf{x}$ using (22).

The Foldy–Lax equation (24) can be written in matrix form as

$$\mathbf{H}\mathbf{U} = \mathbf{U}^i, \tag{25}$$

where

$$\mathbf{U} \equiv [u(\mathbf{x}_1), \ldots, u(\mathbf{x}_N)]^{\mathrm{T}}, \qquad \mathbf{U}^i \equiv [u^i(\mathbf{x}_1), \ldots, u^i(\mathbf{x}_N)]^{\mathrm{T}}, \tag{26}$$

and where the complex matrix $\mathbf{H}$ is given by

$$H_{jj'} = \delta_{jj'} - (1 - \delta_{jj'})\tau_{j'} G_0(\mathbf{x}_j, \mathbf{x}_{j'}). \tag{27}$$

For simplicity, let us assume that the strength of the scatterers is given by $\tau_j = \epsilon_j \tau k^2$, where the ϵ_j are independent variables taking the values 1 and -1 with equal probability. This is the setting that we will consider in the next section. We then verify that

$$\Sigma_{2D}(\mathbf{x}) = \frac{k^3\tau^2 n_0}{4}, \qquad \sigma_{2D}(\mathbf{x}, \mathbf{k}, \mathbf{k}') = \frac{k^3\tau^2 n_0}{8\pi} \tag{28}$$

in the two-dimensional case and

$$\Sigma_{3D}(\mathbf{x}) = \frac{k^4\tau^2 n_0}{4\pi}, \qquad \sigma_{3D}(\mathbf{x}, \mathbf{k}, \mathbf{k}') = \frac{k^4\tau^2 n_0}{16\pi^2} \tag{29}$$

in the three-dimensional case. These expressions are consistent with (14), provided that we set $L = 1$ and normalize the sound speed $c_0 = 1$. We shall assume that $L = 1$ and $c_0 = 1$ for the rest of the paper.

2.4. Models for the buried inclusions. We now have to model the buried inclusions, both at the level of the Helmholtz equation and of the radiative transfer equation.

The Foldy–Lax model can be generalized to account for the presence of extended objects. Here, we consider impenetrable objects with vanishing Neumann boundary conditions at the inclusion's boundary for the Helmholtz equation. In this setting, (22) becomes

$$u(\mathbf{x}) = u^i(\mathbf{x}) + \sum_{j=1}^{N} \tau_j G_0(\mathbf{x}; \mathbf{x}_j) u(\mathbf{x}_j) + \sum_{l=1}^{M} \int_{\partial\Omega_l} \mathbf{n}_l \cdot \nabla_{\mathbf{y}} G_0(\mathbf{x}; \mathbf{y}) u(\mathbf{y}) dS(\mathbf{y}), \tag{30}$$

where M is the number of extended objects in the domain and Ω_l is the lth inclusion with sufficiently smooth boundary $\partial\Omega_l$ and outer normal vector $-\mathbf{n}_l$ on the boundary. The Foldy–Lax consistent equation now becomes

$$u(\mathbf{x}_j) = u^i(\mathbf{x}_j) + \sum_{j' \neq j}^{N} \tau_{j'} G_0(\mathbf{x}_j; \mathbf{x}_{j'}) u(\mathbf{x}_{j'}) + \sum_{l=1}^{M} \int_{\partial\Omega_l} \mathbf{n}_l \cdot \nabla_{\mathbf{y}} G_0(\mathbf{x}_j; \mathbf{y}) u(\mathbf{y}) dS(\mathbf{y}) \tag{31}$$

for $j = 1, \ldots, N$.

In order to evaluate wave fields at arbitrary points $\mathbf{x}$, we need to solve (30) for points on the boundary of the extended objects and (31) for $\mathbf{x}_j, j = 1, \ldots, N$. Equations (30) and (31) are the new self-consistent Foldy–Lax multiple scattering equations in the case where extended objects are present.

At the transport level, in order to obtain a contribution of order $O(1)$, we need to assume that the inclusion is comparable in size to the overall distance of propagation. This implies that the extended object is large compared to the wavelength ε so that its boundary may be treated as specularly reflecting. In other words, we assume the following specular reflection:

$$a(\mathbf{x},\mathbf{k}) = a(\mathbf{x},\mathbf{k}-2\mathbf{k}\cdot\mathbf{n}(\mathbf{x})\mathbf{n}(\mathbf{x})), \qquad \mathbf{x}\in\partial\Omega_l. \tag{32}$$

The radiative transfer equation (21) holds outside of the inclusions, i.e., on $\mathbb{R}^d\backslash(\cup_l\Omega_l)$.

3. Numerical validations.

Although radiative transfer equations have been used for a long time to describe the energy density of waves in random media, numerical validations of such models are more recent; see, e.g., [10, 33] for simulations in the time domain. The reason is that the propagation of high frequency waves in highly heterogeneous media is computationally quite expensive. In this section, we compare the energy densities of monochromatic waves based on the Foldy–Lax model with the solution of the corresponding radiative transfer equation.

3.1. The wave and transport solvers.

The Foldy–Lax consistent equations (25) form a system of complex-valued algebraic equations with a dense matrix $\mathbf{H}$. We solve the system by a direct solver that utilizes the LU factorization. In the case where extended objects are present in the domain, we need to solve (30) and (31). We approximate the boundary integrals by standard numerical quadrature rules. Since these integrals are weakly singular, we adopt the kernel splitting method developed in [22] to discretize the integrals.

The random medium is generated by distributing the random scatterers according to a Poisson point process of density n_0. We recall that for a bounded volume element V, the number of points in V has the distribution $\mathbb{P}(N_V = k) = \frac{e^{-n_0|V|}(n_0|V|)^k}{k!}$ for $k \geq 0$, where $|V|$ is the (Lebesgue) measure of V. Once a realization of the number k is chosen, the k points are placed in V using a uniform (normalized Lebesgue) distribution on V. A typical distribution of point scatterers is shown in Figure 1.

The transport equation (21) is solved by the Monte Carlo method [39]. We run enough particles to ensure that the statistical error in the simulation is smaller than any other involved quantity. As in [10], to which we refer the reader for more details, we use as much as possible the same random trajectories in the Monte Carlo simulation for the calculations performed with and without the extended objects. This variance reduction technique is necessary to calculate the influence of the inclusions accurately.

Note that the Foldy–Lax model and the radiative transfer equations work for a given (large) frequency. By varying the frequency in the vicinity of a central frequency, time-domain data may be obtained by inverse Fourier transform. In the rest of the paper, we concentrate on frequency domain calculations.

3.2. Numerical results.

To compare the wave and transport models, we compute the total energy on a fixed array of detectors D; see Figure 1. After appropriate normalization, the energies take the form

$$E_W = \int_D |u(\mathbf{x})|^2 d\mathbf{x} \quad \text{and} \quad E_T = \int_D \int_{S^{d-1}} a(\mathbf{x},\hat{\mathbf{k}})d\hat{\mathbf{k}}d\mathbf{x}, \tag{33}$$

where $u(\mathbf{x})$ is the random solution to the wave equation and $a(\mathbf{x},\hat{\mathbf{k}})$ is the energy density of those waves at position $\mathbf{x}$ propagating in direction $\hat{\mathbf{k}}\in S^{d-1}$, where $d=2$ in all our numerical simulations.

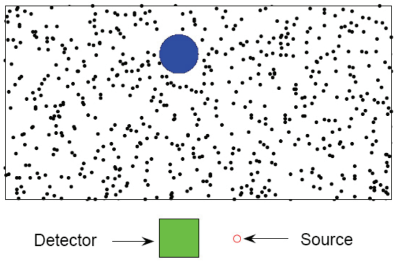


Fig. 1. *Setup for the numerical simulations. The inclusion may be placed inside or outside of the random medium. We show here a typical realization of the distribution of* 1000 *point scatterers.*

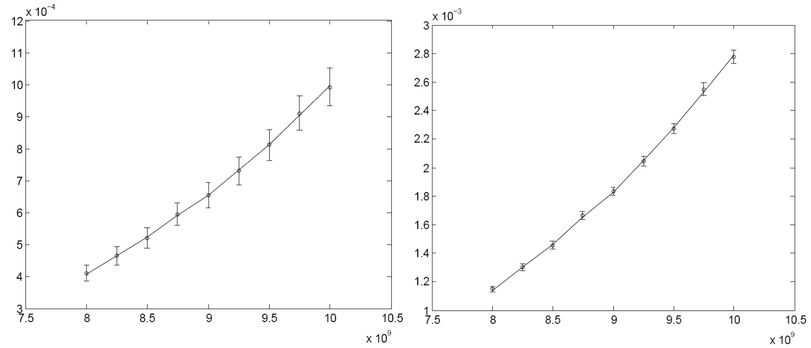


Fig. 2. *Comparison between measured transport and wave data at frequencies between* $\omega = \frac{2\pi}{\lambda}$ *and* 1.25ω *with detectors of size* $40\lambda \times 40\lambda$ *(left) and* $80\lambda \times 80\lambda$ *(right), respectively. Solid line: transport data* E_T*. Circles with error bar: wave data* $\mathbb{E}\{E_W\}$ *and its standard deviation* $\sigma(E_W)$*.*

Since $u(\mathbf{x})$ is a random variable, E_W is also a random variable, which thus depends on the realization of the random medium. Note that E_T, in contrast, is a deterministic quantity. We have therefore two objectives: (i) show that $\mathbb{E}\{E_W\}$ is close to E_T; and (ii) show that the standard deviation $\sigma(E_W)$ of E_W is small. The latter is defined as

$$\sigma(E_W) = \Big(\mathbb{E}\{(E_W - \mathbb{E}\{E_W\})^2\}\Big)^{\frac{1}{2}}. \tag{34}$$

From now on, we use the standard notation σ to denote standard deviations, which should not be confused with the scattering cross-section in (21). We are interested in the behavior of $\sigma(E_W)$ as a function of the mean free path $c_0\Sigma^{-1}$.

Energy measurements and detector size. The first numerical test compares wave and transport data in the absence of any buried inclusion. The setup is as shown in Figure 1. In order to illustrate the behavior of the energy density as a function of frequency, we compare the models for nine frequencies uniformly distributed on $[\omega \;\; 1.25\omega]$, where $\omega = \frac{2\pi}{\lambda}$. The domain of interest is fixed and given by $[0 \;\; 400\lambda] \times [0 \;\; 200\lambda]$. The point source is located at position $(220\lambda, -40\lambda)$. A total of 6000 point scatterers on average (using a Poisson distribution) are randomly distributed in the domain. This corresponds to a correlation length $l_c \approx 3.65\lambda$. The strength of the scatterers is chosen such that the mean free path is equal to 40λ. As the frequency increases, the transport mean free path decreases as the third power of frequency, as can be seen in (28). We show in Figure 2 a comparison between E_W and E_T at different frequencies with two different sizes of the array of detectors. The average

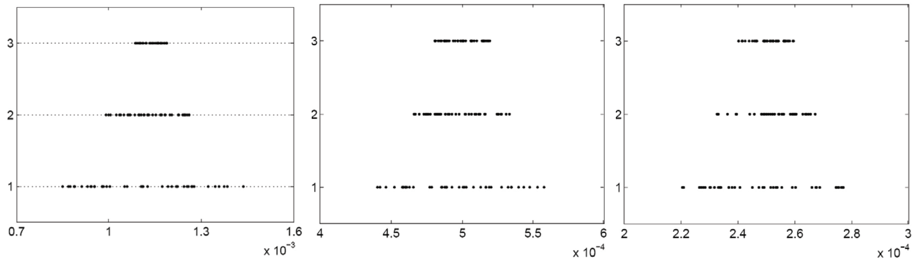


FIG. 3. *Statistical stability of wave data with respect to media properties at three mean free paths:* $c_0\Sigma^{-1} = 30\lambda$ *(left),* $c_0\Sigma^{-1} = 50\lambda$ *(middle), and* $c_0\Sigma^{-1} = 100\lambda$ *(right). Each dot corresponds to the wave energy measured on the array of detectors for one realization with an average of* 6000 *scatterers (top lines labeled* 3*;* $l_c \approx 3.65\lambda$*),* 3000 *scatterers (middle lines labeled* 2*;* $l_c \approx 5.16\lambda$*), and* 1500 *scatterers (bottom lines labeled* 1*;* $l_c \approx 7.30\lambda$*), respectively.*

$\mathbb{E}\{E_W\}$ and standard deviation $\sigma(E_w)$ are calculated based on 40 realizations of the random medium.

We observe that the wave and transport models agree quite well. The ensemble average of the energy density is well captured by the radiative transfer equation. However, radiative transfer models are valid when energy is averaged over a sufficiently large domain compared to the wavelength [4, 8]. When averaging takes place over too small a detector, significant statistical instabilities occur.

These results generalize to the case where an inclusion is present in the random medium. The comparison between the energy densities E_W and E_T is then qualitatively very similar to the case shown in Figure 2.

Statistical stability and density of scatterers. In the next numerical example, we want to address the statistical stability of the transport model with respect to the number of random scatterers. The same average scattering medium, characterized by a given mean free path, may be obtained from a low density of strong scatterers or a high density of weak scatterers in such a way that $\tau^2 n_0$ stays constant. We do not have a theoretical model at present to characterize the statistical stability of the energy E_W when τ and n_0 vary while the product $\tau^2 n_0$ remains constant. Intuitively, however, we expect the random medium to be more mixing, and thus more stable statistically, when the number of scatterers is large simply because the wave fields interact with the underlying structure more often. The following numerical simulations confirm this.

We show in Figure 3 the energy measurements obtained from nine types of random media corresponding to an average of 6000 scatterers ($l_c \approx 3.65\lambda$), 3000 scatterers ($l_c \approx 5.16\lambda$), and 1500 scatterers ($l_c \approx 7.30\lambda$), and to mean free paths $c_0\Sigma^{-1}$ equal to 30λ, 50λ, and 100λ. We observe that, as expected, the standard deviation increases with the correlation length in the medium (as statistical instability increases) and that it increases when the mean free path decreases (as the random medium becomes optically thicker); see the statistics in Table 1.

4. Transport-based imaging in random media. We now examine the capabilities of the radiative transfer model to detect and image inclusions buried in random media. As in the preceding section, all simulations are performed in a two-dimensional setting, which is appropriate for the experimental configuration considered in, e.g., [31]. Note that both the Foldy–Lax model and the Monte Carlo method are independent of dimension, so the proposed numerical method is essentially independent of spatial dimension.

TABLE 1
Average and standard deviation of the wave energy measurements presented in Figure 3.

	$c_0\Sigma^{-1} = 30\lambda$			$c_0\Sigma^{-1} = 50\lambda$			$c_0\Sigma^{-1} = 100\lambda$		
	6000	3000	1500	6000	3000	1500	6000	3000	1500
$\mathbb{E}\{E_W\} \times 10^3$	1.145	1.150	1.150	0.500	0.496	0.495	0.251	0.253	0.249
$\sigma(E_W) \times 10^4$	0.425	0.711	1.156	0.123	0.204	0.383	0.055	0.093	0.181
$\frac{\sigma(E_W)}{\mathbb{E}\{E_W\}} \times 10^2$	3.71	6.18	10.05	2.46	4.11	7.73	2.19	3.69	7.27

As we saw in section 2, the wave energy density is modeled by a radiative transfer equation given by (21) outside of the buried inclusions and by specular reflection conditions (32) at the inclusions' boundary. The inclusions thus become constitutive parameters in the radiative transfer equation, as is the mean free path $c_0\Sigma^{-1}(\mathbf{x})$. In this section, we propose to reconstruct the mean free path and the inclusion in (21)–(32) using energy measurements obtained by solving the Foldy–Lax equations (22)–(24). Although inverse transport models have been used already (see, e.g., [1, 34]), to our knowledge this is the first analysis of reconstructions based on a macroscopic (wavelength-independent) transport model from microscopic (wave- and medium-dependent) wave data.

We consider three slightly different settings as follows: inclusions buried inside a random medium; inclusions separated from the array of detectors by a random medium; and inclusion buried in a random medium and located behind a large blocker that prevents a direct line of sight from the source location. The settings are shown in Figure 4.

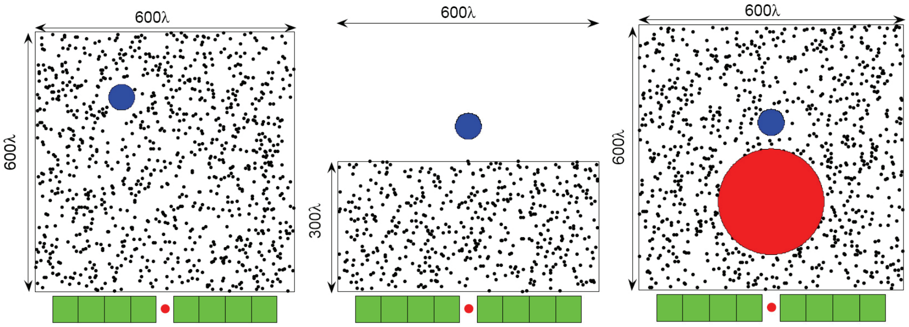


FIG. 4. *Three different setups for the reconstructions: inclusion inside the medium (left), inclusion outside the medium (middle), and inclusion behind a blocker (right). The small circles and the squares represent the source and detector locations, respectively. The small disks are the inclusions to be reconstructed. The large disk represents a blocker, whose location and geometry is assumed to be known.*

The theory of inverse problems for transport equations is relatively well established in the presence of phase-space measurements; see [20]. Here, however, we assume that E_W and E_T are available, not $a(\mathbf{x}, \hat{\mathbf{k}})$ for all $\mathbf{x} \in D$ and $\hat{\mathbf{k}} \in S^{d-1}$. As in reconstructions from knowledge of the Cauchy data in an inverse elliptic problem [26], the recent theoretical result on inverse transport obtained recently in [7] shows that our inverse problem is severely ill-posed. The stability of such inverse problems is notoriously bad, in the sense that noise is drastically amplified during the inversion.

What this means in practice is that the number of degrees of freedom about the random medium and the inclusion that we can possibly retrieve from available data is *small*. The inverse transport problem thus needs to be parameterized. Our choice of a parameterization is the following: we assume that the mean free path $c_0\Sigma^{-1}$ is constant in a rectangular domain, whose geometry is known, and vanishes outside of that domain; and we assume that the inclusion is a disc parameterized by its location $\mathbf{x}$ and its radius R. This hypothesis is slightly relaxed in section 4.3.

4.1. Transport-based imaging method. The reconstruction method for the parameters mentioned above is based on the following least-square minimization procedure. Let $\mathcal{F}$ be the family of parameters we want to reconstruct. We find these parameters by solving the following minimization problem:

$$\mathcal{F}_b = \arg\min_{\mathcal{F}\in\Xi} \mathcal{O}(\mathcal{F}), \tag{35}$$

where $\Xi = [\mathcal{F}_{min}, \mathcal{F}_{max}]$ is a family of a priori box constraints that define a region in which we search for optimal solutions. The objective function that measures the mismatch between measured data and model prediction is given by

$$\mathcal{O}(\mathcal{F}) = \frac{1}{2}\sum_{j=1}^{J} |E_T^j - E_W^j|^2, \tag{36}$$

where E_T^j is the model prediction on detector j, E_W^j the corresponding wave measurement, and J the total number of detectors. We solve this minimization problem by a quasi-Newton minimization algorithm with BFGS updating rules for the Hessian matrix. Box constraints on decision variables are enforced by a gradient projection method. The gradient of the objective function with respect to the parameters to be recovered is calculated by using a finite difference approximation since we have only a few parameters to reconstruct. We refer the interested reader to [16, 32] for details on the BFGS quasi-Newton method and to [34] for an application of the method in inverse transport problems.

In all of the inversions run below, we consider eight detectors of size $62.5\lambda \times 80\lambda$, as depicted in Figure 4. We have mentioned that the energy density was statistically stable only on sufficiently large domains. The size of the detectors thus needs to be sufficiently large to average over local fluctuations. The number of detectors also needs to be sufficiently large to increase the amount of available data. We do not possess a theory for the energy-energy correlations that could guide us in the design of optimal detector arrays. Several scenarios have been tested, and eight is an optimal number of detectors in terms of the statistical stability and the amount of nonredundant information it provides.

The box constraints Ξ have also been chosen to be fairly nonconstraining. For an inclusion of radius $R = 30\lambda$ and location $(300\lambda, 400\lambda)$ in a random medium of size $[0\ 600\lambda]\times[0\ 600\lambda]$, for instance, the constraints on the radius are $R \in [10\lambda, 100\lambda]$ and the constraints on the locations (x, y) are $[50\lambda\ 550\lambda]\times[50\lambda\ 550\lambda]$.

4.2. Imaging under different measurement scenarios. We consider two imaging scenarios: (i) when we have wave energy measurements in the presence of the object; and (ii) when we have energy measurements in the presence *and* in the absence of the inclusion. The former measurements are referred to as *direct* measurements. The latter measurements are referred to as *differential* measurements.

4.2.1. Scenario Ia: Direct measurement. In this first scenario, we are not able to probe the random medium in the absence of an inclusion. We distinguish two subscenarios, which essentially have the same reconstruction capabilities. In scenario Ia, we assume that we have access to measurements in the absence of the inclusion for a given realization of the random medium, which we call medium 1. This allows us to reconstruct the parameters of the random medium, i.e., here the mean free path $c_0\Sigma^{-1}$. We then assume that we have access to measurements in the presence of the inclusion in a medium 2 that is completely uncorrelated to medium 1. This allows us to reconstruct the inclusions' parameters. In scenario Ib, we assume that we have access to the measurements in medium 2 only, and hence in the presence of the inclusion. We have thus to reconstruct all parameters at once. We will present numerical evidence that both scenarios provide very similar reconstruction capabilities. In both cases, the inclusion's influence on the measurements needs to be greater than the noise level coming from our lack of understanding of the specific realization of the random medium. In other words, the inclusion's influence needs be larger than the statistical instability of the radiative transfer model. The reconstruction in scenario Ia is a two-step process, as follows.

Step A. We measure the energy density of waves propagating in one realization of the random medium described above. We then estimate the scattering cross-section Σ by solving the following minimization problem:

$$\Sigma_b = \arg\min_{\Sigma\in\Xi_A} \mathcal{O}(\Sigma), \tag{37}$$

where $\Xi_A = [\Sigma_{min}, \Sigma_{max}]$ is the space in which we seek Σ, and the objective functional is given by

$$\mathcal{O}(\Sigma) = \frac{1}{2}\sum_{j=1}^{J} |E_T^j - E_W^j|^2, \tag{38}$$

where E_T^j is the model prediction detector j, and E_W^j is the corresponding wave energy measurement.

Step B. We now perform the energy measurements in the presence of an inclusion buried in medium 2 uncorrelated with the medium used in Step A. Such a scenario is realistic when we know that medium 2 has statistics similar to medium 1, on which more refined estimates can be obtained before measurements in medium 2 are performed.

We use the scattering coefficient Σ_b (b for best fit) obtained in Step A and image the inclusion from available measurements in medium 2. The position and radius of the inclusion, assumed to be a disc, are obtained by the following minimization:

$$(\mathbf{x}_b, R_b) = \arg\min_{(\mathbf{x},R)\in\Xi_B} \mathcal{O}(\mathbf{x}, R). \tag{39}$$

Here $\mathcal{O}(\mathbf{x}, R)$ is defined as in (36), with E_T^j the transport solutions, with Σ_b the scattering cross-section, and E_W^j the wave energy measurements in medium 2. Here, $\Xi_B \subset \mathbb{R}^{d+1}$ is the constraint set in which the solution $(\mathbf{x}, R)$ is sought.

We show in Figure 5 four typical reconstructions ((A)–(D) from left to right) with this two-step procedure. The reconstructions in (A) and (B) are done for the first configuration in Figure 4, where the medium covers the domain $[0\ 600\lambda]\times[0\ 600\lambda]$ and the inclusion is located inside the medium. A few scatterers around the inclusions are removed from the picture to make the plot clearer. The reconstructions in (C) and (D) are done for the second configuration in Figure 4, where the medium covers the

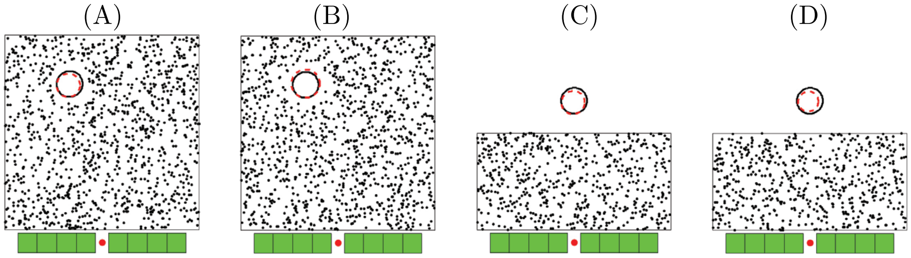


FIG. 5. *Typical reconstruction of an inclusion from direct measurements under scenario* Ia. (A) *and* (C) *are for* $c_0\Sigma^{-1} = 200\lambda$. (B) *and* (D) *are for* $c_0\Sigma^{-1} = 100\lambda$. *Real and reconstructed objects are plotted as solid and dotted circles, respectively.*

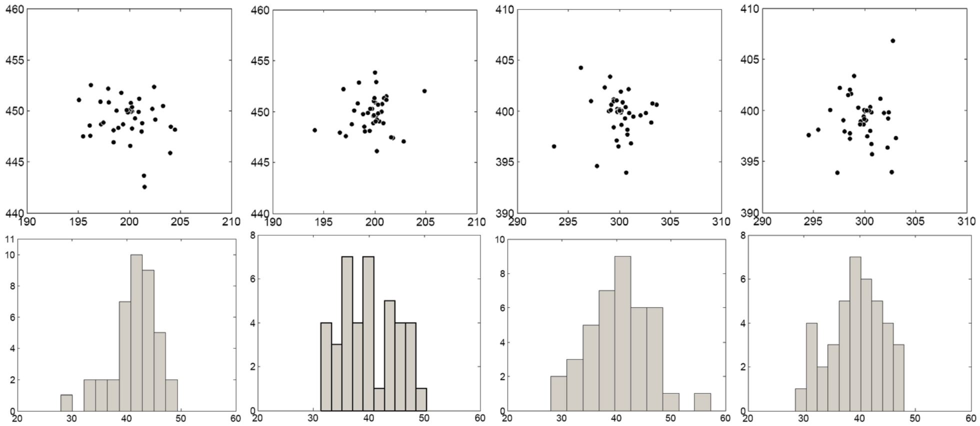


FIG. 6. *Reconstructed parameters for the four cases in Figure* 5 *based on* 40 *realizations. Top row: distribution of the reconstructed inclusion's locations. Bottom row: histogram of the reconstructed radii.*

domain $[0\ 600\lambda]\times[0\ 300\lambda]$ and the inclusion is located outside of the medium. The mean free path (with $c_0 = 1$) for the medium in (A) and (C) is $c_0\Sigma^{-1} = 200\lambda$ and that for the medium in (B) and (D) is $c_0\Sigma^{-1} = 100\lambda$. In all experiments, the correlation length is $l_c \approx 7.75\lambda$, with an average of 6000 rods in experiments (A) and (C) and of 3000 rods in experiments (B) and (D).

The reconstructions are repeated for 40 different realizations of the random medium. The results are presented in Figure 6, where we have plotted the reconstructed locations and radii for the four cases shown in Figure 5. We observe that in all cases, the reconstructions of the inclusion's location are quite accurate. The radii are also good, though not as accurate. Given the smallness (relative to the mean free path) of the inclusions, we do not expect to reconstruct their size very precisely.

We have calculated the first two statistical moments (expectation and standard deviation) of the mean free path, the inclusions' locations (measured by x- and y-coordinates), and the inclusions' radii. The numbers are presented in Table 2. We observe that the reconstruction of the inclusion's location is relatively good, as the variance is quite small. The error in the reconstruction of the inclusion's geometry (i.e., the radius) is, however, significantly larger. The results presented in Table 2 also show that the standard deviations in the x and y variables are somewhat comparable (and on the order of $2-3\lambda$, less than 1% of the distance from the source to the

TABLE 2

Reconstructed mean free paths, locations, and radii from the four cases in Figure 5. *All numbers are in units of the wavelength* λ. ^{a}*Averaged value and standard deviation (numbers in brackets) calculated from* 40 *realizations.*

	Inclusion			Reconstruction^a		
	$c_0\Sigma^{-1}$	location	R	$c_0\Sigma_b^{-1}$	location	R
Case A	200	(200, 450)	40	200.0 [2.7]	(199.8 [2.4], 449.2 [2.1])	41.5 [4.3]
Case B	100	(200, 450)	40	99.8 [2.7]	(199.8 [1.8], 449.8 [1.7])	40.0 [5.1]
Case C	200	(300, 400)	40	200.6 [3.3]	(300.0 [1.8], 399.7 [2.1])	40.4 [6.0]
Case D	100	(300, 400)	40	100.1 [2.9]	(299.8 [1.9], 399.2 [2.4])	39.3 [4.9]

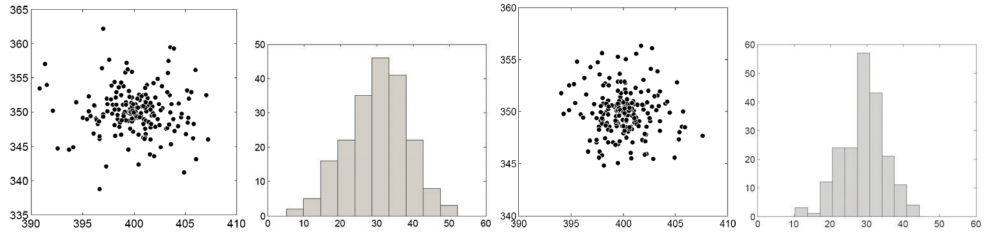


FIG. 7. *Distributions of the reconstructed locations and radii for* 200 *realizations of random media consisting of* N = 3000 *(left two plots) and* N = 6000 *(right two plots) point scatterers on average.*

inclusion). This is an indication that the regime of wave propagation is quite highly mixing so that there is no real privileged direction of propagation.

As we have mentioned in section 3, the statistical stability of the wave energy measurements depends on the number of scatterers for a given mean free path. We now consider the reconstruction of inclusions for two random media with a mean free path equal to $c_0\Sigma^{-1} = 100\lambda$. The first random medium has an average of 6000 point scatterers ($l_c \approx 7.75\lambda$) and the second random medium has an average of 3000 point scatterers ($l_c \approx 10.95\lambda$). The results are presented in Figure 7. We have also calculated the two first statistical moments of the inclusions' location and radius. Results are reported in Table 3. The averaged values are very similar for both reconstructions, which is expected, since radiative transfer is indeed valid for the ensemble averaged energy density. The standard deviation, however, increases when the scatterers become fewer and stronger. In that case, our lack of understanding of the specific realization of the random medium creates large noise in the data, and hence in the reconstructions. Such results are consistent with our numerical analysis of the statistical instability done in section 3.

Let us conclude this section with a remark on the resolution of the method. With random media with 6000 scatterers on average, a mean free path of $c_0\Sigma^{-1} = 100\lambda$ on a square domain of size 600λ, and an inclusion buried 375λ north of the source, we have good reconstructions for radii larger than 20λ. At about 20λ, the "optimal" radius obtained by minimization has a large probability of hitting the box constraints imposed on the radius and is thus meaningless. In the same configuration with a random medium with 3000 scatterers on average, the smallest radius we can confidently reconstruct increases to $R = 30$. These results lead to the following conclusion. Even in the presence of a highly mixing random medium (with 6000 scatterers, which may be large for most practical situations [31]), the inclusion needs to be quite large compared to the wavelength in order for its influence to be larger than

TABLE 3

Reconstructed mean free path, location, and radius from media with $N = 3000$ and $N = 6000$ point scatterers on average. [a]*Averaged value and standard deviation (numbers in brackets) calculated with* 200 *realizations. All numbers are in units of the wavelength λ.*

	Inclusion			Reconstruction[a]		
	$c_0\Sigma^{-1}$	location	R	$c_0\Sigma_b^{-1}$	location	R
$N = 3000$	100	(400, 350)	30	99.8 [4.4]	(400.0 [2.9], 350.2 [3.2])	30.1 [8.4]
$N = 6000$	100	(400, 350)	30	100.0 [2.9]	(400.0 [2.2], 350.0 [2.2])	29.0 [5.9]

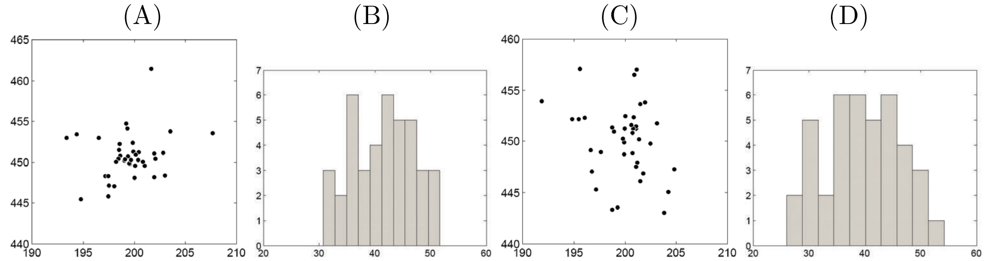


FIG. 8. *Distribution of reconstructed locations and radii with the method in scenario* Ib. *Plots* (A) *and* (B): *for a medium with mean free path $c_0\Sigma^{-1} = 200\lambda$. Plots* (C) *and* (D): *for a medium with $c_0\Sigma^{-1} = 100\lambda$.*

the statistical instability of the wave energy measurements. In such a context, the only solution to improve resolution is to have access to the measurements of scenario II considered below.

4.2.2. Scenario Ib: Reconstructing all at once. We now consider the setting where Σ and the inclusion's parameters are reconstructed in a single step because we do not possess any a priori knowledge of the statistics of the random medium. Compared to scenario Ib, we have one more parameter to reconstruct, and possibly many more parameters when the power spectrum is spatially dependent. However, the reconstruction of these parameters is more adapted to the random medium 2 than in scenario Ia. Reconstructions are based on minimizing the functional,

$$(\Sigma_b, \mathbf{x}_b, R_b) = \arg \min_{(\Sigma, \mathbf{x}, R) \in \Xi} \mathcal{O}(\Sigma, \mathbf{x}, R), \tag{40}$$

where $\Xi \subset \mathbb{R}^3$ is a set of constraints.

This scenario is of equivalent complexity to the previous one since only one additional parameter is added. Moreover, as long as the inclusion is relatively small compared to the size of the domain, it will not affect the reconstruction of Σ in any significant way.

We show in Figure 8 the same reconstructions under scenario Ib as those obtained under scenario Ia in Figures 5(A) and 5(B). The first two statistical moments of the reconstructed locations and radii are listed in Table 4. We observe that the reconstructed mean free paths, inclusion's positions, and radii are very similar to the parameters reconstructed in Figure 5.

4.2.3. Scenario II: Differential measurement. Scenario II relies on much different measurements. We assume that we have access to wave energy measurements in the presence and in the absence of the inclusion, and in both cases for the *same* realization of the random medium (except at the location of the inclusion, where the

Table 4

Reconstructed mean free path, location, and radius in scenario Ib *for two different type of media.* [a] *Averaged value and standard deviation (numbers in brackets) calculated from* 40 *realizations. All numbers are in units of the wavelength* λ.

	Inclusion			Reconstruction[a]		
	$c_0\Sigma^{-1}$	location	R	$c_0\Sigma_b^{-1}$	location	R
Medium 1	200	(200, 450)	40	200.6 [3.1]	(199.5 [2.6], 450.7 [2.8])	41.5 [5.7]
Medium 2	100	(200, 450)	40	99.0 [5.0]	(199.5 [3.4], 450.1 [3.5])	39.7 [7.0]

random scatterers are suppressed). However, we do not know the random medium and thus have to model it macroscopically using a radiative transfer model. As in the preceding case, the macroscopic model is parameterized by a unique parameter, the mean free path $c_0\Sigma^{-1}$.

As we said earlier, differential measurements allow for much more accurate reconstructions. The reason is that the difference of the two measured energies depends only on the inclusion. Therefore, with a kinematic picture in mind, where wave energy packets are replaced by particles scattering in the random media, all the wave packets that do not interact with the inclusion do not contribute to differential measurements. These wave packets are the largest contributors to the statistical instability of the random medium that hampers reconstructions of small objects in scenario I. The measured wave packets that have interacted with the inclusion are also statistically unstable. However, they are in some sense proportional to the inclusion, and in the absence of external measurement noise, differential measurements allow one to reconstruct arbitrarily small objects. They are not immune to the statistical instability, and a statistical instability in the random medium of 10% may result in an error on the location and radius of the inclusion on the order of 10% as well. However, the limit in the size of the objects that can be reconstructed is governed by external measurement noise and no longer by the statistical instability in the medium. The two-step reconstruction process used in scenario Ia applies here as follows.

Step A. We use the measurements in the absence of an inclusion in the medium to estimate the scattering cross-section of the medium. This is done by minimizing (37) as before.

Step B. Once Σ has been found, we reconstruct the location ($\mathbf{x}$) and the radius (R) of the spherical inclusion by minimizing,

$$(\mathbf{x}_b, R_b) = \arg \min_{(\mathbf{x},R)\in\Xi_B} \delta\mathcal{O}(\mathbf{x}, R), \tag{41}$$

where

$$\delta\mathcal{O}(\mathbf{x}, R) = \frac{1}{2}\sum_{j=1}^{J} |\delta E_T^j - \delta E_W^j|^2. \tag{42}$$

Here, δE_T^j and δE_W^j correspond to the difference of energies with and without the inclusion for the transport model and the wave data, respectively, at detector j. In practice, δE_W is calculated by estimating the difference of the solutions to two Foldy–Lax equations, and δE_W is estimated by Monte Carlo using the variance reduction technique introduced in [10]. The role of this variance reduction is to write the difference δE_W as the expectation of an appropriate process, rather than the difference of two expectations, which requires a huge amount of particles to be accurate. We refer the reader to [10] for additional details.

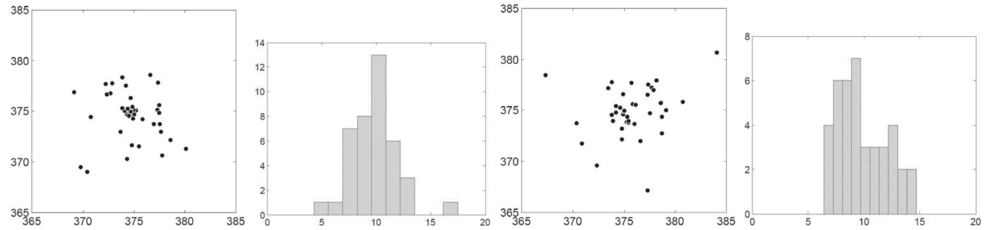


FIG. 9. *Distribution of reconstructed location and radius for* 40 *realizations with differential measurement. Left two plots:* $c_0\Sigma^{-1} = 200\lambda$. *Right two plots:* $c_0\Sigma^{-1} = 100\lambda$.

TABLE 5

Reconstructed mean free paths, locations, and radii based on differential measurements. [a]*Averaged value and standard deviation (numbers in brackets) calculated from* 40 *realizations. All numbers are in units of the wavelength* λ.

	Real			Reconstruction[a]		
	$c_0\Sigma^{-1}$	location	R	$c_0\Sigma_b^{-1}$	location	R
Medium 1	200	(375, 375)	10	199.8 [2.8]	(374.8 [2.4], 374.6 [2.4])	9.9 [2.3]
Medium 2	100	(375, 375)	10	100.1 [3.3]	(375.8 [2.9], 375.0 [2.4])	9.9 [2.2]

Reconstructions based on scenario II have been performed for two different mean free paths $c_0\Sigma^{-1} = 200\lambda$ and $c_0\Sigma^{-1} = 100\lambda$, respectively. The random medium is again formed of an average of 6000 scatterers over $[0\ 600\lambda]\times[0\ 600\lambda]$ so that the correlation length $l_c \approx 7.75\lambda$. The inclusion is located at coordinates $(375\lambda, 375\lambda)$ and has a radius equal to $R = 10\lambda$, which is significantly smaller than what we can reconstruct under scenario I.

We show in Figure 9 the distributions of the reconstructed locations and radii for the two random media.

The averaged value and standard deviation of the mean free path and the inclusion's location and radius are summarized in Table 5. As we can observe, the transport inversion does a relatively good job at locating the inclusion. However, it misses the size of the inclusion by a much larger amount. This is understandable because the specular reflection model may not be totally consistent with the size of an object with radius of order 10λ. For smaller objects, a more accurate radiation model than specular reflection is necessary. We do not consider this issue further here.

Reconstructions based on differential measurement may be used in the monitoring of cluttered areas, where we have access to energy measurements before and after the inclusion is present.

4.3. Imaging the orientation of the inclusions. In the preceding sections, we have obtained satisfactory reconstructions of inclusions based on low-dimensional parameterizations of the inclusion. Since, as we have mentioned in section 4, the inverse transport problem is quite ill-posed, we should not expect to reconstruct any fine geometrical information about the inclusion. To demonstrate this, we consider the reconstruction of half discs from wave energy measurements. Half discs are parameterized by their location, their radius, and the orientation θ of their flat portion (with respect to the x-axis). We use the same minimization techniques as in the preceding sections with this additional parameter θ in the functional in (36) in the case of direct measurements and in (42) in the case of differential measurements.

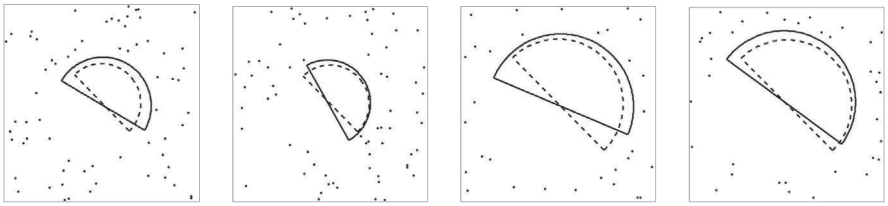


FIG. 10. *Reconstructions of half discs from direct and differential measurements in random media with an average of* 6000 *scatterers and a mean free path* $c_0\Sigma^{-1} = 100\lambda$*. From left to right: A typical reconstruction from differential measurement; an ensemble averaged reconstruction from differential measurement; a typical reconstruction from direct measurement; and an ensemble averaged reconstruction from direct measurement. True inclusions are plotted with dashed lines and reconstructed inclusions with solid lines.*

We show in Figure 10 typical reconstructions obtained from direct and differential measurements. The radius of the inclusion is $R = 30\lambda$ for the case of differential measurement and $R = 50\lambda$ for the case of direct measurement. In both cases, the half disc is centered at position $(375\lambda, 375\lambda)$. The reconstruction of the orientation is relatively accurate. Based on simulations on 20 realizations, the averaged reconstructed angle is 0.68π and the standard deviation is 0.16π for the case with differential measurements. The average angle is 0.81π and the standard deviation is 0.19π for the case with direct measurements. In each case, the exact angle is $\frac{3\pi}{4}$. The standard deviations on the orientation thus correspond to roughly 20% of the error. The reconstruction of the orientation is, however, not as accurate as that of the radius (or equivalently, the volume) and the location of the inclusion. The orientation is a finer geometric property of the inclusion and is thus more difficult to observe. Even with differential measurement, our experience is that we cannot faithfully reconstruct the orientation of half disks with a radius smaller than 25λ. For direct measurement, we can reconstruct the orientation when the radius of the inclusion is larger than 45λ. Below these numbers, the reconstruction of θ becomes extremely noisy.

4.4. Imaging in the presence of blockers.

In all of the above reconstructions, the mean free path is sufficiently small so that the energy leaving the source, hitting the inclusion, and coming back to the detectors without having interacted with the underlying medium, i.e., the energy of the coherent wave field, is relatively small. Inversions based only on the coherent information may thus fail to provide meaningful information about the inclusion. One may, however, use larger wavelengths, which are less affected by the random medium since the mean free path is much larger as the latter scales like λ^3. The assumption in scenarios I and II above is that we have access to wave measurements at frequencies for which there is considerable multiple scattering, i.e., for which the mean free path is relatively small.

There are, however, situations in which the coherent wave field can hardly be used no matter which frequency we consider, for instance, when the inclusion we seek to image is hidden by a large blocker. One can always argue that some energy reaches the hidden inclusion by diffractive effects. Such fields, however, are quite weak according to the geometric theory of diffraction and may well be below noise level as soon as the propagating medium has unknown spatial fluctuations.

In such situations, randomness in the underlying medium may be helpful. In the context considered in this paper of random media with multiple localized scatterers, the wave energy may be modeled by a radiative transfer equation, and both the

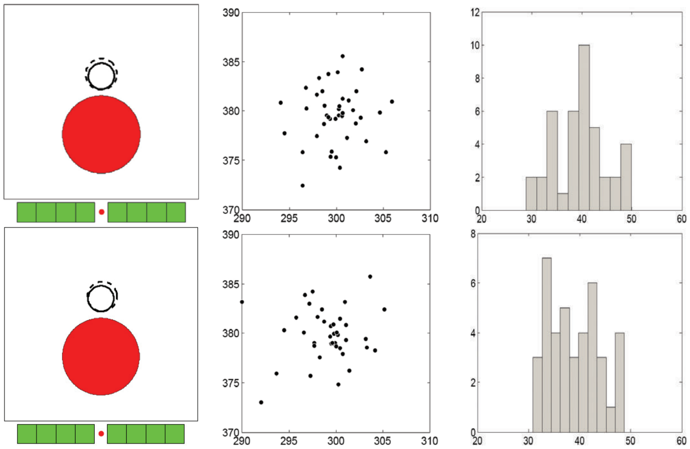


Fig. 11. *Reconstruction of an inclusion hidden behind a large blocker. Top row: reconstruction in scattering media with* $c_0\Sigma^{-1} = 100\lambda$. *Bottom row: reconstruction in scattering media with* $c_0\Sigma^{-1} = 75\lambda$. *From left to right: a typical reconstruction, distribution of the reconstructed locations, and histogram of the reconstructed radius.*

Table 6

Reconstructed mean free path, location, and radius in presence of blockers. [a]*Average and standard deviation (numbers in bracket) calculated from* 40 *reconstructions. All numbers are in units of the wavelength* λ.

	Inclusion			Reconstruction[a]		
	$c_0\Sigma^{-1}$	location	R	$c_0\Sigma_b^{-1}$	location	R
Medium 1	100	(300, 380)	40	101.0 [2.9]	(299.9 [2.6], 379.6 [2.9])	39.7 [5.3]
Medium 2	75	(300, 380)	40	74.8 [3.3]	(299.1 [3.1], 379.9 [2.6])	39.0 [5.2]

blocker and the unknown inclusion may be modeled as constitutive parameters in that equation. We consider here a situation where the blocker is large, spherical, and *known*. It is sufficiently large to block direct paths from the source term to the inclusion. The blocker is treated as any other extended inclusion in the Foldy–Lax and transport models.

We consider the setup shown on the right in Figure 4. The blocker is located at $(300\lambda, 200\lambda)$ and its radius is 120λ. The inclusion's center and radius are $(300\lambda, 380\lambda)$ and $R = 40\lambda$. All reconstructions are done based on differential measurements, i.e., under scenario II. The mean free path is estimated in the presence of the known blocker and the inclusion's parameters minimizing (42). Figure 11 shows reconstructions in random media with (theoretical) mean free paths equal to $c_0\Sigma^{-1} = 100\lambda$ and $c_0\Sigma^{-1} = 75\lambda$, respectively. In each case, there would be a number of scatterers equal to 6000 on average if the blocker was filled with scatterers so that $l_c \approx 7.75\lambda$. We are therefore in the more stable of the two random media considered so far.

The average and standard deviations of the reconstructed parameters are shown in Table 6. We observe quite good reconstruction capabilities. The images obtained in the presence of the blocker are of a quality comparable to those obtained in the absence of the blocker. Since the blocker is known and not so large so that no energy radiated from the source can reach the inclusion and come back to the array of detectors, this is consistent with what one expects from theoretical considerations.

5. Conclusions. We have derived a radiative transfer equation to model the energy density of mono-frequency waves propagating in random media composed of localized scatterers. We have shown the validity of the model based on numerical simulations, provided that the medium is sufficiently mixing. In our context, this means that the number of scatterers needs to be sufficiently large so that sufficient mixing occurs. Otherwise, the energy density becomes statistically less stable, i.e., depends more on the realization of the random medium.

When the medium is sufficiently mixing, the radiative transfer model is sufficiently stable for imaging purposes. Inclusions buried in the random medium are modeled as a constitutive parameter in the transport equation. We have shown numerical evidence that the inverse transport method indeed allows for accurate reconstruction of sufficiently large inclusions from wave energy measurements. Because inverse transport problems are quite ill-posed, in the sense that noise may be drastically amplified during the reconstruction, we have parameterized the inclusion by a small number of parameters, typically its location and its radius for spherical inclusions.

For smaller inclusions whose influence falls below the noise level coming from the statistical instability of the random medium, we have shown that differential measurements, i.e., wave energy measurements in the presence and in the absence of the inclusion, allowed for accurate reconstructions. Because the inclusions are modeled by specular reflections of the wave energy at their boundaries, the inverse model works for inclusions that are significantly larger than the wavelengths. For smaller inclusions, the model needs to be modified and the inclusion treated as a point source in the transport model with an appropriate radiation pattern that depends on geometry (an isotropic radiation pattern for a small spherical inclusion).

The imaging methods developed in this paper have been validated with real-world experimental data. The results are reported elsewhere [5].

Acknowledgments. We would like to thank the anonymous referees for their helpful comments on the first draft of this paper.

REFERENCES

[1] S. R. Arridge, *Optical tomography in medical imaging*, Inverse Problems, 15 (1999), pp. R41–R93.

[2] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[3] G. Bal, *Kinetics of scalar wave fields in random media*, Wave Motion, 43 (2005), pp. 132–157.

[4] G. Bal, *On the self-averaging of wave energy in random media*, Multiscale Model. Simul., 2 (2004), pp. 398–420.

[5] G. Bal, L. Carin, D. Liu, and K. Ren, *Experimental validation of a transport-based imaging method in highly scattering environments*, Inverse Problems, 23 (2007), pp. 2527–2539.

[6] G. Bal, T. Komorowski, and L. Ryzhik, *Self-averaging of Wigner transforms in random media*, Comm. Math. Phys., 242 (2003), pp. 81–135.

[7] G. Bal, I. Langmore, and F. Monard, *Inverse transport with isotropic sources and angularly averaged measurements*, Inverse Probl. Imaging, 2 (2008), pp. 23–42.

[8] G. Bal, G. Papanicolaou, and L. Ryzhik, *Self-averaging in time reversal for the parabolic wave equation*, Stoch. Dyn., 4 (2002), pp. 507–531.

[9] G. Bal and O. Pinaud, *Time-reversal-based detection in random media*, Inverse Problems, 21 (2005), pp. 1593–1620.

[10] G. Bal and O. Pinaud, *Accuracy of transport models for waves in random media*, Wave Motion, 43 (2006), pp. 561–578.

[11] G. Bal and O. Pinaud, *Kinetic models for imaging in random media*, Multiscale Model. Simul., 6 (2007), pp. 792–819.

[12] G. Bal and L. Ryzhik, *Time reversal and refocusing in random media*, SIAM J. Appl. Math., 63 (2003), pp. 1475–1498.

[13] G. Bal and R. Verástegui, *Time reversal in changing environments*, Multiscale Model. Simul., 2 (2004), pp. 639–661.

[14] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Amer., 111 (2002), pp. 230–248.

[15] L. Borcea, G. Papanicolaou, and C. Tsogka, *Interferometric array imaging in clutter*, Inverse Problems, 21 (2005), pp. 1419–1460.

[16] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput., 16 (1995), pp. 1190–1208.

[17] J. D. Benamou, F. Castella, T. Katsaounis, and B. Perthame, *High frequency limit of the Helmholtz equations*, Rev. Mat. Iberoamericana, 18 (2002), pp. 187–209.

[18] S. Chandrasekhar, *Radiative Transfer*, Dover Publications, New York, 1960.

[19] M. Cheney and R. J. Bonneau, *Imaging that exploits multipath scattering from point scatters*, Inverse Problems, 20 (2004), pp. 1691–1711.

[20] M. Choulli and P. Stefanov, *Reconstruction of the coefficients of the stationary transport equation from boundary measurements*, Inverse Problems, 12 (1996), pp. L19–L23.

[21] J. F. Claerbout, *Fundamentals of Geophysical Data Processing: With Applications to Petroleum Prospecting*, Blackwell Scientific, Palo Alto, CA, 1985.

[22] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, New York, 1998.

[23] L. Erdös and H. T. Yau, *Linear Boltzmann equation as the weak coupling limit of a random Schrödinger equation*, Comm. Pure Appl. Math., 53 (2000), pp. 667–735.

[24] M. Fink, *Time reversed acoustics*, Physics Today, 50 (1997), pp. 34–40.

[25] L. O. Foldy, *The multiple scattering of waves*, Phys. Rev., 67 (1945), pp. 107–119.

[26] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[27] A. Ishimaru, *Wave Propagation and Scattering in Random Media*, IEEE Press, New York, 1997.

[28] A. Ishimaru, S. Jaruwatanadilok, and Y. Kuga, *Short pulse detection and imaging of objects behind obscuring random layers*, Waves Random Complex Media, 16 (2006), pp. 509–520.

[29] M. Lax, *Multiple scattering of waves*, Rev. Modern Physics, 23 (1951), pp. 287–310.

[30] D. Liu, G. Kang, L. Li, Y. Chen, S. Vasudevan, W. Joines, Q. H. Liu, J. Krolik, and L. Carin, *Electromagnetic time-reversal imaging of a target in a cluttered environment*, IEEE Trans. Antennas and Propagation, 53 (2005), pp. 3508–3566.

[31] D. Liu, S. Vasudevan, J. Krolik, G. Bal, and L. Carin, *Electromagnetic time-reversal source localization in changing media: Experiment and analysis*, IEEE Trans. Antennas and Propagation, 55 (2007), pp. 344–354.

[32] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.

[33] J. Przybilla, M. Korn, and U. Wegler, *Radiative transfer of elastic waves versus finite difference simulations in two-dimensional random media*, J. Geophys. Res., 111 (2006), B04305.

[34] K. Ren, G. Bal, and A. H. Hielscher, *Frequency domain optical tomography based on the equation of radiative transfer*, SIAM J. Sci. Comput., 28 (2006), pp. 1463–1489.

[35] L. Ryzhik, G. Papanicolaou, and J. B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion, 24 (1996), pp. 327–370.

[36] P. Sebbah, ed., *Waves and Imaging Through Complex Media*, Kluwer, Dordrecht, The Netherlands, 2003.

[37] P. Sheng, *Introduction to Wave Scattering, Localization and Mesoscopic Phenomena*, Academic Press, New York, 1995.

[38] H. Spohn, *Derivation of the transport equation for electrons moving through random impurities*, J. Statist. Phys., 17 (1977), pp. 385–412.

[39] J. Spanier and E. M. Gelbard, *Monte Carlo Principles and Neutron Transport Problems*, Addison-Wesley, Reading, MA, 1969.

[40] J. R. Taylor, *Scattering Theory*, Wiley, New York, 1972.

SIAM J. Appl. Math.
Vol. 68, No. 6, pp. 1763–1782

TIME-LOCAL DISSIPATIVE FORMULATION AND STABLE NUMERICAL SCHEMES FOR A CLASS OF INTEGRODIFFERENTIAL WAVE EQUATIONS[*]

C. CASENAVE[†] AND E. MONTSENY[†]

Abstract. We consider integrodifferential equations of the abstract form $\mathbf{H}(\partial_t)\Phi = \mathbf{G}(\nabla)\Phi + f$, where $\mathbf{H}(\partial_t)$ is a diagonal convolution operator and $\mathbf{G}(\nabla)$ is a linear anti-selfadjoint differential operator. On the basis of an original approach devoted to integral causal operators, we propose and study a time-local augmented formulation under the form of a Cauchy problem $\partial_t\Psi = \mathcal{A}\Psi + \mathcal{B}f$ such that $\Phi = \mathcal{C}\Psi$. We show that under a suitable hypothesis on the symbol $\mathbf{H}(p)$, this new formulation is dissipative in the sense of a natural energy functional. We then establish the stability of numerical schemes built from this time-local formulation, thanks to the dissipation of appropriate discrete energies. Finally, the efficiency of these schemes is highlighted by concrete numerical results relating to a model recently proposed for 1D acoustic waves in porous media.

Key words. integrodifferential equation, partial differential equation, convolution operator, diffusive representation, numerical scheme, Cauchy problem, energy functional, stability condition

AMS subject classifications. 45K05, 65J10, 35L99

DOI. 10.1137/070693710

1. Introduction.

In many physical problems where accurate dynamic models are required, the contribution of some underlying and more or less ill-known distributed phenomena cannot be neglected. Although the precise local description of such phenomena often appears excessively complex or even, in many cases, out of scope, fortunately most of the time their macroscopic dynamic consequences can be taken into account by means of suitable time-operators of a convolution nature, which in fact summarize the collective contribution of many hidden parameters to the global dynamic behavior of the quantities of interest. In that sense, such *integrodifferential* models therefore conciliate accuracy and simplicity, up to the loss of the so-called time-locality property. In opposition to standard Cauchy problems, for which the future is conditioned by the present only, here all past evolution is involved via the time-convolution. In past years, various problems relating to integrodifferential models have been studied in many fields. As examples we can cite [2, 7, 13, 16] in physics, [6, 9, 12] in mathematical analysis or numerical simulation, [1, 11] in control problems, [3, 10] in electrical engineering, [18] in biology, etc.

In the particular context of partial integrodifferential equations, the crucial problem of numerical simulation is in general quite difficult. This is due in one part to the numerical complexity of quadratures of convolution integrals, which generate highly expensive time discretizations, particularly when long memory components are present. Beyond this first heavy shortcoming, the stability of numerical schemes is in general very difficult to get, namely because standard techniques devoted to (ordinary) partial differential equations such as energy dissipation cannot be used for integrodifferential equations. So, the construction of stable numerical schemes remains an important challenge, and it can be expected that some specific methods

[*] Received by the editors June 5, 2007; accepted for publication (in revised form) March 24, 2008; published electronically July 16, 2008.
http://www.siam.org/journals/siap/68-6/69371.html

[†] Laboratoire d'Analyse et d'Architecture des Systèmes, LAAS-CNRS, University of Toulouse, 31077 Toulouse, France (casenave@laas.fr, emontseny@laas.fr).

devoted to analysis and approximation of convolution operators should be of great help. This is the topic of the present paper.

In what follows, we consider partial integrodifferential equations of the abstract form

$$\mathbf{H}(\partial_t)\Phi = \mathbf{G}(\nabla)\Phi + f \quad \text{on } (t,x) \in \mathbb{R}_t^{+*} \times \mathbb{R}_x^n, \tag{1.1}$$

where $\mathbf{H}(\partial_t)$ is an invertible[1] diagonal convolution operator, and $\mathbf{G}(\nabla)$ is an anti-selfadjoint linear differential operator. Many propagation phenomena can be modelled following (1.1). As significant examples, we can mention, for example, electromagnetic waves in dissipative media [13], wave propagation in viscoacoustic media [6], etc. In order to illustrate our results, we will consider in particular the following model of 1D acoustic waves in a porous wall proposed in [5]: $(H_1(\partial_t)u, H_2(\partial_t)P)^T = (-\partial_x P, -\partial_x u)^T + f$, where u and P stand for the velocity and the pressure of the gas, and the symbols H_i take the form $H_1(p) = k\,p + a\sqrt{1+b\,p}$, $H_2(p) = k'p + \frac{c\,p^2}{p+a'\sqrt{1+b'p}}$.

On the basis of an original approach devoted to integral causal operators presented in [14, 15] and successfully applied to various integrodifferential problems, namely in [1, 2, 3, 10], we propose and study a new formulation, both equivalent to (1.1) and time-local, written as the following Cauchy problem:

$$\partial_t\Psi = \mathcal{A}\Psi + \mathcal{B}f \quad \text{on } (t,x,\xi) \in \mathbb{R}_t^{+*} \times \mathbb{R}_x^n \times \mathbb{R}_\xi, \quad \Psi(0,.,.) = 0, \tag{1.2}$$

in such a way that the solution of (1.1) is expressed as $\Phi = \mathcal{C}\Psi$. We show in particular that under a natural hypothesis on the symbol $\mathbf{H}(p)$, the formulation (1.2) is dissipative in the sense of an energy functional derived, in some way, from the one of the standard equation $\partial_t\Phi = \mathbf{G}(\nabla)\Phi$. Following a convenient method introduced in [14], straightforward dissipative approximate versions of (1.2) are deduced by simple discretization of the auxiliary variable ξ. We then study numerical schemes based on classical discretizations relating to the variables t, x, and we establish their stability in the sense of adapted energy functionals inherited from the continuous model.

The paper is organized as follows. Section 2 deals with the time-local formulation of (1.1). It begins with a short presentation of the so-called diffusive representation of causal integral operators introduced in [14]; then, the formulation (1.2) is deduced and its dissipativity is established. In section 3, implicit and explicit numerical schemes for (1.2) are stated and studied from the point of view of stability. Finally, the efficiency of these schemes is highlighted in section 4 by means of some numerical simulations.

2. Time-local formulation of integrodifferential equations.

2.1. Time-local realization of causal convolution operators.

In this section, we present a particular case of a methodology called diffusive representation, introduced and developed in [14] in a general framework.

We consider a causal convolution operator denoted by $K(\partial_t)$, that is, for any continuous function $w : \mathbb{R}^+ \to \mathbb{R}$,

$$(K(\partial_t)w)(t) = \int_0^t k(t-s)\,w(s)\,ds = (k * w)(t); \tag{2.1}$$

the function $K = \mathcal{L}k$ (the Laplace transform of k) is called the symbol of operator $K(\partial_t)$.

[1] We implicitly refer to an underlying algebra of causal convolution operators. For example, for a Cauchy problem on $\mathbb{R}_t^+$ with null initial condition, the inverse of $\mathbf{H}(\partial_t) = \partial_t$ is $\partial_t^{-1} : v \mapsto \int_0^t v\,ds$.

Let $w^t(s) = \mathbf{1}_{]0,t]}(s)\, w(s)$, and let $w_t(s) = w^t(t-s)$ be the so-called history of w. From causality of $K(\partial_t)$, we deduce

$$\left(K(\partial_t)(w - w^t)\right)(t) = 0 \ \forall t;$$

then, we have for any continuous function w,

$$(K(\partial_t)w)(t) = \left[\mathcal{L}^{-1}\left(K\,\mathcal{L}w\right)\right](t) = \left[\mathcal{L}^{-1}\left(K\,\mathcal{L}w^t\right)\right](t). \tag{2.2}$$

We then define

$$\Psi_w(t,p) := \mathrm{e}^{pt}\left(\mathcal{L}w^t\right)(p) = \left(\mathcal{L}w_t\right)(-p); \tag{2.3}$$

by computing $\partial_t \mathcal{L} w_t$, and using Laplace inversion and (2.2), the following lemma can be shown.

Lemma 2.1.

1. *The function Ψ_w is a solution of the differential equation*

$$\partial_t \Psi(t,p) = p\,\Psi(t,p) + w, \ \ t > 0, \ \ \Psi(0,p) = 0, \ \ p \in \mathbb{C}. \tag{2.4}$$

2. *There exists $b_0 \in \mathbb{R}$ such that*

$$\forall b \geqslant b_0,\ \left(K(\partial_t)w\right)(t) = \frac{1}{2\mathrm{i}\pi}\int_{b-\mathrm{i}\infty}^{b+\mathrm{i}\infty} K(p)\,\Psi_w(t,p)\,dp. \tag{2.5}$$

Proof.

1. From (2.3), we have $\Psi_w(t,p) := \mathrm{e}^{pt}\int_0^t \mathrm{e}^{-ps}\, w(s)ds$, and so

$$\partial_t \Psi_w(t,p) = p\,\mathrm{e}^{pt}\int_0^t \mathrm{e}^{-ps}\, w(s)ds + \mathrm{e}^{pt}\,\mathrm{e}^{-pt}\, w(t).$$

2. From (2.2), there exists $b_0 \in \mathbb{R}$ such that for any $b \geqslant b_0$,

$$\left(K(\partial_t)w\right)(t) = \frac{1}{2\mathrm{i}\pi}\int_{b-\mathrm{i}\infty}^{b+\mathrm{i}\infty} \mathrm{e}^{pt}\, K(p)\ \left(\mathcal{L}w^t\right)(p)\,dp = \frac{1}{2\mathrm{i}\pi}\int_{b-\mathrm{i}\infty}^{b+\mathrm{i}\infty} K(p)\,\Psi_w(t,p)\,dp. \qquad \square$$

We denote by Ω the holomorphic domain of K. Let γ be a simple arc closed at infinity and included in $\mathbb{C}^- = \mathbb{R}^- + \mathrm{i}\mathbb{R}$. We denote by Ω_γ^+ the exterior domain defined by γ, and denote by Ω_γ^- the complementary of $\overline{\Omega_\gamma^+}$ (see Figure 2.1). By use of standard techniques (Cauchy theorem, Jordan lemma), the following lemma can be shown.

Lemma 2.2. *For $\gamma \subset \Omega$ such that K is holomorphic in Ω_γ^+, if $K(p) \to 0$ when $p \to \infty$ in Ω_γ^+, then, for any closed simple arc $\tilde{\gamma}$ in Ω_γ^+ such that $\gamma \subset \Omega_{\tilde{\gamma}}^-$ (see Figure 2.1),*

$$\left(K(\partial_t)w\right)(t) = \frac{1}{2\mathrm{i}\pi}\int_{\tilde{\gamma}} K(p)\,\Psi_w(t,p)\,dp. \tag{2.6}$$

We now suppose that γ, $\tilde{\gamma}$ are defined by functions of $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R};\mathbb{C})$, also denoted γ, $\tilde{\gamma}$. From classical techniques, the following has been shown in [14].

Theorem 2.3. *Under the hypothesis of Lemma 2.2, if in addition the possible singularities of K on γ are simple poles or branching points in the neighborhood of which $|K \circ \gamma|$ is locally integrable, then*

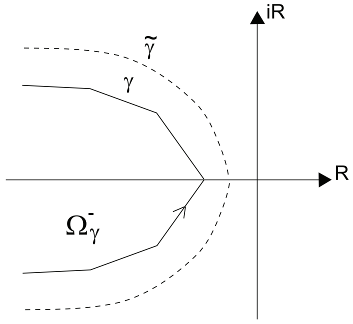


FIG. 2.1. *Example of γ and $\tilde{\gamma}$ arcs.*

1. *with* $\tilde{\nu} = \frac{\tilde{\gamma}'}{2i\pi} K \circ \tilde{\gamma}$ *and* $\tilde{\psi}(t,.) = \Psi_w(t,.) \circ \tilde{\gamma}$,

$$(K(\partial_t)w)(t) = \int_{\mathbb{R}} \tilde{\nu}(\xi)\, \tilde{\psi}(t,\xi)\, d\xi; \tag{2.7}$$

2. *if* $\tilde{\gamma}_n \to \gamma$ *in* $W^{1,\infty}_{\text{loc}}$, *then* $\frac{\tilde{\gamma}_n'}{2i\pi} K \circ \tilde{\gamma}_n \to \nu$ *in the sense of measures;*
3. $\psi(t,.) = \Psi_w(t,.) \circ \gamma$ *is the unique solution of the following Cauchy problem on* $(t,\xi) \in \mathbb{R}^{*+} \times \mathbb{R}$*:*

$$\partial_t \psi(t,\xi) = \gamma(\xi)\, \psi(t,\xi) + w(t),\ \psi(0,\xi) = 0 \tag{2.8}$$

and

$$(K(\partial_t)w)(t) = \langle \nu, \psi(t,.)\rangle\,. \tag{2.9}$$

For convenience in what follows, we will indifferently denote by $\langle \nu, \psi\rangle$ or $\int \nu\, \psi\, d\xi$ the duality product between a continuous function ψ and a measure ν (in particular, for Dirac measures, $\psi(a) = \int \delta_a \psi\, d\xi$).

Remark 1. In the limit case $\gamma(\xi) = -|\xi|$, we have $\Omega^-_\gamma = \varnothing$. The above results remain valid, and we deduce from the symmetry of the problem that there exists a measure μ such that

$$\int_{-\infty}^{+\infty} \nu\, \psi\, d\xi = \int_0^{+\infty} \mu\, \psi\, d\xi.$$

This particular case will be useful in practice when K is holomorphic in $\mathbb{C} \setminus \mathbb{R}^-$.

DEFINITION 2.4 (see [14]). *The measure ν defined in Theorem* (2.3) *is called the γ-symbol of operator $K(\partial_t)$.*

In many cases, the arc γ can be constrained to satisfy a suitable additional condition which makes (2.8) of diffusive type [14]. The main advantage of the input-output formulation (2.8), (2.9) lies in its time-local nature which allows us to use classical methods devoted to Cauchy problems. In particular, stable and efficient schemes for the numerical resolution of (1.1) can be straightforwardly built from discretizations of problem (2.8) following standard techniques. This is the topic of the following sections.

2.2. Application to a class of partial integrodifferential equations. We consider the problem

$$\mathbf{H}(\partial_t)\Phi = \mathcal{G}\Phi + f \quad \text{on } \mathbb{R}_t^+ \times \Omega \quad \Omega \subset \mathbb{R}_x^m, \tag{2.10}$$

where $\Phi = (\Phi_1, \dots, \Phi_M)^T$ is the unknown, $\mathbf{H}(\partial_t)$ is an invertible causal convolution operator of the form

$$\mathbf{H}(\partial_t) = \begin{bmatrix} H_1(\partial_t) & & \\ & \ddots & \\ & & H_M(\partial_t) \end{bmatrix}, \tag{2.11}$$

and $\mathcal{G} = \mathbf{G}(\nabla)$ is a differential operator assumed to be anti-selfadjoint, that is,

$$\mathcal{G}_{ij}^* = -\mathcal{G}_{ji}, \tag{2.12}$$

where $\mathcal{G}_{ij}^*$ is defined by

$$\left(\mathcal{G}_{ij}\, u \middle|\, v\right)_{L^2(\Omega)} = \left(u \,\middle|\mathcal{G}_{ij}^*\, v\right)_{L^2(\Omega)} \quad \forall u, v \in \mathcal{D}(\Omega).$$

As usual, suitable boundary conditions associated with $\mathcal{G}$, not expressed here, can complete the model (2.10). The γ_i-symbols ν_i of operators $H_i(\partial_t)^{-1}$ are assumed to be positive measures. Note that this property appears as physically realistic in the sense of an energy balance, as it will be highlighted later.

By expressing (2.10) under the form $\Phi = \mathbf{H}(\partial_t)^{-1}\left(\mathcal{G}\Phi + f\right)$, we formally deduce from the results of section 2.1, under a suitable hypothesis on $H_i^{-1}(\partial_t)$, the following diffusive time-local formulation of (2.10):

$$\partial_t\psi(t,x,\xi) = \gamma(\xi)\psi(t,x,\xi) + \mathcal{G}\left\langle \nu, \psi(t,x,.)\right\rangle + f(t,x), \quad \psi(0,.) = 0, \tag{2.13}$$

$$\Phi(t,x) = \left\langle \nu, \psi(t,x,.)\right\rangle, \tag{2.14}$$

where $\psi := (\psi_1, \dots, \psi_M)^T$, $\gamma := \mathrm{diag}(\gamma_1, \dots, \gamma_M)$, $\nu := \mathrm{diag}(\nu_1, \dots, \nu_M)$, and $\langle \nu, \psi\rangle := (\langle \nu_1, \psi_1\rangle, \dots, \langle \nu_M, \psi_M\rangle)^T$.

Let us now consider the functional

$$\psi \longmapsto E_\psi = \frac{1}{2}\sum_i \iint \nu_i \left|\psi_i\right|^2 d\xi\, dx = \frac{1}{2}\iint \psi^T \nu\, \overline{\psi}\, d\xi\, dx\,;$$

thanks to the positivity of ν_i, the functional E_ψ is positive. We have the following.

PROPOSITION 2.5. *For any ψ solution of* (2.13)*, and at any t such that $f(t,\cdot) = 0$, the functional E_ψ verifies*

$$\frac{dE_\psi(t)}{dt} \leqslant 0.$$

Proof.

$$\begin{aligned}
\frac{dE_\psi(t)}{dt} &= \frac{1}{2}\left(\iint \left(\partial_t\psi\right)^T \nu\, \overline{\psi}\, d\xi dx + \iint \psi^T \nu\, \overline{\partial_t\psi}\, d\xi dx\right)\\
&= \iint \psi^T \nu \mathrm{Re}(\gamma)\, \overline{\psi}\, d\xi dx + \frac{1}{2}\left(\int \left\langle \nu, \psi\right\rangle^T\, \overline{\mathcal{G}\left\langle \nu, \psi\right\rangle}\, dx + \int \left(\mathcal{G}\left\langle \nu, \psi\right\rangle\right)^T \overline{\left\langle \nu, \psi\right\rangle}\, dx\right)\\
&= \iint \psi^T \nu \mathrm{Re}(\gamma)\, \overline{\psi}\, d\xi dx\\
&\quad + \frac{1}{2}\sum_{i,j}\left[\left(\mathcal{G}_{ij}\left\langle \nu_j, \psi_j\right\rangle \,\middle|\, \left\langle \nu_i, \psi_i\right\rangle\right)_{L^2(\Omega)} + \left(\left\langle \nu_j, \psi_j\right\rangle \,\middle|\, \mathcal{G}_{ji}\langle \nu_i, \psi_i\rangle\right)_{L^2(\Omega)}\right].
\end{aligned}$$

Because $\mathcal{G}$ is anti-selfadjoint, we then have

$$\frac{dE_\psi(t)}{dt} = \iint \psi^T \nu \mathrm{Re}(\gamma)\, \overline{\psi}\, d\xi\, dx = \sum_i \iint \nu_i \mathrm{Re}(\gamma_i)\, |\psi_i|^2\, d\xi \leqslant 0. \qquad \square$$

Therefore, the time-local problem (2.13) is dissipative in the sense of the positive functional E_ψ. At this stage, standard methods of semigroup theory can be investigated to study the well-posedness of this Cauchy problem in the associated energy Hilbert space[19], from which will follow the well-posedness of problem (2.10) as a simple consequence.

In practice, the numerical resolution of problems such as (2.10) presents major difficulties due to the nonlocal nature of $\mathbf{H}(\partial_t)^{-1}$. So, we focus here on the construction and analysis of numerical schemes for (2.13), from which approximate solutions of (2.10) will be directly deduced. We mainly study the stability property, which holds most of the technical difficulties.

3. Numerical schemes for (2.13).

First note that in any case, it follows from (2.14) that, in the sense of suitable topologies not specified here, approximations of a Φ solution of (2.10) will be straightforwardly obtained from discrete approximations $\tilde{\psi}$ of a ψ solution of (2.13) under the generic form,

$$\Phi(t_n, x_k) \simeq \tilde{\Phi}(t_n, x_k) = \sum_l \alpha_l\, \tilde{\psi}(t_n, x_k, \xi_l).$$

So, we build and study some numerical schemes for (2.10). A general technique for ξ-discretization presented in [14] is first introduced, followed by the statement of fundamental properties of generic x-discretizations, inherited from the properties of operator $\mathcal{G}$. Then, we consider different ways of using time discretization which define different classes of implicit and explicit schemes.

3.1. $\boldsymbol{\xi}$-discretization (see [14]).

Consider $\mathcal{K}$ a Hilbert space such that $\psi(t, x, .) \in \mathcal{K}$, and consider $\mathcal{K}_L$ a sequence of subspaces of $\mathcal{K}$ of dimension L such that $\overline{\cup_L \mathcal{K}_L}^{\mathcal{K}} = \mathcal{K}$. Given a mesh $\{\xi_l\}_{l=1:L}$, consistent approximations $\widetilde{\psi}_L \in \mathcal{K}_L$ of ψ are then defined by

$$\widetilde{\psi}_L(\xi) = \sum_{l=1}^{L} \psi(\xi_l) \Lambda_l(\xi),$$

where Λ_l are finite element functions belonging to $\mathcal{K}_L$ in such a way that

$$\left\| \widetilde{\psi}_L - \psi \right\|_{\mathcal{K}} \underset{L\to\infty}{\longrightarrow} 0.$$

We then deduce the finite-dimensional approximate state formulation of (2.13),

$$\partial_t \psi(t, x, \xi_l) = \gamma(\xi_l)\psi(t, x, \xi_l) + \mathcal{G} \sum_j C_j \psi(t, x, \xi_j), \quad l = 1 : L, \tag{3.1}$$

where

$$C_l = \mathrm{diag}(c_{l1}, \ldots, c_{lM}),\ c_{li} := \int \nu_i(\xi)\, \Lambda_l(\xi) d\xi.$$

Note that, in practice, only a few tens of ξ_l are necessary to correctly approximate each operator $H_i(\partial_t)^{-1}$. More details on the ξ-discretization of diffusive state realizations of convolution operators can be found in [14].

In addition, for consistency with positivity of measures ν_i, we will suppose that

$$c_{li} \geqslant 0;$$

this property, which will play a central role, is satisfied namely if $\Lambda_l \geqslant 0$. The energy functional associated with (3.1) is then

$$E_\psi^L(t) = \frac{1}{2}\sum_{i,l}\int c_{li}\,|\psi_i(t,x,\xi_l)|^2\,dx = \frac{1}{2}\sum_{l}\int \psi(t,x,\xi_l)^T C_l\,\overline{\psi(t,x,\xi_l)}\,dx$$

and verifies, in the same way as previously,

$$\begin{aligned} \frac{dE_\psi^L(t)}{dt} &= \sum_l \int \psi(t,x,\xi_l)^T \mathrm{Re}(\gamma(\xi_l))\,C_l\,\overline{\psi(t,x,\xi_l)}\,dx \\ &= \sum_{l,i}\int c_{li}\,\mathrm{Re}(\gamma_i(\xi_l))\;|\psi_i(t,x,\xi_l)|^2\,dx \leqslant 0. \end{aligned} \tag{3.2}$$

3.2. *x*-discretization. In formulation (3.1), $\mathcal{G}_{ij}$ is a differential operator; it is approximate on a mesh $\{x_k\}_{k=1:K} \subset \mathbb{R}^m$ by

$$\left(\mathcal{G}_{ij}\Phi\right)(x_q) \simeq \sum_{k=1}^{K} g_{ij}^{qk}\Phi(x_k) \quad \forall q = 1:K, \tag{3.3}$$

where the coefficients g_{ij}^{qk} define the approximation under consideration (for example, finite differences [17], finite elements, or even more general Galerkin methods up to suitable technical adaptations [4]). By denoting $\widetilde{\Phi} := (\Phi(x_1),\ldots,\Phi(x_K))^T$, (3.3) can be written in a more condensed way, as follows:

$$\left(\left(\mathcal{G}_{ij}\Phi\right)(x_1),\ldots,\left(\mathcal{G}_{ij}\Phi\right)(x_K)\right)^T \simeq G_{ij}\widetilde{\Phi},$$

where we denote by G_{ij} the matrix with terms g_{ij}^{qk}. In what follows, for simplicity $\widetilde{\Phi}$ will be denoted Φ.

Because the operator $\mathcal{G}$ is anti-selfadjoint, it is natural to consider approximations which preserve this property. So the block matrix G with block elements $G_{ij} \in \mathcal{M}_{K,K}(\mathbb{R})$ must be antisymmetric, that is,

$$G_{ij}^T = -G_{ji}. \tag{3.4}$$

In what follows, we will denote by $S_{G_{ij}}$ the quantity

$$S_{G_{ij}} := \max\left(\max_k \sum_q \left|g_{ij}^{qk}\right|, \max_q \sum_k \left|g_{ij}^{qk}\right|\right).$$

The Euclidian scalar product in $\mathbb{C}^K$ and the associated norm will be denoted

$$(X|\,Y) = \sum_{k=1}^{K} X_k\overline{Y_k}, \text{ and } \|X\| = \sqrt{\sum_{k=1}^{K}|X_k|^2}.$$

3.3. Stability analysis for an implicit scheme. We propose the following class of time-implicit schemes, based on a Cranck–Nicholson time discretization:

$$\frac{\psi_i^{n+1}(\xi_l)-\psi_i^n(\xi_l)}{\Delta t} = \gamma_i(\xi_l)\,\frac{\psi_i^{n+1}(\xi_l)+\psi_i^n(\xi_l)}{2} + \sum_{k,j} G_{ik}c_{jk}\,\frac{\psi_k^{n+1}(\xi_j)+\psi_k^n(\xi_j)}{2} + f_i^n, \tag{3.5}$$

where

$$\psi_i^n(\xi_l) = (\psi_i(n\Delta t, x_1, \xi_l), \dots, \psi_i(n\Delta t, x_K, \xi_l))^T$$

and $f_i^n = (f_i(n\Delta t, x_1), \dots, f_i(n\Delta t, x_K))^T$. In a more condensed way, (3.5) can be written

$$\frac{\psi^{n+1}(\xi_l)-\psi^n(\xi_l)}{\Delta t} = \Gamma_l\,\frac{\psi^{n+1}(\xi_l)+\psi^n(\xi_l)}{2} + G\sum_j Q_j\,\frac{\psi^{n+1}(\xi_j)+\psi^n(\xi_j)}{2} + f^n, \tag{3.6}$$

where $\psi^n(\xi_l) = (\psi_1^n(\xi_l)^T, \dots, \psi_M^n(\xi_l)^T)^T$, $f^n = (f_1^{nT}, \dots, f_M^{nT})^T$, $\Gamma_l = \mathrm{diag}(\gamma_i(\xi_l)\, I_K)$, $Q_j = \mathrm{diag}(c_{jk}\, I_K)$, and G is the antisymmetric block matrix defined above.

Let us now consider the quantity

$$E^n = \sum_l \Big(\psi^n(\xi_l)\,\Big|\, Q_l\psi^n(\xi_l)\Big) = \sum_{i,l} c_{li}\,|\psi_i^n(\xi_l)|^2 .$$

Note that, thanks to the positivity of coefficients c_{li}, E^n is an energy candidate for (3.6). We have the following.

THEOREM 3.1. *The implicit scheme* (3.6) *is stable.*

Proof.

$$\begin{aligned}
&E^{n+1} - E^n\\
&= \sum_l \big(Q_l(\psi^{n+1}(\xi_l) + \psi^n(\xi_l))\big|\, \psi^{n+1}(\xi_l) - \psi^n(\xi_l)\big) + \sum_l 2\mathrm{i}\,\mathrm{Im}\big(Q_l\psi^{n+1}(\xi_l)\big|\, \psi^n(\xi_l)\big)\\
&= \sum_l \frac{\Delta t}{2} \big(Q_l(\psi^{n+1}(\xi_l)+ \psi^n(\xi_l))\big|\, \Gamma_l(\psi^{n+1}(\xi_l)+ \psi^n(\xi_l))\big) + \sum_l 2\mathrm{i}\,\mathrm{Im}\big(Q_l\psi^{n+1}(\xi_l)\big|\, \psi^n(\xi_l)\big)\\
&\qquad + \frac{\Delta t}{2}\sum_{l,j} \big(Q_l(\psi^{n+1}(\xi_l) + \psi^n(\xi_l))\big|\, GQ_j(\psi^{n+1}(\xi_j) + \psi^n(\xi_j))\big) .
\end{aligned}$$

Because G is antisymmetric, we have

$$\sum_{l,j} \big(Q_l(\psi^{n+1}(\xi_l) + \psi^n(\xi_l))\big|\, GQ_j(\psi^{n+1}(\xi_j) + \psi^n(\xi_j))\big) = 0,$$

so

$$\begin{aligned}
&E^{n+1} - E^n\\
&= \sum_l \frac{\Delta t}{2} \big(Q_l(\psi^{n+1}(\xi_l)+ \psi^n(\xi_l))\big|\, \Gamma_l(\psi^{n+1}(\xi_l)+ \psi^n(\xi_l))\big) + \sum_l 2\mathrm{i}\,\mathrm{Im}\big(Q_l\psi^{n+1}(\xi_l)\big|\, \psi^n(\xi_l)\big)\\
&= \frac{\Delta t}{2}\sum_{i,l} \gamma_i(\xi_l)c_{li}\,\big|\psi_i^{n+1}(\xi_l) + \psi_i^n(\xi_l)\big|^2 + \sum_l 2\mathrm{i}\,\mathrm{Im}\big(Q_l\psi^{n+1}(\xi_l)\big|\, \psi^n(\xi_l)\big) .
\end{aligned}$$

As $E^{n+1} - E^n$ is real, we have

$$E^{n+1} - E^n = \frac{\Delta t}{2}\sum_{i,l} c_{li}\,\mathrm{Re}(\gamma_i(\xi_l))\ \big|\psi_i^{n+1}(\xi_l) + \psi_i^n(\xi_l)\big|^2 \leqslant 0. \qquad \square$$

3.4. Stability analysis for explicit schemes. In this section, we propose a class of two-step explicit schemes of the form

$$\psi_i^{n+1}(\xi_l) = a_{li}\,\psi_i^{n-1}(\xi_l) + b_{li}\sum_k G_{ik}\sum_j b_{jk}\,\psi_k^n(\xi_j) + b_{li}\,f_i^n, \tag{3.7}$$

where $a_{li} \in \mathbb{C}$, $|a_{li}| < 1$, and $b_{jk} \in \mathbb{R}_+^*$ depend both on time approximation and $\gamma_i(\xi_l)$ choices, and G is the antisymmetric block matrix associated with operator $\mathcal{G}$.

Let us now study the stability of (3.7). We consider the functional

$$E^n = \sum_{i,l} \|\psi_i^n(\xi_l)\|_2^2 + \mathrm{Re}\left(\psi_i^{n+1}(\xi_l)|\psi_i^{n-1}(\xi_l)\right).$$

LEMMA 3.2. *If*

$$\mathrm{Re}(a_{li}) - \frac{b_{li}}{2}\sum_{k,j} b_{jk}S_{G_{ik}} > 0 \quad \forall i,l, \tag{3.8}$$

then there exists $K > 0$ such that

$$E^n \geqslant K\sum_{i,l}\|\psi_i^n(\xi_l)\|^2.$$

Proof. We have

$$E^n = \sum_{i,l}\|\psi_i^n(\xi_l)\|_2^2 + \sum_{i,l}\mathrm{Re}(a_{li})\left\|\psi_i^{n-1}(\xi_l)\right\|_2^2 + \sum_{i,l,k,j} b_{li}b_{jk}\,\mathrm{Re}\left(G_{ik}\,\psi_k^n(\xi_j)|\,\psi_i^{n-1}(\xi_l)\right).$$

Moreover, by using the following relation:

$$\forall\alpha\in\mathbb{R},\ \forall u,v\in\mathbb{C}^K,\quad \alpha\,\mathrm{Re}\,(u|v) = \frac{|\alpha|}{2}(\|u\|^2 + \|v\|^2 - \|u - \mathrm{sign}(\alpha)v\|^2), \tag{3.9}$$

we get

$$\begin{aligned}\mathrm{Re}\left(G_{ik}\,\psi_k^n(\xi_j)|\,\psi_i^{n-1}(\xi_l)\right) \geqslant \sum_{p,q}\frac{|g_{ik}^{pq}|}{2}\left|\psi_k^n(\xi_j,x_q) + \mathrm{sign}(g_{ik}^{pq})\psi_i^{n-1}(\xi_l,x_p)\right|^2 \\ - \frac{S_{G_{ik}}}{2}\left\|\psi_i^{n-1}(\xi_l)\right\|^2 - \frac{S_{G_{ik}}}{2}\|\psi_k^n(\xi_j)\|^2;\end{aligned}$$

so, as $S_{G_{ik}} = S_{G_{ki}}$,

$$\begin{aligned}E^n \geqslant \sum_{i,l}\left(1 - \frac{b_{li}}{2}\sum_{k,j}b_{jk}S_{G_{ik}}\right)\|\psi_i^n(\xi_l)\|^2 + \sum_{i,l}\left(\mathrm{Re}(a_{li}) - \frac{b_{li}}{2}\sum_{k,j}b_{jk}S_{G_{ik}}\right)\left\|\psi_i^{n-1}(\xi_l)\right\|^2 \\ + \sum_{i,j,k,l,p,q} b_{li}b_{jk}\frac{|g_{ik}^{pq}|}{2}\left|\psi_k^n(\xi_j,x_q) + \mathrm{sign}(g_{ik}^{pq})\psi_i^{n-1}(\xi_l,x_p)\right|^2. \qquad \square\end{aligned}$$

Remark 2. Note that condition (3.8) necessarily implies $\mathrm{Re}(a_{li}) > 0$; the hypothesis $|a_{li}| < 1$ is motivated by the term $a_{li}\,\psi_i^{n-1}(\xi_l)$ of (3.7).

Remark 3. Conditions of Lemma 3.2 are necessary conditions that link Δt (in a_{li} and b_{ik}) and the space discretization step (in $S_{G_{ik}}$).

Let us now consider the quantity

$$\mathcal{E}^n = E^n + E^{n-1},$$

which, under the conditions of Lemma 3.2, defines an energy candidate for (3.7). Then, we have the following theorem for stability of the class of explicit schemes (3.7).

THEOREM 3.3. *Under the conditions of Lemma* 3.2, *and if for any* k, j,

(3.10)

$$|a_{jk}|^2+\frac{b_{jk}}{2}\sum_{i,l}\|G_{ik}\|^2\, b_{li}\left(\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|+b_{li}\sum_{p,q}b_{qp}\left(|a_{li}-\overline{a_{jk}}|+|a_{li}-\overline{a_{qp}}|\right)\right)\leqslant 1$$

and

$$\text{(3.11)} \qquad \operatorname{Re}(a_{jk})(|a_{jk}|^2-1)+\frac{b_{jk}}{2}\sum_{i,l}b_{li}\left||a_{jk}|^2+a_{jk}^2-\overline{a_{li}}a_{jk}-1\right|\leqslant 0,$$

then the scheme (3.7) *is stable.*

Proof. After computations and reorganization, we have

$$\begin{aligned}
&\mathcal{E}^{n+1}-\mathcal{E}^n = E^{n+1}-E^{n-1}\\
&=\sum_{i,l}|a_{li}|^2\left\|\psi_i^{n-1}(\xi_l)\right\|^2+\sum_{i,l,k,j}b_{li}b_{jk}\operatorname{Re}\left(a_{li}\psi_i^{n-1}(\xi_l)\right|G_{ik}\,\psi_k^n(\xi_j)\big)\\
&\quad+\sum_{i,l,k,j}b_{li}b_{jk}\operatorname{Re}\left(G_{ik}\,\psi_k^n(\xi_j)\left|\psi_i^{n+1}(\xi_l)\right)+\sum_{i,l}|a_{li}|^2\operatorname{Re}\left(\psi_i^n(\xi_l)\left|\psi_i^{n-2}(\xi_l)\right)\right.\right.\\
&\quad+\sum_{i,l,k,j}b_{li}b_{jk}\operatorname{Re}\left(a_{li}\psi_i^n(\xi_l)\left|G_{ik}\psi_k^{n-1}(\xi_j)\right)+\sum_{i,l,k,j}b_{li}b_{jk}\operatorname{Re}\left(G_{ik}\,\psi_k^{n+1}(\xi_j)\right|\psi_i^n(\xi_l)\right)\\
&\quad-\sum_{i,l}\left(\|\psi_i^{n-1}(\xi_l)\|^2+\operatorname{Re}\left(\psi_i^n(\xi_l)\left|\psi_i^{n-2}(\xi_l)\right)\right)\right..
\end{aligned}$$

As $G_{ik}=-G_{ki}^T$, we have

$$\begin{aligned}
\mathcal{E}^{n+1}-\mathcal{E}^n &= \sum_{i,l}(|a_{li}|^2-1)\left\|\psi_i^{n-1}(\xi_l)\right\|^2+\sum_{i,l}(|a_{li}|^2-1)\operatorname{Re}\left(\psi_i^n(\xi_l)\left|\psi_i^{n-2}(\xi_l)\right)\right.\\
&\quad+\sum_{i,l,k,j}b_{li}b_{jk}(a_{li}-a_{jk})\operatorname{Re}\left((a_{li}-\overline{a_{jk}})\psi_i^n(\xi_l)\right|G_{ik}\psi_k^{n-1}(\xi_j)\big)\\
&=\sum_{i,l}(|a_{li}|^2-1)\left\|\psi_i^{n-1}(\xi_l)\right\|^2+\sum_{i,l}(|a_{li}|^2-1)\operatorname{Re}(a_{li})\left\|\psi_i^{n-2}(\xi_l)\right\|^2\\
&\quad+\sum_{i,l,k,j}b_{li}b_{jk}\operatorname{Re}\left(\left(|a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right)\psi_i^{n-2}(\xi_l)\right|G_{ik}\,\psi_k^{n-1}(\xi_j)\Big)\\
&\quad+\sum_{i,l,k,j,p,q}b_{li}^2b_{jk}b_{qp}\operatorname{Re}\left((a_{li}-\overline{a_{jk}})G_{ip}\psi_p^{n-1}(\xi_q)\right|G_{ik}\psi_k^{n-1}(\xi_j)\big).
\end{aligned}$$

By using (3.9) and the following relation:

$$\operatorname{Re}(\beta u|v)=\frac{1}{2}\left(|\beta|\,\|v\|_2^2+|\beta|\,\|u\|_2^2-\left\|\sqrt{\beta}u+\overline{\sqrt{\beta}}v\right\|_2^2\right),$$

and after reorganization, we obtain

$$\begin{aligned}
&\mathcal{E}^{n+1}-\mathcal{E}^{n}=\sum_{i,l}(|a_{li}|^2-1)\left\|\psi_i^{n-1}(\xi_l)\right\|^2\\
&+\sum_{i,l}\left(\mathrm{Re}(a_{li})(|a_{li}|^2-1)+\frac{b_{li}}{2}\sum\nolimits_{k,j}b_{jk}\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|\right)\left\|\psi_i^{n-2}(\xi_l)\right\|^2\\
&-\frac{1}{2}\sum_{i,l,k,j,p,q}b_{li}^2b_{jk}b_{qp}\left\|\sqrt{a_{li}-\overline{a_{jk}}}G_{ip}\,\psi_p^{n-1}(\xi_q)+\overline{\sqrt{a_{li}-\overline{a_{jk}}}}G_{ik}\psi_k^{n-1}(\xi_j)\right\|^2\\
&-\frac{1}{2}\sum_{i,l,k,j}b_{li}b_{jk}\left\|\sqrt{|a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1}\,\psi_i^{n-2}(\xi_l)+\overline{\sqrt{|a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1}}\,G_{ik}\psi_k^{n-1}(\xi_j)\right\|^2\\
&+\frac{1}{2}\sum_{i,l,k,j}b_{li}b_{jk}\left(\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|+\sum_{p,q}b_{li}b_{qp}(|a_{li}-\overline{a_{jk}}|+|a_{li}-\overline{a_{qp}}|)\right)\left\|G_{ik}\psi_k^{n-1}(\xi_j)\right\|^2.
\end{aligned}$$

By using the property $\left\|G_{ik}\psi_k^{n-1}(\xi_j)\right\|\leqslant\|G_{ik}\|\left\|\psi_k^{n-1}(\xi_j)\right\|$, we then get

$$\begin{aligned}
\mathcal{E}^{n+1}-\mathcal{E}^{n}\leqslant\sum_{k,j}\Bigg(&(|a_{jk}|^2-1)+\frac{b_{jk}}{2}\sum_{i,l}\|G_{ik}\|^2\,b_{li}\Bigg[\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|\\
&+\ b_{li}\sum_{p,q}b_{qp}(|a_{li}-\overline{a_{jk}}|+|a_{li}-\overline{a_{qp}}|)\Bigg]\Bigg)\left\|\psi_k^{n-1}(\xi_j)\right\|^2\\
&+\sum_{i,l}\left(\mathrm{Re}(a_{li})(|a_{li}|^2-1)+\frac{b_{li}}{2}\sum_{k,j}b_{jk}\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|\right)\left\|\psi_i^{n-2}(\xi_l)\right\|^2\\
&-\frac{1}{2}\sum_{i,l,k,j}b_{li}b_{jk}\left\|\sqrt{|a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1}\,\psi_i^{n-2}(\xi_l)+\overline{\sqrt{|a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1}}\,G_{ik}\psi_k^{n-1}(\xi_j)\right\|^2\\
&-\frac{1}{2}\sum_{i,l,k,j,p,q}b_{li}^2b_{jk}b_{qp}\left\|\sqrt{a_{li}-\overline{a_{jk}}}G_{ip}\,\psi_p^{n-1}(\xi_q)+\overline{\sqrt{a_{li}-\overline{a_{jk}}}}G_{ik}\psi_k^{n-1}(\xi_j)\right\|^2.
\end{aligned}$$

So, if for any k,j,

$$|a_{jk}|^2+\frac{b_{jk}}{2}\sum_{i,l}\|G_{ik}\|^2\,b_{li}\left(\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|+b_{li}\sum_{p,q}b_{qp}\,(|a_{li}-\overline{a_{jk}}|+|a_{li}-\overline{a_{qp}}|)\right)\leqslant 1$$

and for any i,l,

$$\mathrm{Re}(a_{li})(|a_{li}|^2-1)+\frac{b_{li}}{2}\sum_{k,j}b_{jk}\left||a_{li}|^2+a_{li}^2-\overline{a_{jk}}a_{li}-1\right|\leqslant 0,$$

then $\mathcal{E}^{n+1}\leqslant\mathcal{E}^{n}$, from which we deduce $E^{n+1}\leqslant E^{n-1}$. Consequently, we have $E^n\leqslant\max(E^0,E^1)$; from Lemma 3.2, the scheme is stable. □

In section 4, where a concrete application is presented, we will consider as follows two particular explicit schemes of the form (3.7), based on two time discretizations (the first one is rather classical, and the second can be expected to be more precise):

• In the first scheme, the time derivative is approximated by centered finite differences; we then get

$$\psi_i^{n+1}(\xi_l)=(1+2\Delta t\,\gamma_i(\xi_l))\,\psi_i^{n-1}(\xi_l)+2\Delta t\sum_k G_{ik}\sum_j c_{jk}\,\psi_k^n(\xi_j)+2\Delta t\,f_i^n,$$

which, after the change of variable $\widetilde{\psi}_i^{n+1}(\xi_l) = \psi_i^{n+1}(\xi_l)\sqrt{\frac{c_{li}}{2\Delta t}}$, is rewritten under the form (3.7) with

$$a_{li} = 1 + 2\,\Delta t\,\gamma_i(\xi_l) \text{ and } b_{jk} = \sqrt{2\,\Delta t\, c_{jk}}. \tag{3.12}$$

• The second scheme is based on another time discretization, described in the appendix and can be considered in the case where $\gamma_i(\xi)$ is real (e.g., $\gamma_i(\xi) = -\xi$). It is written

$$\psi_i^{n+1}(\xi_l) = e^{\gamma_i(\xi_l)2\Delta t}\,\psi_i^{n-1}(\xi_l) + \frac{e^{\gamma_i(\xi_l)2\Delta t} - 1}{\gamma_i(\xi_l)}\left(\sum_k G_{ik}\sum_j c_{jk}\,\psi_k^n(\xi_j) + f_i^n\right); \tag{3.13}$$

after the change of variable $\widetilde{\psi}_i^{n+1}(\xi_l) = \psi_i^{n+1}(\xi_l)\sqrt{\frac{c_{li}\gamma_i(\xi_l)}{e^{\gamma_i(\xi_l)2\Delta_t}-1}}$, (3.13) is rewritten under the form (3.7) with

$$a_{li} = e^{\gamma_i(\xi_l)2\Delta_t} \text{ and } b_{jk} = \sqrt{c_{jk}\,\frac{e^{\gamma_k(\xi_j)2\Delta_t} - 1}{\gamma_k(\xi_j)}}. \tag{3.14}$$

The stability of those particular schemes is obtained as a corollary of the general stability theorem, Theorem 3.3.

COROLLARY 3.4. *Under the conditions of Lemma* 3.2, *and if* Δt *is small enough, the two schemes* (3.7), (3.12) *and* (3.7), (3.14) *are stable.*

Proof. For the first scheme, we have

$$a_{li} = 1 + 2\Delta t\,\gamma_i(\xi_l) \text{ and } b_{jk} = \sqrt{2\Delta t c_{jk}},$$

so that, by supposing Δt small enough, conditions (3.10) and (3.11) are, respectively, equivalent to

$$1 + 4\Delta t\ \mathrm{Re}(\gamma_k(\xi_j)) \leqslant 1 \quad\text{and}\quad 4\Delta t\ \mathrm{Re}(\gamma_k(\xi_j)) \leqslant 0,$$

which are both verified thanks to the property $\mathrm{Re}\,\gamma \subset \mathbb{R}^-$.

For the second scheme, we have

$$a_{li} = e^{\gamma_i(\xi_l)2\Delta_t} \text{ and } b_{jk} = \sqrt{c_{jk}\,\frac{e^{\gamma_k(\xi_j)2\Delta_t} - 1}{\gamma_k(\xi_j)}},$$

so if Δt is small enough, then

$$a_{li} \sim 1 + 2\Delta t\,\gamma_i(\xi_l) \text{ and } b_{jk} \sim \sqrt{2\Delta t c_{jk}},$$

and the same analysis as for the first scheme can be made. □

4. Application to a porous wall model.

4.1. Problem under consideration. In the context of aircraft motor noise reduction in the aerospace industry, a specific porous wall was proposed in [5] for absorption of a wide part of the energy of incident acoustic waves. The following frequency model of such a material has been established from analysis of harmonic propagating waves:

$$\text{(4.1)} \qquad \left\{ \begin{array}{l} e\,\mathrm{i}\omega\,\rho_{\text{eff}}\,(\mathrm{i}\omega)\;\hat{u} + \partial_x \hat{P} = 0 \\ e\,\mathrm{i}\omega\,\chi_{\text{eff}}\,(\mathrm{i}\omega)\;\hat{P} + \partial_x \hat{u} = 0 \end{array} \right. \quad \text{with} \quad \left\{ \begin{array}{l} \rho_{\text{eff}}\,(\mathrm{i}\omega) = \rho\,(1 + a\,\frac{\sqrt{1+b\,\mathrm{i}\omega}}{\mathrm{i}\omega}) \\ \chi_{\text{eff}}\,(\mathrm{i}\omega) = \chi\,(1 - \beta\,\frac{\mathrm{i}\omega}{\mathrm{i}\omega + a'\sqrt{1+b'\mathrm{i}\omega}}), \end{array} \right.$$

where $\widehat{u}$ and $\widehat{P}$ designate the Fourier transforms of the velocity and the pressure in the porous medium, e denotes the thickness of the porous wall,[2] $\rho_{\text{eff}}\,(\mathrm{i}\omega)$ and $\chi_{\text{eff}}\,(\mathrm{i}\omega)$ are, respectively, the effective density of Pride, Morgan, and Gangi [16] and the effective compressibility of Lafarge [8], and $\rho = \rho_0\,\alpha_\infty$, $\chi = \frac{1}{P_0}$, $a = \frac{8\mu}{\rho_0 \Lambda^2}$, $a' = \frac{8\mu}{\rho_0 \Lambda'^2}$, $b = \frac{1}{2a}$, $b' = \frac{1}{2a'}$, $0 < \beta = \frac{\gamma - 1}{\gamma} < 1$. The physical parameters ρ_0, P_0, μ, γ, α_∞, Λ, Λ' are, respectively, the density and pressure at rest, the dynamic viscosity, the specific heat ratio, the tortuosity, the high frequency characteristic length of the viscous incompressible problem, and the high frequency characteristic length of the thermal problem. Note that all these parameters are positive by nature.

The aim of this section is to perform temporal simulations of these equations, based on the schemes previously studied. In the time domain, (4.1) can be written (by replacing $p = i\omega$ by ∂_t)

$$\text{(4.2)} \qquad \begin{bmatrix} H_1(\partial_t) & 0 \\ 0 & H_2(\partial_t) \end{bmatrix} \begin{pmatrix} u \\ P \end{pmatrix} = \begin{bmatrix} 0 & -\partial_x \\ -\partial_x & 0 \end{bmatrix} \begin{pmatrix} u \\ P \end{pmatrix}$$

with

$$H_1(p) = e\,\rho\,(p + a\,\sqrt{1 + b\,p}) \text{ and } H_2(p) = e\,p\,\chi\,\left(1 - \beta\,\frac{p}{p + a'\sqrt{1 + b'p}}\right).$$

The analytic continuations of functions $H_1(p)^{-1}$ and $H_2(p)^{-1}$ are clearly decreasing at infinity and holomorphic in $\mathbb{C} \setminus \mathbb{R}^-$. So, from Theorem 2.3, the time-local formulation (2.13) of (4.2) with $\gamma_i(\xi) = -\,|\xi|$ is valid. It takes the form

$$\text{(4.3)} \qquad \left\{ \begin{array}{l} \partial_t \psi(t,x,\xi) = \begin{bmatrix} -\xi & 0 \\ 0 & -\xi \end{bmatrix} \psi(t,x,\xi) + \begin{bmatrix} 0 & -\partial_x \\ -\partial_x & 0 \end{bmatrix} \begin{pmatrix} \langle \nu_1, \psi_1(t,x,.) \rangle \\ \langle \nu_2, \psi_2(t,x,.) \rangle \end{pmatrix}, \\ u = \langle \nu_1, \psi_1(t,x,.) \rangle\,, \\ P = \langle \nu_2, \psi_2(t,x,.) \rangle\,. \end{array} \right.$$

After computations, the γ-symbol ν_i associated with the operator $H_i(\partial_t)^{-1}$ is expressed (δ denotes the Dirac measure)

$$\nu_1(\xi) = \frac{a}{\pi\,e\,\rho}\,\frac{\sqrt{b\,\xi - 1}}{\xi^2 + \frac{a\,\xi}{2} - a^2}\,\mathbf{1}_{\xi > 2a} + k_1\,\delta(\xi - \xi_1),$$

$$\nu_2(\xi) = \frac{a'\,\beta}{\pi\,e\,\chi}\,\frac{\sqrt{b'\,\xi - 1}}{\xi^2\,(1 - \beta)^2 + \frac{a'}{2}\xi - a'^2}\,\mathbf{1}_{\xi > 2a'} + \frac{1}{e\,\chi}\delta(\xi) + k_2\,\delta(\xi - \xi_2)$$

[2] In the model, the unit of length for x is e, so $x \in \,]0,1[$.

with

$$\xi_1 = \frac{a(\sqrt{17}-1)}{4} > 0, \quad \xi_2 = \frac{a'(\sqrt{1+16(1-\beta)^2}-1)}{4(1-\beta)^2} > 0,$$
$$k_1 = \frac{\sqrt{17}-1}{e\,\rho\sqrt{17}} > 0, \quad k_2 = \frac{\beta(\sqrt{1+16(1-\beta)^2}-1)}{e\,\chi(1-\beta)\sqrt{1+16(1-\beta)^2}} > 0.$$

For the ξ-discretization, we consider the classical interpolation functions

$$\Lambda_l(\xi) = \frac{\xi-\xi_{l-1}}{\xi_l-\xi_{l-1}}\mathbf{1}_{[\xi_{l-1},\xi_l]}(\xi) + \frac{\xi_{l+1}-\xi}{\xi_{l+1}-\xi_l}\mathbf{1}_{]\xi_l,\xi_{l+1}]}(\xi),$$

and coefficients c_{li} are computed by a simple quadrature of $\int \nu_i(\xi)\Lambda_l(\xi)d\xi$.

4.2. Numerical schemes. In this example, $\mathcal{G} = \left[\begin{smallmatrix} 0 & -\partial_x \\ -\partial_x & 0 \end{smallmatrix}\right]$. We use centered finite differences to approximate the derivative operator ∂_x, so the matrix of the x-discretization G is given by

$$G = \begin{bmatrix} 0 & G_{12} \\ G_{21} & 0 \end{bmatrix} \text{ with } G_{12} = G_{21} = \frac{1}{2\Delta x}\begin{bmatrix} 0 & -1 & & & \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ & & & 1 & 0 \end{bmatrix}.$$

Note that this matrix is antisymmetric, so that the schemes studied in section 3 can be used. Then, we consider the following:

- the implicit scheme

$$\begin{cases} \dfrac{\psi_1^{n+1}(\xi_l)-\psi_1^n(\xi_l)}{\Delta t} = -\xi_l\dfrac{\psi_1^{n+1}(\xi_l)+\psi_1^n(\xi_l)}{2} + G_{12}\sum_j c_{j2}\dfrac{\psi_2^{n+1}(\xi_j)+\psi_2^n(\xi_j)}{2} + f_1^n, \\ \dfrac{\psi_2^{n+1}(\xi_l)-\psi_2^n(\xi_l)}{\Delta t} = -\xi_l\dfrac{\psi_2^{n+1}(\xi_l)+\psi_2^n(\xi_l)}{2} + G_{21}\sum_j c_{j1}\dfrac{\psi_1^{n+1}(\xi_j)+\psi_1^n(\xi_j)}{2} + f_2^n; \end{cases} \tag{4.4}$$

- the two particular explicit schemes of the form

$$\begin{cases} \psi_1^{n+1}(\xi_l) = a_{l1}\,\psi_1^{n-1}(\xi_l) + b_{l1}\,G_{21}\sum_j b_{j2}\,\psi_2^n(\xi_j) + b_{l1}f_1^n, \\ \psi_2^{n+1}(\xi_l) = a_{l2}\,\psi_2^{n-1}(\xi_l) + b_{l2}\,G_{12}\sum_j b_{j1}\,\psi_1^n(\xi_j) + b_{l2}f_2^n, \end{cases} \tag{4.5}$$

respectively, obtained with

$$a_{li} = 1-2\Delta t\,\xi_l,\ b_{jk} = \sqrt{2\Delta t\,c_{jk}} \quad \text{and} \quad a_{li} = e^{-\xi_l 2\Delta_t},\ b_{jk} = \sqrt{c_{jk}\,\frac{e^{-\xi_j 2\Delta_t}-1}{-\xi_j}}.$$

4.3. Physical interpretation of stability conditions. Obviously, to correctly simulate wave propagation phenomena, explicit schemes must necessarily have a numerical influence velocity at least equal to the maximal velocity of wave fronts in the medium under consideration. When this is not the case, a consistent explicit scheme cannot be convergent and is therefore unstable. So, it can be expected that the stability conditions of section 3.4 applied to (4.5) can be in some way interpreted in terms of high frequency wave velocity. More precisely, is the sufficient stability condition for (4.5) "optimal" in the sense that it is close to the necessary condition mentioned above? This is studied in the present section.

Let us compute the expression of the high frequency wave velocity of model (4.2), denoted by c. We have

$$\left\{\begin{array}{l} u = -H_1(\partial_t)^{-1}\partial_x P, \\ P = -H_2(\partial_t)^{-1}\partial_x u, \end{array}\right. \tag{4.6}$$

so we get $u = H_1(\partial_t)^{-1}H_2(\partial_t)^{-1}\,\partial_x^2 u$. Moreover [14],

$$H_i(\mathrm{i}\omega)^{-1} = \int \frac{\nu_i(\xi)}{\mathrm{i}\omega+\xi}\,d\xi = \frac{1}{\mathrm{i}\omega}\int \frac{\nu_i(\xi)}{1+\frac{\xi}{\mathrm{i}\omega}}\,d\xi, \quad i=1,2.$$

So, when $\omega \to +\infty$,

$$H_i(\mathrm{i}\omega)^{-1} \sim \frac{1}{\mathrm{i}\omega}\int \nu_i(\xi)\,d\xi.$$

The equation $u = H_1(\partial_t)^{-1}H_2(\partial_t)^{-1}\,\partial_x^2 u$ therefore behaves at high frequency as equation $\partial_t^2 u = c^2\,\partial_x^2 u$, with

$$c = \sqrt{\int \nu_1(\xi)\,d\xi \int \nu_2(\xi)\,d\xi}.$$

Similarly, we denote by c_d the high frequency wave velocity of the continuous model obtained after ξ-discretization of (4.2), in which $H_i(\mathrm{i}\omega)$ is replaced by its approximation $\widetilde{H}_i(\mathrm{i}\omega)$ [14] as follows:

$$\widetilde{H}_i(\mathrm{i}\omega)^{-1} = \sum_j \frac{c_{ji}}{\mathrm{i}\omega+\xi_j} = \frac{1}{\mathrm{i}\omega}\sum_j \frac{c_{ji}}{1+\frac{\xi_j}{\mathrm{i}\omega}}, \quad i=1,2. \tag{4.7}$$

We have, when $|\omega| \to +\infty$,

$$\widetilde{H}_i(\mathrm{i}\omega)^{-1} \sim \frac{1}{\mathrm{i}\omega}\sum_j c_{ji},$$

which leads to a high frequency behavior of the form $\partial_t^2 u = c_d^2\,\partial_x^2 u$ with

$$c_d = \sqrt{\sum_j c_{j1}\sum_j c_{j2}}.$$

Thanks to the expression of c_{li}, we then have, if $\widetilde{H}_i^{-1}$ is sufficiently close to H_i^{-1},

$$c_d \simeq c. \tag{4.8}$$

Moreover, $S_{G_{12}} = S_{G_{21}} = \frac{1}{\Delta x}$, so the stability conditions of section 3.4 applied to (4.5) are Δt small enough and

$$\forall (i,k) \in \{(1,2),(2,1)\}, \quad \forall l = 1:L, \quad a_{li} - \frac{b_{li}}{2\Delta x}\sum_j b_{jk} > 0. \tag{4.9}$$

For the first explicit scheme, (4.9) is expressed as

$$\frac{\Delta x}{\Delta t} > v_d := \max_{(i,k)} \max_{l} \frac{\sqrt{c_{li}}}{1-2\Delta t\, \xi_l} \sum_j \sqrt{c_{jk}}, \tag{4.10}$$

where $\frac{\Delta x}{\Delta t}$ is the numerical influence velocity of the scheme. For the second explicit scheme, the order one approximation leads to the same condition. Then, we have the following result:

PROPOSITION 4.1. $v_d \geqslant c_d$.

Proof. Without loss of generality, we can consider that

$$v_d = \max_{(i,k)} \max_{l} \frac{\sqrt{c_{li}}}{1-2\Delta t\, \xi_l} \sum_j \sqrt{c_{jk}} = \max_{l} \frac{\sqrt{c_{l1}}}{1-2\Delta t\, \xi_l} \sum_j \sqrt{c_{j2}}.$$

So we have

$$\begin{aligned} c_d^2 = \sum\nolimits_j c_{j1} \sum\nolimits_j c_{j2} &\leqslant \max_l \sqrt{c_{l1}} \sum\nolimits_j \sqrt{c_{j1}}\, \max_l \sqrt{c_{l2}} \sum\nolimits_j \sqrt{c_{j2}} \\ &\leqslant \max_l \frac{\sqrt{c_{l1}}}{1-2\Delta t\, \xi_l} \sum\nolimits_j \sqrt{c_{j2}}\, \max_l \frac{\sqrt{c_{l2}}}{1-2\Delta t\, \xi_l} \sum\nolimits_j \sqrt{c_{j1}} \\ &\leqslant \left(\max_l \frac{\sqrt{c_{l1}}}{1-2\Delta t\, \xi_l} \right)^2 \left(\sum\nolimits_j \sqrt{c_{j2}} \right)^2 = v_d^2. \qquad \square \end{aligned}$$

As expected, we deduce from (4.10) and Proposition 4.1 that the numerical influence velocity of the scheme necessarily satisfies

$$\frac{\Delta x}{\Delta t} > c_d.$$

The sufficient stability condition (4.10) is of course not necessary. However, in the numerical results of section 4.4, the gap between this condition and the instability of the scheme is small; then, this condition is quasi optimal in this case.

Remark 4. The numerical velocity v_d of (4.5) could also be compared to that of a (theoretical) scheme, in which the variable ξ remains continuous, by considering the continuous equivalent of the quantity $v_d = \max_{(i,k)} \max_l \frac{\sqrt{c_{li}}}{1-2\Delta t\, \xi_l} \sum_j \sqrt{c_{jk}}$ in (4.10). Namely, by supposing by simplicity that ν_i are positive and continuous functions[3] with bounded support, that $\Delta_\xi = \xi_{l+1} - \xi_l$ is constant and with $\Lambda_l = \mathbf{1}_{[\xi_l, \xi_{l+1}]}$, there exists $\nu'_{li} \in [\nu_i(\xi_l), \nu_i(\xi_{l+1})]$ such that $c_{li} = \int_{\xi_l}^{\xi_{l+1}} \nu_i(\xi)\, d\xi = \nu'_{li}\, \Delta\xi$; so,

$$\sqrt{c_{li}} \sum_j \sqrt{c_{jk}} = \sqrt{\nu'_{li}} \sum_j \sqrt{\nu'_{jk}}\, \Delta\xi \simeq \sqrt{\nu'_{li}} \int \sqrt{\nu_k(\xi)}\, d\xi,$$

and therefore, with Δt such that $1 - 2\Delta t\, \xi \geqslant 0$ for any $\xi \in \operatorname{supp} \nu_1 \cup \operatorname{supp} \nu_2$,

$$v_d \simeq \max_{(i,k)} \max_{l} \frac{\sqrt{\nu'_{li}}}{1-2\Delta t\, \xi_l} \int \sqrt{\nu_k(\xi)}\, d\xi \leqslant v := \max_{(i,k)} \sup_{\xi} \frac{\sqrt{\nu_i(\xi)}}{1-2\Delta t\, \xi} \int \sqrt{\nu_k}\, d\xi.$$

Then, similarly to Proposition 4.1, it can be easily shown that $v > c$. Note, however, that besides the boundedness of supp ν_i, which is a quite unrealistic hypothesis, this estimation can be in some cases excessively pessimistic.

[3] Dirac and L^1_{loc} components could be similarly treated up to suitable technical adaptations.

4.4. Numerical results. We give in this section some numerical results obtained with the explicit schemes. The values of parameters are [5]

$$\Lambda = \Lambda' = 0.1\,10^{-3}\,\mathrm{m},\quad \rho_0 = 1.2\,\mathrm{kg.m}^{-3},\quad P_0 = 10^5\,\mathrm{Pa},$$
$$\mu = 1.8\,10^{-5}\,\mathrm{kg.m}^{-1}.\mathrm{s}^{-1},\quad \gamma = 1.4,\ \alpha_\infty = 1.3,\quad e = 5\,10^{-2}\,\mathrm{m}.$$

The frequency responses of the approximations of $H_i(\partial_t)^{-1}$ obtained with (4.7) are given in Figure 4.1. Only 15 (resp., 20) ξ_l are used to approximate $H_1(\partial_t)^{-1}$ (resp., $H_2(\partial_t)^{-1}$) in a range of six decades with good accuracy.

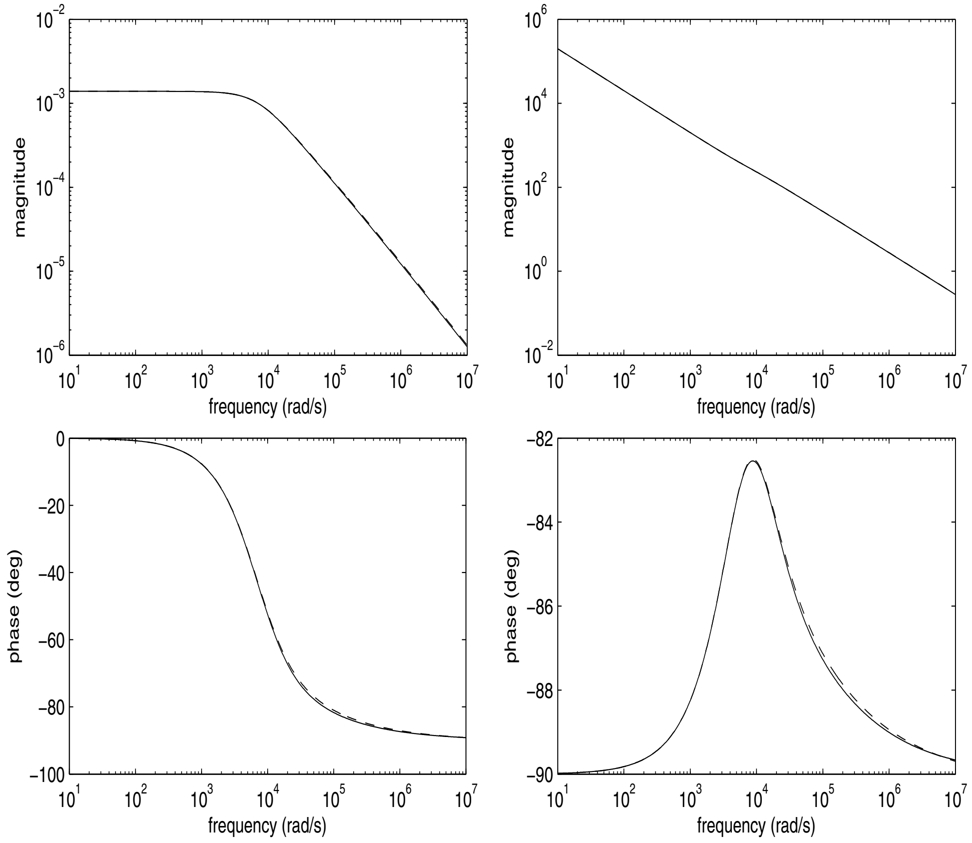


FIG. 4.1. *Exact (—) and approximate (- - -) frequency responses of operators $H_1(\partial_t)^{-1}$ (left) and $H_2(\partial_t)^{-1}$ (right).*

For illustration, the evolution of P obtained from simulation with explicit schemes is shown in Figure 4.2 (the two curves are superposed); the x-domain of (4.2) is $\Omega =]0,1[$ and the boundary conditions are

$$P(t,0) = (1-\cos(2\pi f\,t))\,\mathbf{1}_{[0,\frac{1}{f}]}(t),\qquad u(t,1) = 0,$$

with $f = 5\,kHz$. We can clearly observe the dissipation and dispersion due to operator $\mathbf{H}(\partial_t)$.

In Figure 4.3 we can see, at a particular time, the functions ψ_1 which are involved in the synthesis of u.

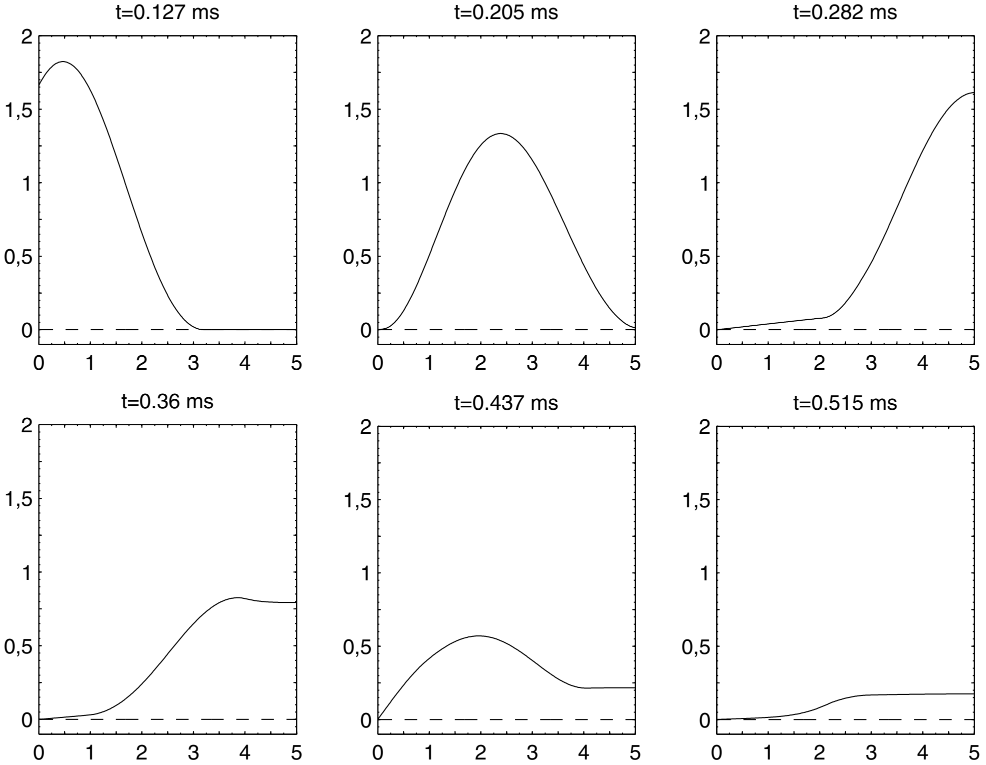


FIG. 4.2. *Evolution of* $\widetilde{P} = \sum_l b_{l2}\psi_2(\xi_l)$ *(N.B.: the unit of length for the x-axis is* $10^{-2}m$*).*

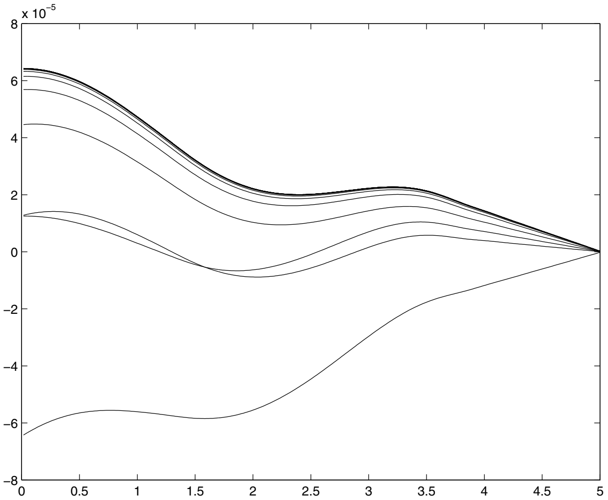


FIG. 4.3. *Functions* $\psi_1(t, ., \xi_l)$, $l = 1 : L$, *at time* $t = 1.3\,ms$.

4.5. Comparison between experimental and theoretical stability conditions. In the case of explicit schemes, the (sufficient) stability conditions are as follows:

Condition 1. Δt small enough.

Condition 2. $a_{li} - \frac{b_{li}}{2\Delta x} \sum_k b_{kj} > 0$.

We propose to compare from numerical simulations the stability condition, Condition 2, with the experimental one. In the same conditions as previously, for different values of $\frac{\Delta t}{\Delta x}$, we test the experimental stability of the schemes and verify whether Condition 2 is satisfied or not. The results are presented in Table 4.1 (resp., in Table 4.2) for the first (resp., the second) scheme. In the two cases, the results confirm that Condition 2 is a sufficient stability condition. We can notice that the interval of $\frac{\Delta t}{\Delta x}$ values, for which the scheme is stable even if Condition 2 is not verified, is small ($1.48\,10^{-4}$ to $1.98\,10^{-4}$); then, this condition is in fact "almost necessary."

Finally, to make the link with section 4.3, we can remark that the experimental stability bounds are intimately linked to propagation velocities. Indeed, the values of velocities defined in section 4.3 are (in length unit e per second)

$$c = 5992,\ c_d = 5038, \text{ and } v_d = 6856,$$

which correspond to the physical values $c = 299.6$ m.s^{-1}, $c_d = 251.9$ m.s^{-1}, and $v_d = 342.8$ m.s^{-1}. We can remark that, as expected, the schemes become unstable when $\frac{\Delta x}{\Delta t} \leqslant \frac{1}{1.98\,10^{-4}} \simeq c_d$, that is, when the numerical propagation velocity is less than the model's one.

TABLE 4.1
First explicit scheme.

Value of $\Delta t/\Delta x$	Condition 2	Stability
$\leqslant 1.4710^{-4}$	verified	yes
from 1.4810^{4} to 1.98710^{-4}	not verified	yes
$\geqslant 1.9810^{-4}$	not verified	no

TABLE 4.2
Second explicit scheme.

Value of $\Delta t/\Delta x$	Condition 2	Stability
$\leqslant 1.4710^{-4}$	verified	yes
from 1.4810^{4} to 1.9810^{-4}	not verified	yes
$\geqslant 1.9910^{-4}$	not verified	no

Appendix A. A particular time discretization. For a linear differential system in $\mathbb{C}^M$,

$$\partial_t \varphi = A\varphi + Bw, \quad \varphi(0) = 0,$$

the solution φ is given by

$$\varphi(t) = \int_0^t e^{A(t-s)} B\, w(s)\, ds.$$

For w constant in $[t - \Delta t, t + \Delta t]$, we have

$$\varphi(t+\Delta t) = \int_0^{t-\Delta t} e^{A(t+\Delta t-s)} B\, w(s)\, ds + \int_{t-\Delta t}^{t+\Delta t} e^{A(t+\Delta t-s)}\, ds B\, w(t) = F\varphi(t-\Delta t) + Gw(t),$$

with $F = e^{2\Delta t A}$ and $G = A^{-1}(e^{2\Delta t A} - I)B$. So we get the following numerical scheme:

$$\varphi^{t+\Delta t} = F\varphi^{t-\Delta t} + Gw^t.$$

Note that this scheme is especially useful in the case where A is diagonal.

REFERENCES

[1] J. Audounet, F. A. Devy-Vareta, and G. Montseny, *Pseudo-invariant diffusive control*, in Proceedings of the 14th Annual International Symposium of Mathematical Theory of Networks and Systems (MTNS 2000), Perpignan, France, 2000, (CD-Rom).

[2] J. Audounet, V. Giovangigli, and J.-M. Roquejoffre, *A threshold phenomenon in the propagation of a point source initiated flame*, Phys. D, 121 (1998), pp. 295–316.

[3] P. Bidan, T. Lebey, G. Montseny, and J. Saint-Michel, *Transient voltage distribution in inverter fed motor windings: Experimental study and modeling*, IEEE Trans. Power Electronics, 16 (2001), pp. 92–100.

[4] B. Cockburn, G. Karniadakis, and C. Shu, eds., *Discontinuous Galerkin Methods. Theory, Computation, and Applications*, Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000.

[5] S. Gasser, *Etude des propriétés acoustiques et mécaniques d'un matériau métallique poreux modèle à base de sphères creuses de nickel*, Ph.D. thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2003.

[6] J.-P. Groby and C. Tsogka, *A time domain method for modeling wave propagation phenomena in viscoacoustic media*, in Proceedings of the Sixth International Conference on Mathematical and Numerical Aspects of Wave Propagation (WAVES 2003), Jyväskylä, Finland, Springer, Berlin, 2003, pp. 911–915.

[7] B. I. Henry and S. L. Wearne, *Existence of Turing instabilities in a two-species fractional reaction-diffusion system*, SIAM J. Appl. Math., 62 (2002), pp. 870–887.

[8] D. Lafarge, *Propagation du son dans les matériaux poreux à structure rigide saturés par un fluide viscothermique*, Ph.D. thesis, Université du Maine, Le Mans, France, 1993.

[9] T. A. M. Langlands and B. I. Henry, *The accuracy and stability of an implicit solution method for the fractional diffusion equation*, J. Comput. Phys., 205 (2005), pp. 719–736.

[10] L. Laudebat, P. Bidan, and G. Montseny, *Modeling and optimal identification of pseudodifferential electrical dynamics by means of diffusive representation* I. *Modeling*, IEEE Trans. on Circuits Syst. I Regul. Pap., 51 (2004), pp. 1801–1813.

[11] M. Lenczner and G. Montseny, *Diffusive realization of operator solutions of certain operational partial differential equations*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 737–740.

[12] D. Levadoux and B. Michielsen, *Analysis of a boundary integral equation for high-frequency Helmholtz problems*, in Proceedings of the Fourth International Conference on Mathematical and Numerical Aspects of Wave Propagation (Golden, Colorado), SIAM, Philadelphia, 1998, pp. 565–567.

[13] A. Lorenzi and F. Messina, *Identification problems for Maxwell integro-differential equations related to media with cylindric symmetries*, J. Inverse Ill-Posed Prob., 11 (2003), pp. 411–437.

[14] G. Montseny, *Représentation diffusive*, Hermes Science, Paris, 2005.

[15] G. Montseny, *Diffusive representation for operators involving delays*, in Applications of Time Delay Systems, J.-J. Loiseau and J. Chiasson, eds., Springer-Verlag, Berlin, 2007, pp. 217–232.

[16] S. R. Pride, F. D. Morgan, and A. F. Gangi, *Drag forces of porous-medium acoustics*, Phys. Rev. B, 47 (1993), pp. 4964–4978.

[17] A. Taflove and S. C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 2nd ed., Artech House, Boston, MA, 2000.

[18] C. M. Topaz and A. L. Bertozzi, *Swarming patterns in a two-dimensional kinematic model for biological groups*, SIAM J. Appl. Math., 65 (2004), pp. 152–174.

[19] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1965.

SIAM J. Appl. Math.
Vol. 68, No. 6, pp. 1783–1806

AN INTERACTION THEORY FOR SCATTERING BY DEFECTS IN ARRAYS[*]

I. THOMPSON[†] AND C. M. LINTON[†]

Abstract. Wave scattering by an array of bodies that is periodic except for a finite number of missing or irregular elements is considered. The field is decomposed into contributions from a set of canonical problems, which are solved using a modified array scanning method. The resulting interaction theory for defects is very efficient and can be used to construct the field in a large number of different situations. Numerical results are presented for several cases, and particular attention is paid to the amplitude with which surface waves are excited along the array. We also show how other approaches can be incorporated into the theory so as to increase the range of problems that can be solved.

Key words. arrays, defects, scattering, surface waves

AMS subject classifications. 78A45, 78A50, 42A99

DOI. 10.1137/070703144

1. Introduction.

Wave scattering by arrays of bodies is of fundamental importance in numerous engineering and physics applications. Here we are concerned with the effect of one or more defects in an infinite, periodic array. This problem is of significant current interest in several fields, including elastodynamics [19] and phononic [7, 28] and photonic [1, 26, 5] crystals. The presence of defects leads to a significant increase in difficulty in determining the scattered field, because the geometry is no longer periodic. In particular, Rayleigh–Bloch (RB) surface waves (also known as array guided surface waves) are excited if the array geometry and physical parameters are such that these modes can exist. RB waves propagate without loss along an array, and are evanescent in other directions. They are known to occur in a wide variety of situations [15, 21, 18, 10]. One of the key goals of this article is to develop an efficient and accurate method for the determination of the amplitude with which they are excited. The theory is presented in a form that can be directly interpreted in a number of different physical contexts. These include the acoustic case, in which the wavenumber k is the ratio of the angular frequency ω to the speed of sound c, and the interaction of linear water waves with bottom mounted, surface penetrating cylinders, in which case k is the positive solution to the dispersion relation $k \tanh kh = \omega^2/g$, g being the acceleration due to gravity and h the quiescent fluid depth. For acoustics, Dirichlet and Neumann boundary conditions are used to model sound hard and sound soft bodies, respectively, whereas Neumann conditions are appropriate for solid bodies immersed in water. The method is also applicable in the electromagnetic and elastodynamic cases, provided that the overall vector wave problem decouples into separate scalar components.

Our first step in obtaining the field scattered by a defective array is to decompose the solution into contributions arising from a set of simpler, canonical problems. This is achieved by modifying the field generated when a wave interacts with a periodic

[*] Received by the editors September 19, 2007; accepted for publication (in revised form) March 24, 2008; published electronically July 16, 2008.
http://www.siam.org/journals/siap/68-6/70314.html

[†] Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, UK (i.thompson@lboro.ac.uk, c.m.linton@lboro.ac.uk). The first author's research was supported by EPSRC grant EP/C510941/1.

array, so as to eliminate a finite number of elements, or replace these with bodies of different sizes, shapes, or surface compositions. The procedure is independent of the type of wavefunctions used to represent the field (i.e., cylindrical, spherical, etc.) and is therefore presented in a general form in section 2. The canonical problems are independent of the defect configuration and all aspects of the incident field, except the wavenumber; they need not be solved again if these parameters are changed. In order for the decomposition to be useful in a specific case, the relevant canonical problems must be solved accurately and efficiently. The boundary conditions on the surface of the array elements come into play at this stage, and therefore we must apply an appropriate multiple scattering theory. This requires the use of certain results concerning the periodic array, and these are readily available for problems involving cylindrical wavefunctions; a summary is given in section 3. The canonical problems for this case are then solved in section 4 using a special Fourier series. This approach is closely related to the array scanning method [27, 16], which is typically used in problems involving excitation by an aperiodic field, and in particular for the analysis of antenna arrays [3, 4]. The idea is to create a periodic incident field by introducing an array of phase-shifted sources, and then to integrate over a single period of the phase shift so as to eliminate all but one of the sources. The procedure used in section 4 is similar, but its effect is rather different, and we shall refer to it as the "modified array scanning method" (MASM). Instead of eliminating sources, the integration, which must be performed using quadrature, enables us to replace one member of a periodic array with a source. This is the most computationally intensive part of the technique. Nevertheless, important parameters such as RB wave amplitudes can be efficiently calculated to near machine precision. In contrast, other techniques such as the filtering approach used for a related problem in [11] have limited accuracy. Technical details regarding the method used to evaluate the relevant integrals are given in the appendix. This method is chosen for simplicity and is open to improvement.

Considered together, the decomposition into canonical problems and the MASM are similar to the "fictitious source superposition method" which was originally used for a study of photonic crystals with a single defect [26]. This was later extended in [5] to account for situations where more than one defect is present. Our formulation, which is a generalization of earlier work in [22], is rather different and automatically includes the case of multiple defects. Indeed, by first reducing to canonical problems, we obtain an "interaction theory for defects" by means of which the solutions for a wide variety of cases can be constructed at very little computational expense.

A representative sample of the numerical results that can be obtained is given in section 5. We also demonstrate how the methods in sections 2–4 can be combined with other approaches, such as infinite array subtraction [11] and the large array approximation method used in [24], to widen the class of problems that can be considered.

2. General theory.

In this section we will show how the problem of wave scattering by a defective array can be reduced to a set of simpler, canonical problems. This is achieved using a procedure that is independent of the shape of the scatterers and the boundary conditions that are to be applied on their surfaces. We therefore present the theory from a general perspective, although for clarity we deal with the case of a one-dimensional array in the two-dimensional setting. The extensions to higher array dimensions and to three dimensions in space is straightforward, requiring only that scalar indices are replaced by appropriate multi-indices.

Thus, consider an array of scatterers which is periodic, except for a finite number of missing, or possibly irregular, elements. The elements are labeled by an index $p \in \mathbb{Z}$, and the defects correspond to those values for which p is a member of the finite defect set $\mathcal{D}$. If $p \notin \mathcal{D}$, we shall say that scatterer p is regular. A (one-dimensional) lattice of points $\mathbf{r}_p$ is defined so that $\mathbf{r} = \mathbf{r}_p$ lies inside scatterer p if this body is present in the array. The field in the vicinity of each scatterer is then expanded about the point $\mathbf{r} = \mathbf{r}_p$ as a sum of incoming and outgoing wavefunctions. The former are regular for all $\mathbf{r}$, whereas the latter are singular at $\mathbf{r} = \mathbf{r}_p$ and regular elsewhere. The choice of $\mathbf{r}_p$ is of course not unique.

In the region exterior to the scatterers, all wavefields ϕ must satisfy the Helmholtz equation

$$(\nabla^2 + k^2)\phi = 0. \tag{2.1}$$

The array is excited by the incident wave ϕ^{i}, and the total field is obtained by adding the scattered response. Hence,

$$\phi^{\mathrm{t}} = \phi^{\mathrm{i}} + \phi^{\mathrm{s}}, \tag{2.2}$$

where ϕ^{s} can be expanded in the form

$$\phi^{\mathrm{s}}(\mathbf{r}; \mathcal{D}) = \sum_m \sum_p A_m^p(\mathcal{D}) \mathcal{H}_m^p(\mathbf{r}). \tag{2.3}$$

Here, the notation $\mathcal{H}_m^p$ represents an outgoing wavefunction of order m that is singular at $\mathbf{r} = \mathbf{r}_p$ and regular elsewhere. Where no limits are placed on an index it is to be understood that this ranges over all possible values. The radiation condition stipulates that ϕ^{s} cannot include any contributions that are incoming from the far field, or that increase in magnitude as the observer moves toward infinity. Initially, we consider defects that consist of missing scatterers, in which case we must have

$$A_m^p(\mathcal{D}) = 0, \quad p \in \mathcal{D}, \tag{2.4}$$

so that there are no singularities in the field. Later we will show how the theory can be modified to account for irregular scatterers, which is slightly more difficult.

The pivotal idea behind our procedure is to modify $\phi^{\mathrm{s}}(\mathbf{r}, \emptyset)$ (i.e., the scattered field that occurs when there is no defect) by cancelling the singularities at $\mathbf{r} = \mathbf{r}_p$ for each $p \in \mathcal{D}$. The resulting wavefield does not include any radiation from the scatterers for which $p \in \mathcal{D}$, and no longer satisfies the boundary conditions on their surface. In this way, the influence of these array elements is eliminated. The boundary conditions on the surface of the regular scatterers are still satisfied, as is the radiation condition.

At a later stage, it is necessary to apply a multiple scattering theory in order to satisfy the boundary conditions on the scatterer surfaces. This requires that, in some region containing the surface of scatterer p, the total field can be represented in the form

$$\phi^{\mathrm{t}}(\mathbf{r}; \mathcal{D}) = \phi_p^{\mathrm{i}}(\mathbf{r}; \mathcal{D}) + \phi_p^{\mathrm{r}}(\mathbf{r}; \mathcal{D}), \tag{2.5}$$

where

$$\phi_p^{\mathrm{i}}(\mathbf{r}; \mathcal{D}) = \sum_m I_m^p(\mathcal{D}) \mathcal{J}_m^p(\mathbf{r}) \tag{2.6}$$

and

$$\phi_p^{\mathrm{r}}(\mathbf{r};\mathcal{D}) = \sum_m A_m^p(\mathcal{D})\mathcal{H}_m^p(\mathbf{r}). \tag{2.7}$$

Here, $\mathcal{J}_m^p$ represents a regular wavefunction of order m and ϕ_p^{i} is the total field incoming toward the point $\mathbf{r}_p$. It consists of the incident wave and the radiation from all of the other scatterers. The second term on the right-hand side of (2.5) represents the field outgoing from scatterer p. The relationships between the expansions (2.3) and (2.5)–(2.7) can be found in [14, Chapters 2 and 3] for wavefunctions in a number of separable geometries. The crucial point here is the nature of the regions where the series appearing in (2.6) and (2.7) converge and therefore represent valid solutions to the Helmholtz equation. The expansion of the incoming field (2.6) is valid inside a simply connected region that contains the point $\mathbf{r}_p$. In fact, if we are to apply a multiple scattering theory based on the expansions (2.5)–(2.7), this region must contain the whole of scatterer p. Thus, the field incoming toward a particular body can be extended to the entire region inside that body, and there it continues to represent a valid solution to the Helmholtz equation. The same cannot be said for the field radiating from a particular body (equation (2.7)) because $\mathcal{H}_m^p(\mathbf{r})$ is singular at the point $\mathbf{r} = \mathbf{r}_p$. Note that the use of (2.5)–(2.7) to represent the field at the surface of the scatterers imposes a geometrical restriction. For cylindrical and spherical wavefunctions, the maximum distance from $\mathbf{r}_p$ to the surface of scatterer p must be less than $|\mathbf{r}_p - \mathbf{r}_{p\pm 1}|$ [14, sections 2.5, 3.12].

As a starting point, for the case where $\mathcal{D} = \emptyset$, we have

$$\phi^{\mathrm{s}}(\mathbf{r};\emptyset) = \sum_m \sum_p A_m^p(\emptyset)\mathcal{H}_m^p(\mathbf{r}), \tag{2.8}$$

and we will assume that the coefficients $A_m^p(\emptyset)$ are known, since this is a periodic geometry, and so the solution can be obtained relatively easily. Now, construct the field $\phi^{\mathrm{s}}(\mathbf{r};\mathcal{D})$ by writing

$$\phi^{\mathrm{s}}(\mathbf{r};\mathcal{D}) = \phi^{\mathrm{s}}(\mathbf{r};\emptyset) + \psi(\mathbf{r};\mathcal{D}), \tag{2.9}$$

and observe that $\psi(\mathbf{r};\mathcal{D})$ must satisfy the boundary conditions on the regular scatterers because $\phi^{\mathrm{s}}(\mathbf{r};\mathcal{D})$ and $\phi^{\mathrm{s}}(\mathbf{r};\emptyset)$ do so independently. From (2.3), (2.4), and (2.8) we have the explicit representation

$$\psi(\mathbf{r};\mathcal{D}) = \sum_m \sum_{p\notin\mathcal{D}} \left[A_m^p(\mathcal{D}) - A_m^p(\emptyset)\right]\mathcal{H}_m^p(\mathbf{r}) - \sum_m \sum_{p\in\mathcal{D}} A_m^p(\emptyset)\mathcal{H}_m^p(\mathbf{r}). \tag{2.10}$$

By considering the last term on the right-hand side (which is known) as an incident field, and the other terms as the associated scattered response, it is now seen that $\psi(\mathbf{r};\mathcal{D})$ is the total field that occurs when an array with scatterers absent for $p \in \mathcal{D}$ is excited by a distribution of sources located at the points $\mathbf{r} = \mathbf{r}_p$, $p \in \mathcal{D}$. We shall refer to $\mathcal{H}_m^p(\mathbf{r})$ as the source of order m with unit amplitude, located at the point $\mathbf{r} = \mathbf{r}_p$.

Rather than solve for $\psi(\mathbf{r};\mathcal{D})$ directly, we can reduce the problem to a set of simpler, canonical problems by considering each source term in (2.10) separately. Thus, introduce the potential $\psi_n^q(\mathbf{r})$, which represents the total field that occurs when a periodic array has a single element (labeled by q) removed and replaced by a unit source of order n. Crucially, if $q \in \mathcal{D}$, then $\psi_n^q(\mathbf{r})$ satisfies the boundary conditions on the surface of all the regular scatterers. Now $\psi(\mathbf{r}; \mathcal{D})$ clearly consists entirely of waves that are outgoing from the array, and therefore we can expand it into the form

$$\psi_n^q(\mathbf{r}) = \mathcal{H}_n^q(\mathbf{r}) + \sum_m \sum_{p \neq q} C_{m,n}^{p,q} \, \mathcal{H}_m^p(\mathbf{r}). \tag{2.11}$$

Here, we have introduced the convention that the indices to the right of the comma describe the source, in this case referring to order n and position q. It is convenient to simplify such expressions by defining

$$C_{m,n}^{q,q} = \delta_{mn}, \tag{2.12}$$

so that the first term on the right-hand side can be taken inside the series. To avoid any possible misinterpretation, we emphasize that (2.11) does not represent a homogeneous solution to the periodic (i.e., defect-free) array problem because the appropriate boundary condition on the surface of scatterer q is not satisfied.

Next, we represent $\psi(\mathbf{r}; \mathcal{D})$ as a linear combination of the potentials $\psi_n^q(\mathbf{r})$, $q \in \mathcal{D}$; thus

$$\psi(\mathbf{r}; \mathcal{D}) = \sum_n \sum_{q \in \mathcal{D}} a_n^q \psi_n^q(\mathbf{r}). \tag{2.13}$$

If we substitute from (2.11) into (2.13) and rearrange the summations, we obtain

$$\psi(\mathbf{r}; \mathcal{D}) = \sum_m \sum_{p \in \mathcal{D}} a_m^p \mathcal{H}_m^p(\mathbf{r}) + \sum_m \sum_p \sum_n \sum_{\substack{q \in \mathcal{D} \\ q \neq p}} a_n^q C_{m,n}^{p,q} \, \mathcal{H}_m^p(\mathbf{r}). \tag{2.14}$$

Comparing this with (2.10), we find that

$$a_m^p + \sum_n \sum_{\substack{q \in \mathcal{D} \\ q \neq p}} a_n^q C_{m,n}^{p,q} = -A_m^p(\emptyset), \quad p \in \mathcal{D}, \tag{2.15}$$

which is a linear system of equations for the coefficients a_m^p, and

$$\sum_n \sum_{q \in \mathcal{D}} a_n^q C_{m,n}^{p,q} = A_m^p(\mathcal{D}) - A_m^p(\emptyset), \quad p \notin \mathcal{D}, \tag{2.16}$$

which then serves to determine the unknowns $A_m^p(\mathcal{D})$. Equation (2.15) is an "interaction theory for defects," which is similar in nature to the standard interaction theories for multiple bodies. If only a single scatterer is absent from the array, we retrieve $a_m^p = -A_m^p(\emptyset)$ so as to cancel the radiation emanating from $\mathbf{r} = \mathbf{r}_p$, as we should expect. A useful simplification now occurs if the array consists of periodic repetitions of a single body. In this case the potentials ψ_m^p are identical up to a spatial shift, and we need only determine ψ_m^0. In terms of the coefficients $C_{m,n}^{p,q}$, we have

$$C_{m,n}^{p,q} = C_{m,n}^{p-q,0} \tag{2.17}$$

and so there is a single canonical problem to solve for each value of m.

Finally, consider defects consisting of scatterers that are in some way different from the other elements of the array. In this case, the method operates by replacing members of the periodic array for which $p \in \mathcal{D}$ with irregular bodies. In contrast to the case of absent scatterers, $A_m^p(\mathcal{D})$ is generally nonzero for $p \in \mathcal{D}$. The singularities at $\mathbf{r} = \mathbf{r}_p$ are no longer cancelled; instead they are adjusted so that for $p \in \mathcal{D}$, the expansion (2.7) represents a solution to the Helmholtz equation in the region exterior to the new element. Consequently, the point $\mathbf{r} = \mathbf{r}_p$ must lie inside scatterer p for $p \in \mathcal{D}$ (as it does for $p \notin \mathcal{D}$). The field $\psi(\mathbf{r}; \mathcal{D})$ can still be constructed from a linear combination of the solutions to the same canonical problems, but in place of (2.15), we now have

$$a_m^p + \sum_n \sum_{\substack{q \in \mathcal{D} \\ q \neq p}} a_n^q C_{m,n}^{p,q} = A_m^p(\mathcal{D}) - A_m^p(\emptyset), \quad p \in \mathcal{D}. \tag{2.18}$$

Equation (2.16) is unaffected. The presence of the additional unknowns $A_m^p(\mathcal{D})$ on the right-hand side of (2.18) is countered by the need to apply a boundary condition on the surface of the irregular scatterers, and in section 5 we shall see how this works in practice. While it is evident that replacing scatterers is more complicated than eliminating them, the increase in difficulty is marginal. Essentially this is because the extra requirement is to determine the field incoming toward $\mathbf{r} = \mathbf{r}_p$ for $p \in \mathcal{D}$, but this is no more difficult than determining the field incoming toward a regular scatterer, which is always necessary.

A major advantage of this method over a more direct approach is as follows. Had we simply applied an interaction theory to the defective array problem, we would be faced with the inversion of a linear system of equations containing infinite sums over the spatial indices. These have a very slow rate of convergence and present serious difficulties in obtaining accurate results, even with the aid of modern computing power. In contrast, (2.15), (2.16), and (2.18) contain only *finite* spatial sums. The infinite order summation is of less concern, particularly at low frequencies, because as $|m|$ is increased the coefficients A_m^p converge rapidly to zero. Even in cases where the scatterers are almost in contact, the convergence of the order sum is much more rapid than that of the spatial sum; the former can be truncated at a relatively small value of $|m|$. Of course, it remains to solve the canonical problems, and these involve infinite linear systems containing spatial sums. However, these possess symmetries that are not present in the overall problem, and as mentioned earlier, solutions to one set of canonical problems can be used to construct the field for a number of different cases. Thus, the decomposition described above is useful even in problems where the MASM cannot be used effectively.

3. Array problems involving cylindrical wavefunctions. In order to solve the canonical problems that arise in the interaction theory for defects, we must deal with the boundary conditions on the scatterer surfaces. It is therefore necessary to present subsequent material for a specific geometry, and since the theory of linear arrays is well established for the case of cylindrical wavefunctions, this is a natural choice. Here, we collect some results from pre-existing literature in this area that will be needed later. It should be noted that the essential principles upon which the method depends remain unchanged if wavefunctions from another separable geometry are used. We will assume that the scatterers themselves are circular so as to present the theory in the simplest possible form; however we will indicate how scatterers of a different shape can be considered through the incorporation of transfer matrices.

Let all lengths be scaled on the distance between the centers of consecutive lattice points, with these located at $\mathbf{r}_p = (p, 0)$ in the (x, y) plane. According to the chosen scaling, the radius a of the regular scatterers must satisfy the inequality $a \leq 0.5$. The expansion (2.3) now takes the form

$$\phi^{\mathrm{s}}(\mathbf{r}; \mathcal{D}) = \sum_m \sum_p A_m^p(\mathcal{D})\, H_m^{(1)}(kr_p)\mathrm{e}^{\mathrm{i}m\theta_p}, \tag{3.1}$$

where (r_p, θ_p) is a set of polar coordinates with its origin at the center of scatterer p (see Figure 3.1), and $H_m^{(1)}(\cdot)$ denotes a Hankel function of the first kind with order m. This choice of outgoing wavefunction (rather than $H_m^{(2)}(\cdot)$) corresponds to an implicit time-harmonic factor $\mathrm{e}^{-\mathrm{i}\omega t}$. We also have a decomposition of the form (2.5)–(2.7), with

$$\phi_p^{\mathrm{i}}(r_p, \theta_p; \mathcal{D}) = \sum_m I_m^p(\mathcal{D})\, J_m(kr_p)\mathrm{e}^{\mathrm{i}m\theta_p} \tag{3.2}$$

and

$$\phi_p^{\mathrm{r}}(r_p, \theta_p; \mathcal{D}) = \sum_m A_m^p(\mathcal{D})\, H_m^{(1)}(kr_p)\mathrm{e}^{\mathrm{i}m\theta_p}, \tag{3.3}$$

where $J_m(\cdot)$ is the Bessel function of order m. As before, $\phi_p^{\mathrm{i}}(\mathbf{r}; \mathcal{D})$ represents the total field incoming toward scatterer p, and this consists of the incident wave and the radiation from all of the other scatterers. The expansion (3.2) is a valid representation for $\phi_p^{\mathrm{i}}(\mathbf{r}; \mathcal{D})$, provided that $r_p < 1$. In general, a transfer matrix appropriate to the geometry of the scatterers relates the coefficients $I_m^p(\mathcal{D})$ and $A_m^p(\mathcal{D})$, but for circular scatterers, orthogonality leads to a matrix that is diagonal. Consequently, we can write

$$A_m^p(\mathcal{D}) + Z_m I_m^p(\mathcal{D}) = 0, \tag{3.4}$$

where Z_m is a scattering coefficient which is given by

$$Z_m = J_m(ka)/\, H_m^{(1)}(ka) \tag{3.5}$$

for Dirichlet boundary conditions, or

$$Z_m = J_m'(ka)/\, H_m^{(1)\prime}(ka) \tag{3.6}$$

for Neumann conditions. Other expressions for Z_m can be used to model different situations, such as impedance boundary conditions.

Scattering problems of this type can be separated into components that are symmetric and antisymmetric about $y = 0$ by decomposing the incident field ϕ^{i} into an even (subscript "+") and an odd (subscript "−") function of y; thus

$$\phi_{\pm}^{\mathrm{i}}(x, y) = \frac{1}{2}\left[\phi^{\mathrm{i}}(x, y) \pm \phi^{\mathrm{i}}(x, -y)\right]. \tag{3.7}$$

If the array is excited by incident wave $\phi_{\pm}^{\mathrm{i}}(x, y)$, then the resulting coefficients A_m^p and I_m^p satisfy the identity

$$\mathcal{U}_{-m}^p = \pm(-1)^m \mathcal{U}_m^p. \tag{3.8}$$

This often leads to useful simplifications, and also to an increase in performance when inverting linear systems. For brevity, we will give equations for the complete

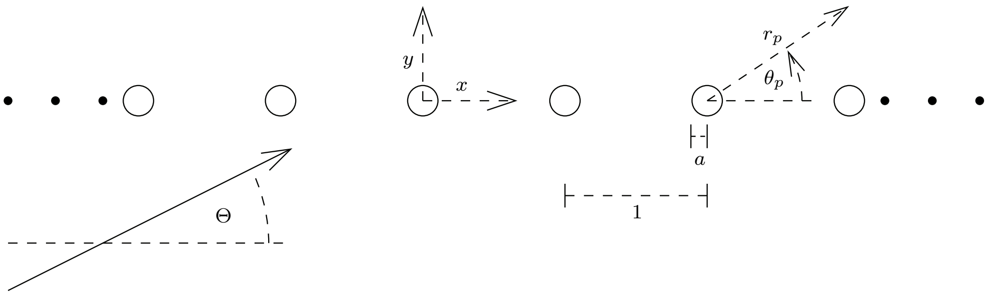


FIG. 3.1. *Schematic diagram of a periodic array with circular scatterers centered at the points* $(p,0)$ *in the* (x,y) *plane, and a plane wave incident at angle* Θ.

wavefield, and it is to be understood that these can always be decomposed in the manner described above.

To conclude this section, we will now review the theory of periodic arrays, starting with the method for obtaining the coefficients $A_m^p(\emptyset)$ in the case where the incident field is the plane wave

$$\phi^{\mathrm{i}}(x,y) = \mathrm{e}^{\mathrm{i}k(x\cos\Theta + y\sin\Theta)}; \tag{3.9}$$

see Figure 3.1. Values for $A_m^p(\emptyset)$ are of course required as a starting point, and the technique used to obtain them provides some motivation for the MASM developed in section 4 to solve the canonical problems. First of all, we require a second set of equations relating the coefficients $A_m^p(\emptyset)$ and $I_m^p(\emptyset)$. This will form a closed system, when combined with (3.4), and can be obtained using Graf's addition theorem [14, section 2.5]. For the specific case under consideration here, we have

$$I_m^p(\emptyset) = \mathrm{i}^m \mathrm{e}^{\mathrm{i}pk\cos\Theta}\mathrm{e}^{-\mathrm{i}m\Theta} + \sum_v \sum_{j\neq p} A_v^j(\emptyset) X_{v-m}^{p-j}\, H_{v-m}^{(1)}(k|p-j|), \tag{3.10}$$

where $X_v^p = \operatorname{sgn}(p)^v$. Given that the only difference between the field at the point (x,y) and that at $(x+j,y)$, $j\in\mathbb{Z}$, is a phase shift due to the incident plane wave, this can be simplified by seeking a solution for which

$$A_m^p(\emptyset) = A_m^0(\emptyset)\mathrm{e}^{\mathrm{i}pk\cos\Theta}. \tag{3.11}$$

Enforcing the boundary conditions via (3.4), and then making use of (3.11), we obtain

$$A_m^0(\emptyset) + Z_m \sum_v A_v^0(\emptyset)\sigma_{v-m}(k\cos\Theta) = -Z_m \mathrm{i}^m \mathrm{e}^{-\mathrm{i}m\Theta}, \tag{3.12}$$

which is a linear system involving only an order sum. The function $\sigma_n(t)$ is a Schlömilch series of order n, i.e.,

$$\sigma_n(t) = \sum_{j=1}^{\infty} \left[\mathrm{e}^{-\mathrm{i}jt} + (-1)^n \mathrm{e}^{\mathrm{i}jt}\right] H_n^{(1)}(kj). \tag{3.13}$$

If the values of k and Θ are such that the Schlömilch series are divergent, the values for $A_m^p(\emptyset)$ can be obtained as in [12]. Note that $\sigma_{-n}(t) = (-1)^n\sigma_n(t)$.

The Schlömilch series is a type of lattice sum, and the capacity to evaluate these accurately and efficiently is crucial to the analysis of wave interactions with arrays.

For the case under consideration here, the well-known Twersky formulae [25, 8] can be used. The singularity structure of $\sigma_n(t)$ must be considered when applying the MASM, and so we note that

$$\sigma_n(t) = b_n(t) + 2(-\mathrm{i})^{n+1}\left[\mu_n^0(t) + \sum_{j=1}^{\infty}\left(\mu_n^j(t) + \mu_{-n}^{-j}(t) - \frac{\delta_{n0}}{\pi j}\right)\right], \tag{3.14}$$

where $b_n(t)$ is an entire function that can be expressed as a finite sum of Bernoulli polynomials, and

$$\mu_n^j(t) = \frac{[t + 2j\pi - \gamma(t+2j\pi)]^n}{k^n\gamma(t+2j\pi)}. \tag{3.15}$$

The function $\gamma(t)$ is defined for real t via

$$\gamma(t) = \begin{cases} \sqrt{t^2-k^2} & : |t| \geq k, \\ -\mathrm{i}\sqrt{k^2-t^2} & : |t| < k. \end{cases} \tag{3.16}$$

For $n = 0$, 1, and 2, the summand in (3.14) is $O(j^{-3})$ as $j \to \infty$; for larger n it is $O(j^{-5})$ or smaller. The rate of convergence can easily be accelerated by expanding the summand in (3.14) for large j. Where derivatives are required, the formula

$$\frac{\mathrm{d}\mu_n^j}{\mathrm{d}t} = \frac{-\mu_n^j(t)}{\gamma(t+2j\pi)}\left[n + \frac{t+2j\pi}{\gamma(t+2j\pi)}\right] \tag{3.17}$$

can be used. The infinite summation in the resulting formula for $\sigma_n'(t)$ has a summand that is $O(j^{-3})$ as $j \to \infty$ for $n = 0$ and $n = 1$, and $O(j^{-5})$ or smaller for larger values of n. Again, the convergence can be accelerated where necessary.

An important property of infinite periodic arrays is their capacity to support RB surface waves in some circumstances. These propagate without loss along the array and decay exponentially in other directions. The presence of RB waves corresponds to the existence of nontrivial homogeneous solutions to the periodic array problem with the form

$$\phi_{RB}^{\mathrm{t}}(\mathbf{r}) = \sum_m \sum_p \widetilde{B}_m \mathrm{e}^{\mathrm{i}p\widetilde{\beta}}\, H_m^{(1)}(kr_p)\mathrm{e}^{\mathrm{i}m\theta_p}, \tag{3.18}$$

where $\widetilde{\beta} \in \mathbb{R}$ is an arbitrary phase shift. The coefficients $\widetilde{B}_m$ satisfy the same system of equations as $A_m^0(\emptyset)$ (i.e., (3.12)), but with the right-hand side set to zero and $k\cos\Theta$ replaced by $\widetilde{\beta}$; thus

$$\widetilde{B}_m + Z_m \sum_v \widetilde{B}_v \sigma_{v-m}(\widetilde{\beta}) = 0, \tag{3.19}$$

in which $\widetilde{B}_m \neq 0$ for at least one m. A straightforward method for finding the appropriate values for $\widetilde{\beta}$ is given in [6]. The associated coefficients $\widetilde{B}_m$ are then normalized so that

$$\sum_m |\widetilde{B}_m|^2 = 1. \tag{3.20}$$

Given the evident 2π-periodicity of the Schlömilch series (3.13), distinct solutions to (3.19) can occur only for $\widetilde{\beta} \in [0, 2\pi)$. Full details of the parameter ranges for which RB modes have been found are given in [23]. Here we summarize the important details.

If the surface of the scatterers is subject to a Dirichlet boundary condition, then RB waves do not occur [2]. On the other hand, if a Neumann boundary condition is in use, then up to two distinct modes are known to exist. One of these is symmetric about $y = 0$; this can occur for scatterers of any size, for a range of wavenumbers $0 < k < k^{\mathrm{s}}_{\max} < \pi$. The other is an antisymmetric mode which exists in the range $k^{\mathrm{a}}_{\min} < k < k^{\mathrm{a}}_{\max} < \pi$, but only if $a \gtrsim 0.403$. The cut-off values depend upon the scatterer radius a. Outside the given ranges for k, the RB wave is replaced by a mode that is evanescent in x. In both the symmetric and antisymmetric cases, the principal value for $\widetilde{\beta}$ lies in the interval (k, π) and corresponds to a right-propagating wave. The associated left-propagating mode has the phase shift $2\pi - \widetilde{\beta}$ in place of $\widetilde{\beta}$ and the coefficient $(-1)^m \widetilde{B}_m$ in place of $\widetilde{B}_m$. As $k \to k_{\max}$, $\widetilde{\beta} \to \pi$, i.e., the RB modes become standing waves. The amplitude with which RB modes are excited is a key parameter in the solution, and obtaining this is a major goal of our analysis. In what follows, we will assume that exactly one type of RB mode occurs (i.e., symmetric or antisymmetric). It is not difficult to modify our subsequent analysis if this is not the case. In a problem where the incident wave has been decomposed using (3.7), there is at most one mode for each component of the solution.

4. Canonical problems.

In order to proceed, we must determine ψ^0_n, i.e., the total field that occurs when scatterer 0 is replaced by a unit source of order n. In this case, we have the expansion

$$\psi^0_n(\mathbf{r}) = \sum_m \sum_p C^{p,0}_{m,n} \, H^{(1)}_m(kr_p) \mathrm{e}^{\mathrm{i}m\theta_p}, \tag{4.1}$$

where

$$C^{0,0}_{m,n} = \delta_{mn}, \tag{4.2}$$

as in (2.12). A useful symmetry relation can be obtained by changing x to $-x$ and y to $-y$ (and therefore $r_p \to r_{-p}$ and $\theta_p \to \pi + \theta_{-p}$) in (4.1). After applying (4.2) and comparing the result to (4.1), we find that

$$C^{-p,0}_{m,n} = (-1)^{m+n} C^{p,0}_{m,n}. \tag{4.3}$$

As before, a linear system for the unknown coefficients can be obtained by locally expanding ψ^0_n about the point $r_p = 0$; thus

$$\psi^0_n(r_p, \theta_p) = \sum_m \Big[K^{p,0}_{m,n} \, J_m(kr_p) + C^{p,0}_{m,n} \, H^{(1)}_m(kr_p) \Big] \mathrm{e}^{\mathrm{i}m\theta_p}. \tag{4.4}$$

An expression for the incoming field coefficients $K^{p,0}_{m,n}$ in terms of the outgoing coefficients $C^{p,0}_{m,n}$ can be deduced from (3.10) by simply omitting the term due to plane wave forcing. We find that

$$K^{p,0}_{m,n} = \sum_v \sum_{j \neq p} C^{j,0}_{v,n} X^{p-j}_{v-m} \, H^{(1)}_{v-m}(k|p-j|), \tag{4.5}$$

and the boundary condition gives

$$C^{p,0}_{m,n} + Z_m K^{p,0}_{m,n} = 0, \quad p \neq 0. \tag{4.6}$$

The MASM can now be used to obtain an expression for $C^{p,0}_{m,n}$. The principal idea is derived from the original array scanning method [27, 16, 23], in which the unknown coefficients are represented as Fourier integrals. First of all, introduce damping by writing

$$k = \mathrm{Re}[k] + \mathrm{i}\epsilon, \tag{4.7}$$

where $\epsilon > 0$. This ensures the convergence of the summations over the spatial index in subsequent equations. Once the solutions are obtained, we can take the limit $\epsilon \to 0$ to retrieve the time-harmonic field. Next, define the function $f_{m,n}(t)$ by writing

$$f_{m,n}(t) = \mathrm{i} \sum_p C^{p,0}_{m,n} \mathrm{e}^{-\mathrm{i}pt}, \tag{4.8}$$

so that we have

$$C^{p,0}_{m,n} = \frac{1}{2\pi \mathrm{i}} \int_0^{2\pi} f_{m,n}(t) \mathrm{e}^{\mathrm{i}pt} \, \mathrm{d}t. \tag{4.9}$$

One motivation for this choice of representation is that the spatial dependence of the integral is such that if we substitute (4.9) into (4.5), the sum over j will become a Schlömilch series as in (3.13). Indeed, combining (4.5), (4.6), and (4.9), we find that

$$\int_0^{2\pi} \left[f_{m,n}(t) + Z_m \sum_v f_{v,n}(t) \sigma_{v-m}(t) \right] \mathrm{e}^{\mathrm{i}pt} \, \mathrm{d}t = 0, \quad p \neq 0. \tag{4.10}$$

A second motivation for (4.8) is that the integration in (4.9) facilitates a simple means by which the left-hand side of (4.10) can be made to vanish for all $p \neq 0$. If we now write

$$f_{m,n}(t) + Z_m \sum_v f_{v,n}(t) \sigma_{v-m}(t) = \mathcal{F}_{m,n}(t), \tag{4.11}$$

then (4.10) becomes

$$\int_0^{2\pi} \mathcal{F}_{m,n}(t) \mathrm{e}^{\mathrm{i}pt} \, \mathrm{d}t = 0, \quad p \neq 0. \tag{4.12}$$

By considering the Fourier series expansions of $\mathcal{F}_{m,n}(t)$, it becomes clear that (4.11) can be satisfied if and only if these functions are constants, which we denote by $\mathcal{F}_{m,n}$. The values for these are fixed by setting $p = 0$ in (4.9) and imposing the requirement (4.2); hence

$$\frac{1}{2\pi \mathrm{i}} \int_0^{2\pi} f_{m,n}(t) \, \mathrm{d}t = \delta_{mn}. \tag{4.13}$$

Note that the system of equations (4.11) contains only an order sum, and also that the source order n does not affect the operator on the left-hand side, which is of exactly the same form as those appearing in (3.12) and (3.19), with the variable t taking the place of the parameters $k \cos \Theta$ and $\widetilde{\beta}$.

In order to determine the coefficients $\mathcal{F}_{m,n}$, we introduce the function $g_{m,n}(t)$ as the solution to the linear system (4.11), but with the right-hand side replaced by δ_{mn}, i.e.,

$$g_{m,n}(t) + Z_m \sum_v g_{v,n}(t) \sigma_{v-m}(t) = \delta_{mn}. \tag{4.14}$$

Since the right-hand side is known, this system of equations can be inverted numerically for any value of t at which both $\sigma_n(t)$ and $g_{v,n}(t)$ are analytic. If (4.14) is multiplied by $\mathcal{F}_{n,u}$ and then summed over all integers n, we see that $g_{m,n}(t)$ is related to $f_{m,n}(t)$ via

$$\sum_v g_{m,v}(t)\mathcal{F}_{v,n} = f_{m,n}(t). \tag{4.15}$$

Integrating (4.15) yields

$$\frac{1}{2\pi \mathrm{i}} \sum_v \mathcal{F}_{v,n} \int_0^{2\pi} g_{m,v}(t)\,\mathrm{d}t = \delta_{mn}, \tag{4.16}$$

in view of (4.13). In principle, therefore, the solutions to the canonical problems are now available—take the limit $\epsilon \to 0$ in (4.7) and then apply quadrature to compute the integrals in (4.16). This latter step is discussed in the appendix. This done, the resulting linear system can be inverted to yield $\mathcal{F}_{m,n}$. However, taking the limit $\epsilon \to 0$ in (4.7) will cause singularities to appear on the real line, and so we must determine the correct indentations for the path of integration.

First, for any $k > 0$, there exists $\lambda \in \mathbb{Z}$ such that $\mathrm{Re}[k_\lambda] \in [0, 2\pi]$, where

$$k_\lambda = k + 2\lambda\pi. \tag{4.17}$$

Equation (3.14) shows that the function $\sigma_n(t)$ has a branch point at $t = k_\lambda$; another is located at $t = 2\pi - k_\lambda$. Note that $\mathrm{Im}[k_\lambda] = \epsilon$, and $\mathrm{Im}[2\pi - k_\lambda] = -\epsilon$. The functions $f_{m,n}(t)$ and $g_{m,n}(t)$ will inherit these singularities via (4.11) and (4.14), respectively. The special case in which $k_\lambda = 2\pi - k_\lambda = \pi$ can be handled by adjusting the path of integration in (4.9) to run from $-\pi$ to π.

A second important possibility is that, after taking the limit $\epsilon \to 0$ in (4.7), there may exist real values of t at which the matrix of known coefficients appearing on the left-hand side of (4.11) and (4.14) is singular. These correspond to the existence of RB waves, as discussed in section 3; at $t = \widetilde{\beta}$ the left-hand side of (4.11) (and also (4.14)) is identical to that of (3.19). In general, the Fredholm alternative permits solutions at $t = \widetilde{\beta}$ and $t = 2\pi - \widetilde{\beta}$ if the functions $f_{m,n}(t)$ and $g_{m,n}(t)$ possess simple poles at these points. Numerical results in [6] show that $\mathrm{d}\widetilde{\beta}/\mathrm{d}k > 0$, and so when we add damping using (4.7) the pole at $t = \widetilde{\beta}$ moves above the real line, and that at $t = 2\pi - \widetilde{\beta}$ moves below. If we now let $\epsilon \to 0$ in (4.7) so as to retrieve the time-harmonic solution, we find that the correct indentations for the path of integration are those shown in Figure 4.1. This is the only configuration that leads to a purely outgoing scattered field in the limit $\sqrt{x^2 + y^2} \to \infty$. The residues at the poles of $f_{m,n}(t)$ determine the amplitudes of any RB waves that are excited, and these make a contribution to $C^{p;0}_{m,n}$ that does not decay in the limit $|p| \to \infty$. We now calculate these, using the method in [23]. First, multiply (4.11) by $t - \widetilde{\beta}$ and then take the limit $t \to \widetilde{\beta}$. The residue of the function $f_{m,n}(t)$ at the pole must satisfy the resulting homogeneous linear system, which is identical to (3.19), and hence

$$\operatorname*{Res}_{t=\widetilde{\beta}} f_{m,n}(t) = c_n \widetilde{B}_m, \tag{4.18}$$

for some constant c_n. Essentially, the coefficients $\widetilde{B}_m$ describe the shape of the RB wave, and c_n is the amplitude. The same procedure can then be applied with $\widetilde{\beta}$

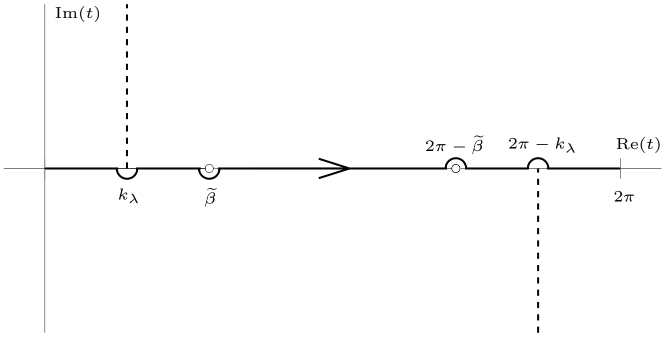


FIG. 4.1. *Singularity structure and indented contour of integration in the t plane. The poles at $t = \widetilde{\beta}$ and $t = 2\pi - \widetilde{\beta}$ do not occur if Dirichlet boundary conditions are imposed on the surface of the regular scatterers, or if $\lambda \neq 0$ (i.e., $k > \pi$).*

replaced by $2\pi - \widetilde{\beta}$, (and $\widetilde{B}_m$ by $(-1)^m \widetilde{B}_m$) and the symmetry relation (4.3) then shows that

$$f_{m,n}(t) = \widehat{f}_{m,n}(t) + c_n \widetilde{B}_m \left[\frac{1}{t - \widetilde{\beta}} - \frac{(-1)^{m+n}}{t + \widetilde{\beta} - 2\pi} \right], \tag{4.19}$$

where $\widehat{f}_{m,n}(t)$ is analytic at $t = \widetilde{\beta}$ and $t = 2\pi - \widetilde{\beta}$. Finally, substitute (4.19) into (4.11), transfer the terms with denominator $t - \widetilde{\beta}$ to the right-hand side, and take the limit $t \to \widetilde{\beta}$ using L'Hôpital's rule as appropriate. We can now apply the Fredholm alternative to the resulting linear system. The left-hand side consists of a singular matrix, multiplied by a vector of bounded functions. Therefore, a solution can exist if and only if the right-hand side is orthogonal to the (nontrivial) solution to the homogeneous adjoint problem [20, eqns. (5.7)–(5.9)]. The latter is easily shown to be $\widetilde{B}_m / Z_m^*$ [23], leading to the following equation for c_n:

$$\sum_m \frac{\widetilde{B}_m^* \mathcal{F}_{m,n}}{Z_m} = c_n \sum_m \widetilde{B}_m^* \sum_v \widetilde{B}_v \sigma'_{v-m}(\widetilde{\beta}). \tag{4.20}$$

Here, the superscript "$*$" denotes the complex conjugate, and the prime a derivative with respect to the argument. The residues of the function $g_{m,n}(t)$ can be calculated in exactly the same way; simply replace f with g in (4.18) and (4.19) and $\mathcal{F}_{m,n}$ with δ_{mn} in (4.20).

The asymptotic behavior of $C_{m,n}^{p,0}$ in the limit $|p| \to \infty$ can be obtained by noting that $f_{m,n}(t)$ is 2π-periodic, this property being inherited from the Schlömilch series via (4.11). Consequently, if the path of integration in (4.9) is closed in the upper half plane, the contributions from $t = \mathrm{i}u$ and $t = 2\pi + \mathrm{i}u$, $u > 0$, cancel each other. Therefore, as $p \to \infty$, we have

$$C_{m,n}^{p,0} \sim c_n \widetilde{B}_m \mathrm{e}^{\mathrm{i}p\widetilde{\beta}} + C_{m,n} \frac{\mathrm{e}^{\mathrm{i}kp}}{p^{3/2}} + O(p^{-5/2}). \tag{4.21}$$

Here, the second term on the right-hand side is the dominant contribution from the branch point at $t = k_\lambda$. The dependence upon p can be deduced by using the method in [12] to show that $f_{m,n}(t)$ remains finite as $t \to k_\lambda$. Given that $t = k_\lambda$ is a branch point of square root type, the result follows. In principle, one can also obtain a formula for the coefficient $C_{m,n}$ using a similar technique, but this is somewhat involved. The behavior of $C^{p,0}_{m,n}$ in the limit $p \to -\infty$ can be deduced by closing the contour of integration in (4.9) in the lower half plane; alternatively, the symmetry relation (4.3) can be used.

A final point concerns the field incoming toward the source, which must be calculated if we are dealing with defects that do not consist of absent scatterers. From (4.4) it is seen that this amounts to finding the value of $K^{0,0}_{m,n}$, which can be achieved by setting $p = 0$ in (4.5). Both the spatial sum and the order sum can be evaluated exactly. Thus, on using (4.9), we have

$$K^{0,0}_{m,n} = \frac{1}{2\pi \mathrm{i}} \int_0^{2\pi} \sum_v f_{v,n}(t)\sigma_{v-m}(t)\ \mathrm{d}t. \tag{4.22}$$

Equation (4.11) reduces this to an integral whose value is known in view of (4.13), the result being

$$Z_m K^{0,0}_{m,n} = -\delta_{mn} - \mathrm{i}\mathcal{F}_{m,n}. \tag{4.23}$$

5. Illustrative results. In this section we present some numerical results for a variety of different situations. We also show how the interaction theory for defects can be combined with the infinite array subtraction technique developed in [11], and the large array approximation used in [24] to validate results, and extend the range of applicability. Particular attention is paid to the determination of the amplitude with which RB waves are excited by the defects. Accurate computation by more direct numerical methods is difficult (see the appendix, and also [11]), but our approach is numerically efficient, and we are able to compute the amplitudes for all possible k and Θ. In view of the number of cases that can be solved, we have not carried out a comprehensive parameter survey, but instead we have attempted to provide a representative sample of the types of result that can be obtained. In performing any such calculations, the rapidly convergent order summations that occur throughout our analysis must be truncated at some suitable value, which depends on the size of ka. This must be chosen to be large enough to yield accurate results, but not so large as to unnecessarily increase program execution time or generate near singular linear systems. The truncation levels used by our numerical codes are the same as those reported in [24]. Unless otherwise stated, Neumann boundary conditions are applied on the surface of the regular scatterers.

5.1. Localized defects. In cases where the defects are confined to a small section of the array, all of the relevant integrals can easily be evaluated by quadrature. The asymptotic behavior of $A^p_m(\mathcal{D})$ for large p can be obtained using (2.16), (2.17), (3.11), and (4.21). A formula for large, negative p can be obtained in a similar way by applying the symmetry relation (4.3) in (2.17). We find that, as $p \to \pm\infty$,

$$A^p_m(\mathcal{D}) \sim A^0_m(\emptyset)\mathrm{e}^{\mathrm{i}kp\cos\Theta} + (\pm 1)^m \Gamma^{\pm} \widetilde{B}_m \mathrm{e}^{\mathrm{i}|p|\widetilde{\beta}} + O(|p|^{-3/2}), \tag{5.1}$$

where

$$\Gamma^{\pm} = \sum_n (\pm 1)^n c_n \sum_{q\in\mathcal{D}} a^q_n \mathrm{e}^{\mp\mathrm{i}q\widetilde{\beta}}. \tag{5.2}$$

The quantity Γ^{+} (Γ^{-}) is the complex amplitude coefficient of the right- (left-) propagating RB wave that is excited by the defect. This depends upon the solutions to

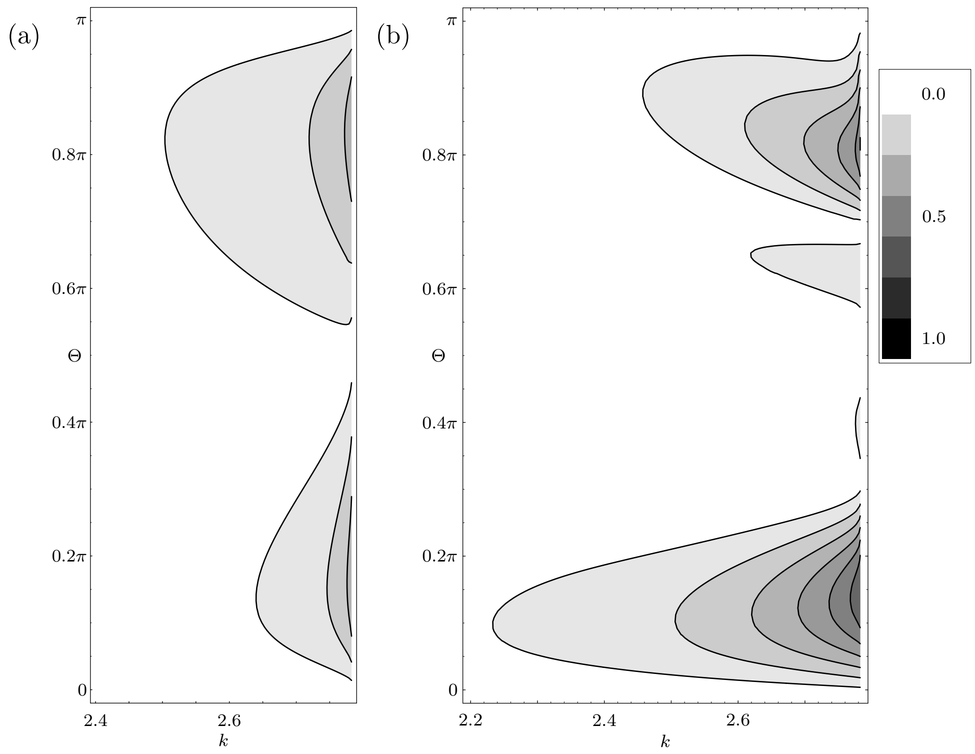


FIG. 5.1. *Contour plots showing the amplitude of the right-propagating RB wave,* $|\Gamma^+|$, *for* $a = 0.25$ *with varying* k *and* Θ. (a) $\mathcal{D} = \{0\}$. (b) $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$.

the canonical problems and the interactions between the defects via the coefficients c_n, and the sum over $\mathcal{D}$, respectively.

Figure 5.1 shows contour plots of $|\Gamma^+|$ with varying k and Θ, for $a = 0.25$ and two different defect sets: $\mathcal{D} = \{0\}$ and $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$. Figure 5.2 shows similar plots, but for the antisymmetric RB wave on an array with $a = 0.49$. In all cases, $|\Gamma^-|$ can be deduced by symmetry. The computation time required to obtain data for figures such as these is greatly reduced by the fact that the canonical problems need only be solved once for each value of k. The general trend for the amplitude to increase with k is consistent with the cases of excitation at an array end [11], and by an aperiodic source [23]. The upper limit for k is the cut-off ($k \approx 2.783$ for $a = 0.25$ and $k \approx 2.971$ for the antisymmetric mode on an array with $a = 0.49$), above which the RB waves cease to exist. For all values of k smaller than those shown, the symmetric mode exists but is excited at a very low amplitude. The antisymmetric mode does not exist for $k \lessapprox 1.796$; for intermediate values up to those that are shown in Figure 5.2 the excitation amplitude is small. The dependence of Γ^+ upon the angle of incidence Θ exhibits a number of interesting features. First, the surface wave is cut off completely as $\Theta \to 0$ and $\Theta \to \pi$. In fact, the total field vanishes in these limits, as demonstrated in [12]; the presence of a finite set of defects has no bearing on this. The cut-off at $\Theta = 0$ is sharper in Figures 5.1(b) and 5.2(b) than it is in Figures 5.1(a) and 5.2(a); this is consistent with the case of excitation at the end of a semi-infinite array, where the cut-off disappears, and the amplitude is generally greatest at head-on incidence [11]. The two-peak structure, and the fact that the relative size

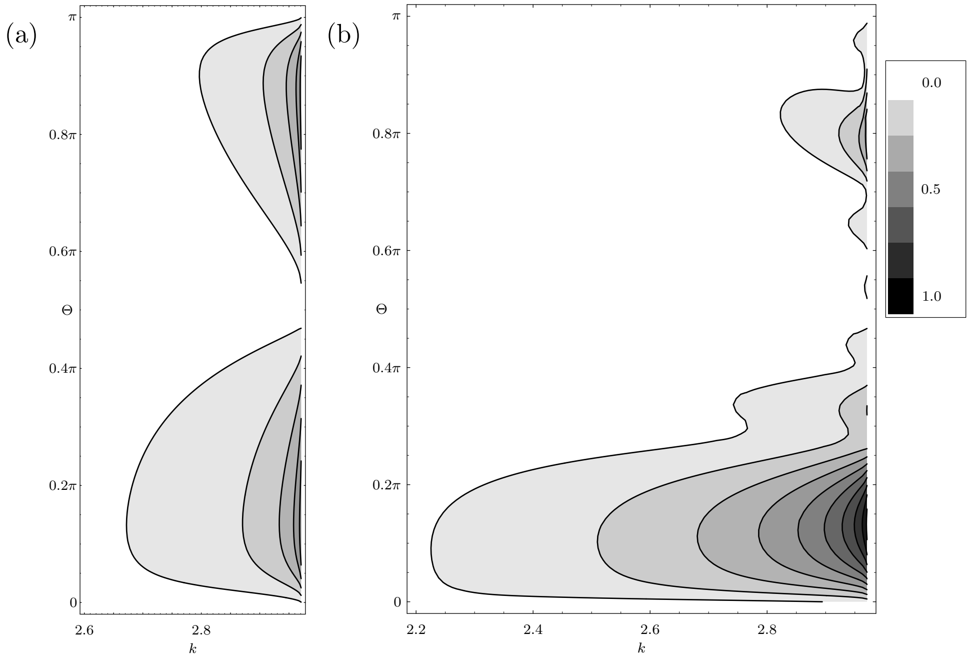


FIG. 5.2. *Contour plots showing the amplitude of the right propagating antisymmetric RB wave,* $|\Gamma^+|$, *for* $a = 0.49$ *with varying* k *and* Θ. (a) $\mathcal{D} = \{0\}$, (b) $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$.

of the peak at $\Theta \approx 0.8\pi$ is reduced for the larger defect sets, is also consistent with the case of excitation at an end. Finally, note that for the case where $\mathcal{D} = \{0\}$, the amplitude of the symmetric right-propagating RB wave is largest when $\Theta \approx 0.8\pi$, which corresponds to an incident field whose x-component is propagating to the left.

The infinite array subtraction methods introduced in [11] provide a useful means of validating results such as those shown in Figures 5.1 and 5.2. If we write

$$D_m^p(\mathcal{D}) = A_m^p(\mathcal{D}) - A_m^p(\emptyset), \tag{5.3}$$

then, on recalling that $A_m^p(\mathcal{D}) = 0$ for $p \in \mathcal{D}$, it is not difficult to use the results in section 3 to show that the coefficients $D_m^p(\mathcal{D})$ satisfy the linear system of equations

$$\begin{aligned} D_m^p(\mathcal{D}) + Z_m \sum_v \sum_{\substack{j \notin \mathcal{D} \\ j \neq p}} D_v^j(\mathcal{D}) X_{v-m}^{p-j}\, H_{v-m}^{(1)}(k|p-j|) \\ = Z_m \sum_v A_v^0(\emptyset) \sum_{j \in \mathcal{D}} \mathrm{e}^{\mathrm{i}jk\cos\Theta} X_{v-m}^{p-j}\, H_{v-m}^{(1)}(k|p-j|), \quad p \notin \mathcal{D}. \end{aligned} \tag{5.4}$$

Note that the right-hand side has been simplified using (3.4) (with $\mathcal{D} = \emptyset$) and (3.10). If no RB waves are present, then we should expect that $D_p \to 0$ as $|p| \to \infty$. On the other hand, if RB waves are present in the solution, their contribution can be isolated by writing

$$D_m^p(\mathcal{D}) = \begin{cases} \widehat{D}_m^p(\mathcal{D}) + \Gamma^- \mathrm{e}^{-\mathrm{i}p\widetilde{\beta}}(-1)^m \widetilde{B}_m & : p \le p_0, \\ \widehat{D}_m^p(\mathcal{D}) & : p_0 < p < p_1, \\ \widehat{D}_m^p(\mathcal{D}) + \Gamma^+ \mathrm{e}^{\mathrm{i}p\widetilde{\beta}} \widetilde{B}_m & : p \ge p_1, \end{cases} \tag{5.5}$$

where p_0 and p_1 are chosen so that the array is regular for $p \leq p_0$ and $p \geq p_1$. Substituting this into (5.4), we find that the coefficients $\widehat{D}^p_m(\mathcal{D})$ satisfy a linear system which has the same left-hand side as (5.4) and correction terms L^p_m and R^p_m (due to the left- and right-propagating RB waves, respectively) added to the right-hand side. A straightforward calculation shows that

$$L^p_m = \begin{cases} -\Gamma^- Z_m \mathrm{e}^{-\mathrm{i}p\widetilde{\beta}} \sum_v (-1)^v \widetilde{B}_v S^{p-p_0}_{v-m}(\widetilde{\beta}) & : p > p_0, \\ \Gamma^- (-1)^m Z_m \mathrm{e}^{-\mathrm{i}p\widetilde{\beta}} \sum_v \widetilde{B}_v S^{1+p_0-p}_{v-m}(-\widetilde{\beta}) & : p \leq p_0 \end{cases} \tag{5.6}$$

and

$$R^p_m = \begin{cases} -\Gamma^+ Z_m \mathrm{e}^{\mathrm{i}p\widetilde{\beta}} \sum_v \widetilde{B}_v S^{p_1-p}_{m-v}(\widetilde{\beta}) & : p < p_1, \\ \Gamma^+ Z_m \mathrm{e}^{\mathrm{i}p\widetilde{\beta}} \sum_v \widetilde{B}_v S^{1+p-p_1}_{v-m}(-\widetilde{\beta}) & : p \geq p_1, \end{cases} \tag{5.7}$$

where

$$S^p_m(\beta) = \sum_{j \geq p} \mathrm{e}^{\mathrm{i}j\beta} H^{(1)}_m(kj); \tag{5.8}$$

this half range Schlömilch series can be efficiently computed using methods in [8]. The fact that the RB wave is a homogeneous solution to the periodic array problem has been used to simplify L^p_m for $p \leq p_0$ and R^p_m for $p \geq p_1$. If we now solve the linear system for $\widehat{D}^p_m(\mathcal{D})$, the solution will decay as $|p| \to \infty$, but *only* if the correct values for the RB amplitudes Γ^- and Γ^+ are used.

As an example, consider the parameter set $a = 0.25$, $k = 2.5$, $\Theta = 0.1\pi$, and $\mathcal{D} = \{0\}$, which is included in Figure 5.1. Figure 5.3 shows a logarithmic plot of D_p for this case, where

$$D_p = \sum_m \left| D^p_m(\mathcal{D}) \right|^2 . \tag{5.9}$$

This provides a simple measure of the difference between the scattered fields in the periodic and defective array problems. The data are obtained by truncating the system (5.4) at $|p| = 100$. Clearly, D_p does not decay as $|p| \to \infty$; instead it oscillates about a fixed value corresponding to the amplitude of the RB wave. As in Figure 5.1, this is stronger to the left of the defect. The quantity $\widehat{D}_p$ is also plotted in Figure 5.3. This is obtained by replacing $D^p_m(\mathcal{D})$ with $\widehat{D}^p_m(\mathcal{D})$ in (5.9). Values for Γ^- ($\approx 0.07613 + 0.02638\mathrm{i}$) and Γ^+ ($\approx -0.04897 + 0.00176\mathrm{i}$) are obtained using (2.15) and (5.2), with the canonical problems solved using the MASM. These values are then used in (5.6) and (5.7). The fact that $\widehat{D}_p$ decays as $|p|$ is increased confirms that these amplitudes are indeed correct.

5.2. Irregular scatterers. In cases where the defects do not consist of absent scatterers, we must close the system of equations for a^p_m (2.18) by applying boundary conditions on the surface of the irregular array elements. We will assume that the irregular scatterers differ from the other array elements in either size, surface composition, or possibly both. In such cases, we can impose the boundary condition for the

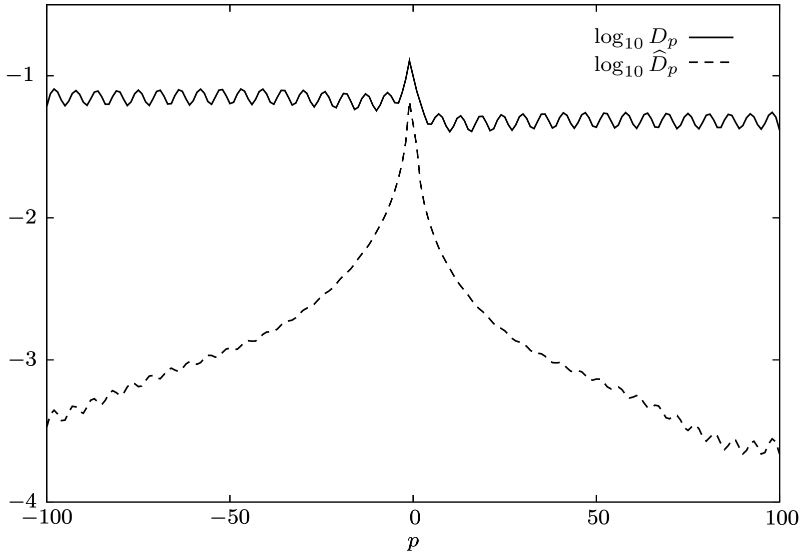


FIG. 5.3. *Logarithmic plot of* D_p *and* $\widehat{D}_p$ *for* $a = 0.25$, $k = 2.5$, $\Theta = 0.1\pi$, *and* $\mathcal{D} = \{0\}$.

irregular scatterers using an equation similar to (3.4), but with a different scattering coefficient $\widehat{Z}_m$; thus

$$A_m^p(\mathcal{D}) + \widehat{Z}_m I_m^p(\mathcal{D}) = 0, \quad p \in \mathcal{D}. \tag{5.10}$$

As before, it is not difficult to incorporate transfer matrices so as to deal with scatterers of a different shape. Equation (5.10) is to be used in conjunction with

$$A_m^p(\mathcal{D}) = A_m^p(\emptyset) + \sum_n \sum_{q\in\mathcal{D}} a_n^q C_{m,n}^{p-q,0}, \tag{5.11}$$

which is obtained from (2.12) and (2.16)–(2.18), and is valid for *all* p. The simplest way to proceed is to deduce an expression for $I_m^p(\mathcal{D})$ from (3.10) by replacing $\emptyset$ with $\mathcal{D}$. If we then use (5.11) to decompose $A_m^p(\mathcal{D})$, the spatial sums in the resulting expression can be evaluated using (3.10) and (4.5), leading to

$$I_m^p(\mathcal{D}) = I_m^p(\emptyset) + \sum_n \sum_{q\in\mathcal{D}} a_n^q K_{m,n}^{p-q,0}. \tag{5.12}$$

The incoming field coefficients $K_{m,n}^{p,0}$ on the right-hand side can then be eliminated using (3.4) (with $\mathcal{D} = \emptyset$), (4.6), and (4.23). This amounts to exploiting the fact that (5.11) decomposes $A_m^p(\mathcal{D})$ into contributions from fields that satisfy the boundary condition for a regular scatterer at $r_p = a$ and contributions for which the local expansion of the incoming field is known from (4.23). We find that

$$-Z_m I_m^p(\mathcal{D}) = A_m^p(\emptyset) + \sum_n \sum_{q\in\mathcal{D}} a_n^q C_{m,n}^{p-q,0} + \mathrm{i} \sum_n a_n^p \mathcal{F}_{m,n}. \tag{5.13}$$

Finally, we can form a closed system for a_m^p by combining (5.13) with (5.10) and (5.11). The resulting expression is

$$-\mathrm{i}\widehat{Z}_m \sum_n a_n^p \mathcal{F}_{m,n} + (Z_m - \widehat{Z}_m) \sum_n \sum_{q\in\mathcal{D}} a_n^q C_{m,n}^{p-q,0} = (\widehat{Z}_m - Z_m) A_m^p(\emptyset), \quad p \in \mathcal{D}. \tag{5.14}$$

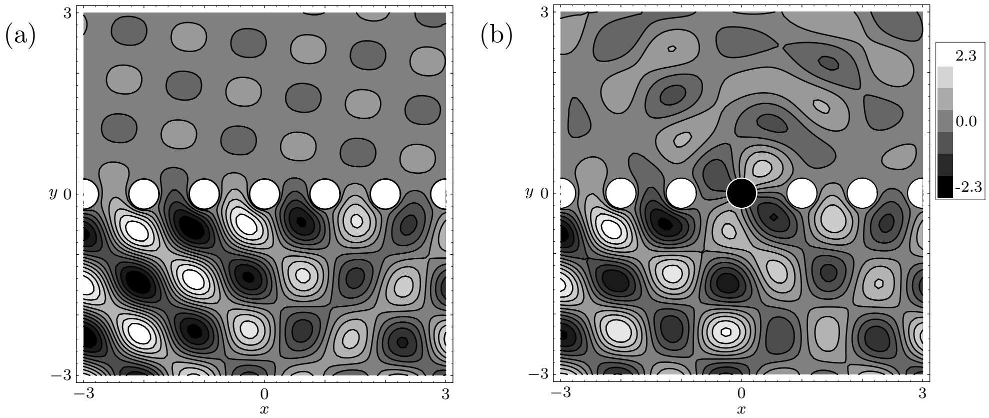


FIG. 5.4. *Contour plots showing* $\mathrm{Re}[\phi^{\mathrm{t}}]$ *for* $a = 0.25$*,* $k = 5.0$*, and* $\Theta = 0.25\pi$*, with Dirichlet boundary conditions applied on the surface of the regular scatterers.* (a) $\mathcal{D} = \emptyset$*.* (b) $\mathcal{D} = \{0\}$*. The defect is a Neumann scatterer, also with* $a = 0.25$*.*

As before, this determines the values of a_m^p; (5.11) can then be used to find values of $A_m^p(\mathcal{D})$ for $p \notin \mathcal{D}$. Note that taking $\widehat{Z}_m = 0$ returns the equation for absent scatterers, and $\widehat{Z}_m = Z_m$ yields $a_m^p = 0$, as we should expect, since then there are no defects.

Figure 5.4 shows contour plots depicting the local effects caused by replacing a single element in a periodic array with an irregular scatterer. The parameters used are $a = 0.25$, $k = 5.0$, and $\Theta = 0.25\pi$, and a Dirichlet boundary condition is applied on the surface of the regular scatterers (shown as white with a black boundary). In Figure 5.4(a), there is no defect, and the quasi-periodic nature of the field is evident. In Figure 5.4(b), the field is modified using the solutions to the canonical source problems, so that the Neumann boundary condition is now satisfied on the surface of scatterer 0 (shown as black with a white boundary). Contour lines intersecting this scatterer do so at a right angle to the surface tangent. The influence of the defect is more significant in the region above the array, because the field in the periodic case is relatively weak here.

For suitable parameters, irregular scatterers also cause RB waves to be excited. Figure 5.5 shows contour plots of $|\Gamma^+|$ for $a = 0.25$ with varying k and Θ, with $\mathcal{D} = \{0\}$ and $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$. The defects consist of Dirichlet scatterers with radius $a = 0.25$. As before, $|\Gamma^-|$ can be deduced by symmetry. The pattern of behavior here is quite different from the case of absent scatterers shown in Figure 5.1. The main qualitative difference lies in the dependence of $|\Gamma^+|$ upon Θ; there is no longer a second peak at $\Theta \approx 0.8\pi$. Elsewhere, the excitation is generally stronger than it is in the corresponding cases in Figure 5.1.

5.3. Widely spaced defects. If the defects are spread over a large section of the array, the evaluation of (4.9) by quadrature is no longer straightforward. This is because we must calculate values for $C_{m,n}^{p-q,0}$ for all $p, q \in \mathcal{D}$, and if $|p - q|$ is large, the integrals are difficult to compute. There are a number of ways to proceed. One possibility is to adopt a mixed strategy, obtaining $\Gamma^{\pm}$ using the MASM, and then solving for the decaying contributions to $C_{m,n}^{p-q,0}$ using the infinite array subtraction technique discussed in section 5.1. This yields approximate values for *all* of the unknown coefficients, and is therefore a particularly attractive idea if results for a large number of different defect sets are to be computed. Alternatively, we can form

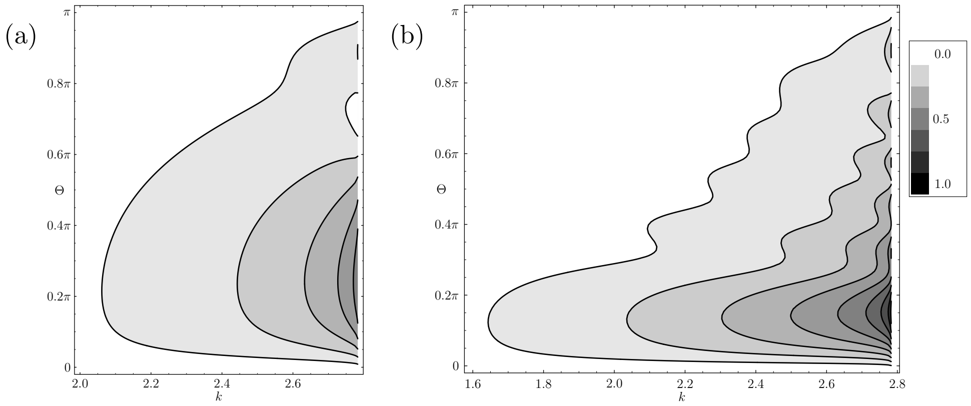


FIG. 5.5. *Contour plots showing the amplitude of the right-propagating RB wave,* $|\Gamma^+|$*, for* $a = 0.25$ *with varying* k *and* Θ*; Dirichlet boundary conditions are applied on the surface of scatterers for which* $p \in \mathcal{D}$*.* (a) $\mathcal{D} = \{0\}$. (b) $\mathcal{D} = \{0, 1, 2, 3, 4, 5\}$.

an approximate interaction theory by neglecting the decaying terms in (4.21) when $p^{3/2} \gg 1$. This approximation was used in generating the data for Figure 5.6; it amounts to assuming that significant interactions between widely spaced defects are caused solely by the RB modes and has been shown to work well in practice in the related case of a long, finite array [24].

The presence of such widely spaced defects in an array can lead to "near-trapping" in the intermediate region. This effect was originally reported in [13] in a study of interactions between water waves and long, finite arrays of bottom-mounted circular cylinders. In this physical context, the force in the x direction exerted on cylinder p by the total field (i.e., the integral of the pressure times the component of the outgoing normal to $r_p = a$ that is parallel to the array), normalized using the force exerted on an isolated cylinder, is given in [9] as

$$X_p = \left| \frac{1}{2Z_1} \left[A_1^p(\mathcal{D}) - A_{-1}^p(\mathcal{D}) \right] \right|. \tag{5.15}$$

Figure 5.6 shows a contour plot of the horizontal force on an array element that is equidistant between two widely spaced defects, each of which consists of a single absent element. The scatterer radius a is 0.25, as in [13] and in the majority of cases in [24]. The wavenumber is varied between 2.7 and the cut-off for RB waves ($k \approx 2.783$), using 1000 data points, and the angle of incidence is varied between 0 and $\pi/2$ using 500 data points. Results for $\Theta > \pi/2$ can be deduced by symmetry. The plot reveals that very large forces occur at certain discrete intervals in k and Θ. The strongest force occurs at a wavenumber that is close to, but not exactly equal to, the cut-off for RB waves. No significant peaks in the force occur for values of k smaller than those shown. The causes of the near-trapping effect are explored in [24] for the case of a finite array; the mechanism here is much the same. Essentially, RB waves generated by one periodicity breaking feature (end or defect) are reflected back by the other. The magnitude of the reflection coefficient increases as $k \to k_{\max}$. Peaks in the force correspond to situations where the interference is predominantly constructive between RB modes excited by the interaction of the incident wave with the defects and those generated by reflection.

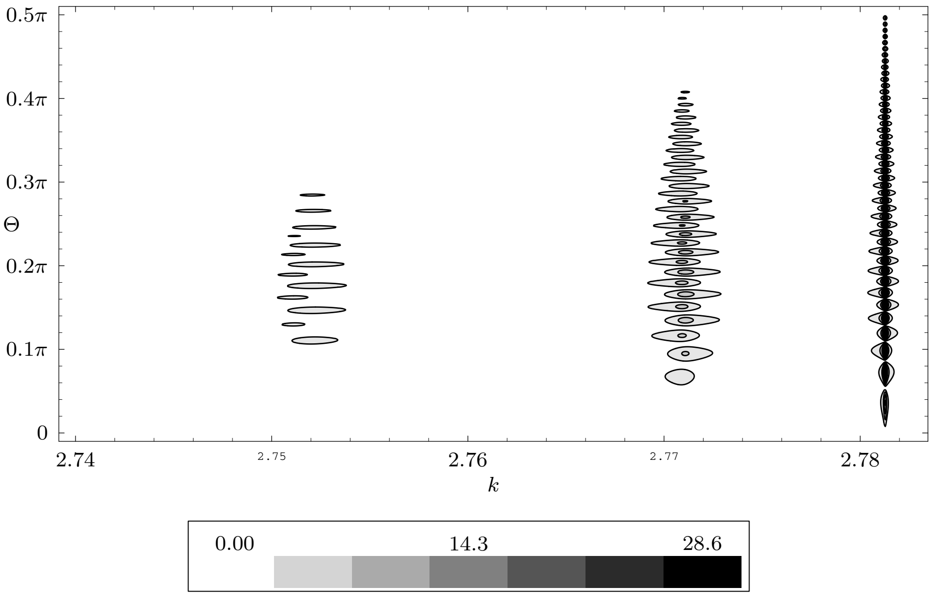


FIG. 5.6. *Contour plot showing the force exerted on scatterer* 51*, with* $a = 0.25$ *and* $\mathcal{D} = \{0, 102\}$*.*

6. Concluding remarks. By reducing the problem of scattering by a defective array to a set of simpler, canonical problems, we have developed an interaction theory for defects in infinite periodic arrays. This is similar in nature to the standard interaction theory for a finite number of bodies. The simplest case is that of an array with one or more absent scatterers. A straightforward extension to the theory that allows irregular scatterers to be considered has also been presented. The MASM is an effective means by which the canonical problems can be solved, and in particular enables important field characteristics such as RB surface wave amplitudes to be efficiently calculated to near machine accuracy. The canonical problems are independent of the defect type and configuration and all aspects of the incident field except the wavenumber, and need not be solved again if these parameters are changed.

Numerical results for various cases have been presented, with particular attention paid to the amplitude with which RB surface waves are excited. The MASM is particularly well-suited to cases in which the defects are localized. For defects that are spread over a larger section of the array, we have shown how other methods such as infinite array subtraction and the large array approximation can be incorporated so as to overcome the difficulties that arise. All of the results that we have presented involve arrays whose elements are circular cylinders. It is not difficult to modify our theory so as to account for other shapes by using transfer matrices. More complicated cases such as fully three-dimensional scattering problems can also be considered, provided that the relevant analogue to the theory of periodic arrays summarized in section 3 is available.

Appendix. Numerical quadrature. The most computationally expensive procedure in applying the interaction theory for defects in arrays is the evaluation of the integrals in (4.16) and (4.9). Quadratures must be performed on a contour whose orientation with respect to the branch points is the same as that shown in Figure 4.1, but in general it is convenient to move the path of integration away from

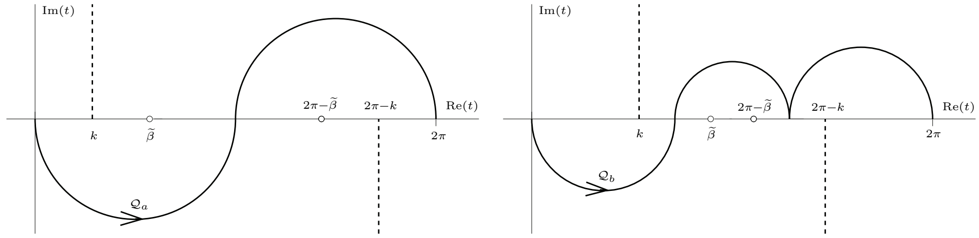


FIG. A.1. *Schematic diagrams of the contours $\mathcal{Q}_a$ and $\mathcal{Q}_b$ used for numerical quadrature.*

the various singularities. Since the residues of the functions $f_{m,m}(t)$ and $g_{m,n}(t)$ can be obtained using (4.20), the orientation with respect to the poles need not be maintained. In choosing an appropriate contour, a number of factors must be taken into consideration. These include the possibility of complex poles, contour length, proximity to the known real line singularities, and the behavior of the exponential term $\mathrm{e}^{\mathrm{i}pt}$ that appears in (4.9). Obviously, the extent to which a computer program can automatically adjust the contour to account for these factors has a significant effect on its overall complexity.

The paths of integration used by our numerical codes when RB modes are present are shown in Figure A.1. These are chosen for their relative simplicity, and we do not claim that they are optimal. For most parameter values, the distance between the two poles is at least as great as the distance between a pole and the nearest branch point, i.e., $2(\pi-\widetilde{\beta}) \geq \widetilde{\beta}-k$, and so we use the contour $\mathcal{Q}_a$, which consists of two semicircular arcs of radius $\pi/2$, centered at $t = \pi/2$ and $t = 3\pi/2$. As $k \to \pi$, $\widetilde{\beta} \to \pi$, and so the two poles move close together. If $2(\pi-\widetilde{\beta}) < \widetilde{\beta}-k$, then we integrate along $\mathcal{Q}_b$, which consists of two semicircular arcs of radius $(\widetilde{\beta}+k)/4$ centered at $t = (\widetilde{\beta}+k)/4$ and $t = 2\pi - (\widetilde{\beta}+k)/4$, and a third arc centered at $t = \pi$ with radius $2\pi - (\widetilde{\beta}+k)$. A residue contribution from the pole at $t = \widetilde{\beta}$ must be included in this case.

To deal with the possibility of complex poles, we introduce the function $d(t)$ as the determinant of the matrix on the left-hand side of (4.11) (also (4.14)), so that poles of $f_m^n(t)$ and $g_m^n(t)$ can occur only at points where $d(t) = 0$. We then numerically apply the principle of the argument [17, page 99] to $\log[d(t)]$ in the finite region(s) of the cut plane enclosed by the original path of integration (Figure 4.1) and the new contour (Figure A.1). Aside from $t = \widetilde{\beta}$ and $t = 2\pi - \widetilde{\beta}$, no poles that interfere with the deformations used here have been found. Additional poles were found on the line $t = \pi + \mathrm{i}u$, $u \in \mathbb{R}$, but only for large values of $|u|$. It should be noted that we have not searched exhaustively across the parameter ranges for a and k. A uniform partition of the contours $\mathcal{Q}_a$ and $\mathcal{Q}_b$ is used by our numerical codes, and the three-point Gaussian formula is applied on each subinterval. In cases involving multiple defects, efficiency can be greatly improved by storing values of $f_{m,n}(t)$ at the partition points used for the largest value of $|p|$ at which the integral in (4.9) must be evaluated and by making repeated use of these.

It is of some interest to compare the accuracy achieved by the MASM with that of the filtering technique [11], which can also be used to solve the canonical problems. The filtering technique requires the truncation and inversion of linear systems involving slowly convergent infinite spatial sums. We should therefore expect the MASM to achieve a superior degree of both accuracy and performance. An ideal parameter

Table A.1

Convergence in $|c_0|$ and performance in computing $|c_n|$ for $a = 0.25$ and $k = 2.5$.

Modified array scanning				Filtering			
NSI	$\lvert c_0\rvert$	% change	time (s)	SPT	$\lvert c_0\rvert$	% change	time (s)
99	0.1361053213		4.24	70	0.1360833934		4.36
149	0.1361053214	0.0000000039	6.40	90	0.1360096190	0.0542126301	8.55
199	0.1361053214	0.0000000003	8.57	110	0.1361554350	0.1072100718	14.33
249	0.1361053214	0.0000000000	10.72	130	0.1361523637	0.0022557485	21.68
299	0.1361053214	0.0000000000	12.90	150	0.1360562815	0.0705696206	31.55

for comparison is the quantity $|c_n|$, i.e., the amplitude of the RB wave that is excited by a source of order n replacing the scatterer centered at the origin. Table A.1 shows typical performance and accuracy figures that can be achieved by the two methods. The parameters used are $a = 0.25$ and $k = 1$, which lead to $\widetilde{\beta} \approx 2.586$, and the computations are performed using Fortran 2003 on a machine with a 2.5GHz processor. Note that the times given are those required for the simultaneous computation of c_n for all n up to the order truncation. The abbreviation NSI stands for the number of subintervals into which the contour is divided. The value for the spatial index p at which the linear system used in the filtering method is truncated is denoted by SPT. The dependence of computation time upon NSI is clearly linear, whereas increasing SPT leads to a significant decrease in performance. The results obtained by the two methods are in agreement up to the degree of accuracy that can be expected of the filtering method [11]; this requires the inversion of a large linear system of equations and is susceptible to round-off errors. It is clear that the MASM yields far greater accuracy, and is also much more efficient.

REFERENCES

[1] H. Ammari and F. Santosa, *Guided waves in a photonic bandgap structure with a line defect*, SIAM J. Appl. Math., 64 (2004), pp. 2018–2033.

[2] A.-S. Bonnet-Bendhia and F. Starling, *Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem*, Math. Methods Appl. Sci., 17 (1994), pp. 305–338.

[3] F. Capolino, D. R. Jackson, and D. R. Wilton, *Fundamental properties of the field at the interface between air and a periodic artificial material excited by a line source*, IEEE Trans. Antennas and Propagation, 53 (2005), pp. 91–99.

[4] F. Capolino, D. R. Jackson, and D. R. Wilton, *Mode excitation from sources in two-dimensional EBG waveguides using the array scanning method*, IEEE Microwave and Wireless Components Letters, 15 (2005), pp. 49–51.

[5] K. B. Dossou, L. C. Botten, S. Wilcox, R. C. McPhedran, C. M. de Sterke, N. A. Nicorovici, and A. A. Asatryan, *Exact modelling of generalised defect modes in photonic crystal structures*, Phys. B, 394 (2007), pp. 330–334.

[6] D. V. Evans and R. Porter, *Trapping and near-trapping by arrays of cylinders in waves*, J. Engrg. Math., 35 (1999), pp. 149–179.

[7] A. Khelif, A. Choujaa, B. Djafari-Rouhani, M. Wilm, S. Ballandras, and V. Laude, *Trapping and guiding of acoustic waves by defect modes in a full-band-gap ultrasonic crystal*, Phys. Rev. B, 68 (2003), 214301.

[8] C. M. Linton, *Schlömilch series that arise in diffraction theory and their efficient computation*, J. Phys. A, 39 (2006), pp. 3325–3339.

[9] C. M. Linton and D. V. Evans, *The interaction of waves with arrays of vertical circular cylinders*, J. Fluid Mech., 215 (1990), pp. 549–569.

[10] C. M. Linton and M. McIver, *The existence of Rayleigh-Bloch surface waves*, J. Fluid Mech., 470 (2002), pp. 85–90.

[11] C. M. Linton, R. Porter, and I. Thompson, *Scattering by a semi-infinite periodic array and the excitation of surface waves*, SIAM J. Appl. Math., 67 (2007), pp. 1233–1258.

[12] C. M. Linton and I. Thompson, *Resonant effects in scattering by periodic arrays*, Wave Motion, 44 (2007), pp. 167–175.

[13] H. D. Maniar and J. N. Newman, *Wave diffraction by a long array of cylinders*, J. Fluid Mech., 339 (1997), pp. 309–330.

[14] P. A. Martin, *Multiple Scattering. Interaction of Time-Harmonic Waves with N Obstacles*, Cambridge University Press, Cambridge, UK, 2006.

[15] P. McIver, C. M. Linton, and M. McIver, *Construction of trapped modes for wave guides and diffraction gratings*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 2593–2616.

[16] B. A. Munk and G. A. Burrell, *Plane-wave expansion for arrays of arbitrarily oriented piecewise linear elements and its application in determining the impedance of a single linear antenna in a lossy half-space*, IEEE Trans. Antennas and Propagation, 27 (1979), pp. 331–343.

[17] A. D. Osborne, *Complex Variables and Their Applications*, Int. Math. Ser., Addison Wesley Longman Limited, Harlow, UK, 1999.

[18] R. Porter and D. V. Evans, *Rayleigh-Bloch surface waves along periodic gratings and their connection with trapped modes in waveguides*, J. Fluid Mech., 386 (1999), pp. 233–258.

[19] M. M. Sigalas, *Elastic wave band gaps and defect states in two-dimensional composites*, J. Acoust. Soc. Amer., 101 (1997), pp. 1256–1261.

[20] I. Stakgold, *Green's Functions and Boundary Value Problems*, 2nd ed., Wiley, New York, 1998.

[21] S. V. Sukhinin, *The whispering surface effect*, J. Appl. Math. Mech., 63 (1999), pp. 863–876.

[22] I. Thompson and C. M. Linton, *An embedding method for scattering by defective arrays*, in Proceedings of Waves 2007, Reading, UK, 2007, pp. 203–205.

[23] I. Thompson and C. M. Linton, *On the excitation of a closely spaced array by a line source*, IMA J. Appl. Math., 72 (2007), pp. 476–497.

[24] I. Thompson, C. M. Linton, and R. Porter, *A new approximation method for scattering by long finite arrays*, Quart. J. Mech. Appl. Math., published online March 27, 2008; doi:10.1093/qjmam/hbn006.

[25] V. Twersky, *Elementary function representation of Schlömilch series*, Arch. Rational Mech. Anal., 8 (1961), pp. 323–332.

[26] S. Wilcox, L. C. Botten, R. C. McPhedran, C. G. Poulton, and C. Martijn de Sterke, *Modeling of defect modes in photonic crystals using the fictitious source superposition method*, Phys. Rev. E, 71 (2005), 056606.

[27] C. P. Wu and V. Galindo, *Properties of a phased array of rectangular waveguides with thin walls*, IEEE Trans. Antennas and Propagation, 14 (1966), pp. 163–173.

[28] F. Wu, Z. Liu, and Y. Liu, *Splitting and tuning characteristics of the point defect modes in two-dimensional phononic crystals*, Phys. Rev. E, 69 (2004), 066609.